Forthcoming in: *Inner Speech*, Langland-Hassan & Vicente, eds., OUP.

*This is a late-stage draft. Comments welcome.*

# From Introspection to Essence: The Auditory Nature of Inner Speech [*]

Peter Langland-Hassan

*University of Cincinnati*

## 1. Introduction

*Inner speech always has an auditory-phonological component*. To some this claim is a truism, a platitude of common sense. To others, it is an empirical hypothesis with accumulating support (Loevenbruk *et al.*, this volume). To yet others it is a false dogma (Gauker, 2011, this volume). I defend the claim in this chapter, confining it to adults with ordinary speech and hearing.

To those already convinced that inner speech has an auditory-phonological component, I urge caution and patience. For it is one thing to assert that inner speech often, or even typically, has an auditory-phonological component—quite another to propose that it *always* does. When forced to argue for the stronger point, we stand to make a number of interesting discoveries about inner speech itself and about our means for discriminating it from other psycholinguistic phenomena. Establishing the stronger conclusion also provides new leverage on debates concerning how we should conceive of, diagnose, and explain auditory verbal hallucinations and "inserted thoughts" in schizophrenia. Or so I will argue in this chapter's final section.

In saying that inner speech always has an auditory-phonological component, I mean to say that it has *sensory character* tied to the auditory modality—and to phonemes in particular.[1] Phonemes are the

*Special thanks to Agustín Vicente, Fernando Martínez-Manrique, Franz Knappik, Jordan Ochs, and Daniel Gregory for critical comments that improved this chapter.

[1] I won't offer a tight definition of what it is to have sensory character. My own view is that a mental state's having sensory character is just a matter of its representing certain properties of the world in a fine-grained, nonconceptual manner, distinctive of some sense modality. But other views that reject a strong connection between a mental state's

smallest distinct building blocks of sound in a spoken language relevant to assessing a word's meaning. Differences in the phonemes /l/ and /s/, for instance, allow us to aurally distinguish the words *kill* and *kiss*. There are 44 phonemes in English, combinations of which make up the sound of every English word. To anticipate later discussion: the question of what exactly a phoneme is, or must be—including whether phonemes are inherently *auditory*—is more delicate than I've just let on, and subject to controversy.[2] But for now you grasp my meaning: the inner speech of adults with ordinary speech and hearing always has sensory character related to phonemes, and that sensory character is also auditory in nature. To be clear, this is a claim about one of inner speech's essential components. It leaves the door open to inner speech having additional essential components, and to its having any number of other components non-essentially.

It is not a revolutionary idea that inner speech has an auditory phonological component. Many theorists already work from the premise that one of inner speech's central features is its auditory-phonological character (Carruthers, this volume; Clark, 1998; Jackendoff, 1996; Langland-Hassan, 2014; M. Perrone-Bertolotti, Rapin, Lachaux, Baciu, & Loevenbruck, 2014). At the same time, it is not always obvious *why* people have this view. The first-pass folk psychological characterizations we use to describe inner speech do not make essential reference to sensation or perception. Inner speech is said to be "the little voice in the head"; it is "talking to oneself silently"; it is "verbal thought" or "thinking in words." These characterizations make reference to language and to speaking; but they are, for the most part, neutral on whether the phenomenon has a sensory component. True, the notion of a "voice" in the head suggests audition, as voices are typically *heard*. Yet inner speech is equally said to be silent. Moreover, *producing* a voice (and a sound) is one thing; hearing it is another. There is no incoherence in the idea that inner speech might be the production of a voice in the head, without accompaniment by the kinds of auditory-perceptual states by which speech is normally heard. Nor is there any conceptual confusion in the claim that we "think in words" that themselves lack a sensory component.

Unsurprisingly, then, some have openly rejected the idea that inner speech is itself sensory in nature. Christopher Gauker (2011, present volume) draws a strict distinction between inner speech, on the one hand, and the auditory verbal imagery that often accompanies it, on the other. According to Gauker, inner speech is a means for *thinking in words*. These inner words are often represented by auditory verbal imagery, Gauker maintains, but are themselves are entirely lacking in sensory character.

---

sensory character and what it represents are consistent with the argument I will give. For instance, if we can meaningfully speak of a mental state's *having* auditory-phonological properties (as opposed to *representing* such properties), then my claim is that inner speech always has such properties.

[2] Linguists typically distinguish phonemes from *phones*, where phones are concrete sound events that are heard as a single phoneme. See fn. 6 for further discussion of this distinction as it relates to the arguments put forward here.

Few others draw so strict a distinction between inner speech itself and auditory verbal imagery. Yet it is common to leave the door open to at least *some* inner speech lacking sensory character. Wayne Wu proposes that, while both auditory verbal hallucinations and inner speech are experiences of language, "the latter seems to be more often abstracted from an auditory format, namely without representation of audible properties" (2012, p. 96). Fernyhough and Alderson-Day also play down the centrality of auditory imagery to inner speech: "At its core," they propose, "inner speech is an abstract linguistic code, that shares more resources with overt speech production than does auditory imagery" (2015, p. 24). And Hurlburt, Heavey, & Kelsey emphasize that inner speaking—which they distinguish from "inner hearing"—is, "more a phenomenon of created action than of received audition" (2013, p. 1482).

So, while few will deny that auditory-phonological imagery is closely associated with inner speech (not even Gauker denies an *association*), the depth and nature of the association remain open questions. What, then, can be said in defense of the idea that auditory-phonological imagery is a strictly essential component of inner speech? One style of answer appeals to behavioral and neuroimaging data from tasks that are assumed to draw upon inner speech. These tasks include activities like maintaining a list of words in working memory (Conrad & Hull, 1964), silently judging whether two words rhyme (Geva, Bennett, Warburton, & Patterson, 2011; Langland-Hassan, Faries, Richardson, & Dietz, 2015), and silent reading. If performance on such tasks is influenced by the auditory-phonological features of the words involved, this can be taken as evidence that inner speech has an auditory-phonological component. Similarly, if neuroimaging shows special activation in auditory speech-perception areas during such tasks, an inference that inner speech has an auditory-phonological component again seems warranted (Marcela Perrone-Bertolotti et al., 2012; Yao, Belin, & Scheepers, 2011).

I have no brief against these ways of approaching the question of whether inner speech has an auditory-phonological component. In fact, I recommend them. However, it is important to see their limitations. Chief among them is their scope: what we discover in such studies, at most, is that the cognitive states exploited in a certain experimental task have an auditory phonological component. This conclusion leaves open the possibility, first, that the cognitive states, while language-related, were not *inner speech*. How could they not be? First, if they are not in fact the kinds of states that people typically pick out with commonsense expressions like "talking to yourself silently" or "thinking in words," one could say that the target was missed. Also, it is possible that more than inner speech itself is elicited in such tasks, and that this "something more" is what has the auditory-phonological component. (This would appear to be Gauker's view). Third, granting that the cognitive states exploited during the task were indeed cases of inner speech, there may be many *other* states that we introspectively mark as inner speech—as "thinking in words," or as a "little voice in the head"—yet which lack an auditory-phonological component. For it could be that the task incorporated an exceptional kind of inner speech—

one that was phonologically enhanced, or that activated auditory verbal imagery in a special way. That would be consistent with much, or even most, of ordinary inner speech having no auditory-phonological component at all.

Alternatively, we might consult introspection. We might generate inner speech episodes and simply ask ourselves whether they have auditory sensory character related to phonemes. However, this method has many of the same drawbacks. It may be that, in judging our inner speech to have auditory-phonological sensory character, we are mistaking a common accompaniment of inner speech for something that is essential to it. It could also be that self-conscious reflection on whether our inner speech has auditory-phonological sensory character leads us to generate an unusual, phonologically enriched kind of inner speech—just as focusing on one's own breathing leads to irregular breathing. (See Hurlburt & Heavy (this volume) and Hurlburt & Schwitzgebel (2007), for further reflection on the pitfalls of casual introspection). The method also questionably assumes that we have an introspective grip on what it is to have auditory-phonological sensory character adequate to judging the kinds of borderline cases that inner speech is likely to present.

In any case, the question of whether inner speech has an auditory-phonological component interests me in part because my own introspection doesn't return a clear verdict. Does there *always* seem to be something auditory about my own inner speech? *Maybe*. It's hard to say. I wouldn't bet the farm on it. Nevertheless, I think there is an argumentative route that begins with introspection and leads quickly to the conclusion that inner speech must have an auditory-phonological component—whether it seems that way to introspection or not! Let's consider that argument now.

## 2. Why inner speech must have an auditory-phonological component

*Inner speech is always keyed to a specific natural language.*[3] I will take this claim as bedrock. It is something we can't give up without losing the ability to distinguish inner speech from most other mental phenomena—including visual images, emotions, thoughts of other stripes. Saying that inner speech is keyed to a specific natural language may not suffice to distinguish it from all other mental phenomena; but it's a necessary first step, one that respects our introspective means for identifying inner speech in the first place. We notice a "little voice" saying words *of a specific language*; we sense that we are "thinking in words" *of a specific language*; we notice that we are "talking to ourselves silently," *in a*

[3] Most would say that inner speech always occurs *in* a language, not that it is merely "keyed to" a language. I urge caution on that point. It could be that inner speech episodes *represent* linguistic items—words and sentences—without themselves occurring *in* a linguistic format. (See Langland-Hassan (2014)). Hence my use of the "keyed to" terminology, which aims at neutrality between those options.

*specific language*. From there we can conduct psychophysical and neuroimaging experiments that tell us more about the nature of the phenomenon, so identified.

Whenever we recognize expressions as being expressions of a certain language, we face the question of how we determined them to be expressions *of that language*, and not some other. This is true whether the expression is discerned through vision, audition, touch (as in the case of Braille), or introspection. For any expression *S* that we judge to be keyed to a particular language, there must be some feature of *S* that enabled us to tell whether *S* occurred in English as opposed to, say, French or Spanish.

My argument will be that the only feature inner speech episodes plausibly have that would allow us to swiftly and reliably determine *which language* they are keyed to is their auditory-phonological component. To get to that conclusion, we need to consider what other possibilities there might be. We can start by considering the most salient features of words and sentences and asking whether those features might reveal to us the language in which they occur. Four immediately come to mind: *semantics*, *syntax*, *phonology*, and *graphology*. (A fifth feature, related to the bodily movements through which sentences are *articulated*, will be considered in Section Three).

Let's begin with semantics. To say that a sentence has a semantics is to say that it has a meaning. Sentences can be assessed as true, false, or indeterminate, in virtue of their semantics. Awareness of a sentence's semantic features, however, won't suffice to tell us which language we are encountering. This is because sentences of different languages can have essentially the same semantics. We can translate almost any English sentence to sentences of other languages; when we do so successfully, we create another sentence with the same semantics.[4] The problem is not that some English sentences have the same meaning as a sentence of another language. Rather, it is that *almost every* English sentence has the same meaning as some sentence in almost every other language. Supposing we could "directly see" the meaning of a sentence, simply apprehending that meaning would not tell us which of many possible languages we were encountering.

The same points go for a sentence's syntactical structure. Sentences have a syntax insofar as the words that make them up play certain standard linguistic roles, such as being the subject, verb, or object of the sentence. There are different methods for mapping out the syntactical structure of a sentence, assigning to each word or phrase a particular syntactic role. The syntax trees in Figure 1 show two methods for mapping the syntactic structure (or "frame") of the sentence 'John has finished the work':

[4] Some views of semantics might have it that the words of different languages can never have *precisely* the same meanings. Nevertheless, to the extent that the words customarily used to translate terms from one language to another differ in semantics, these subtle differences are not usually ones that fluent speakers of the two languages would be in a position to notice or explain. That is all that the argument requires.

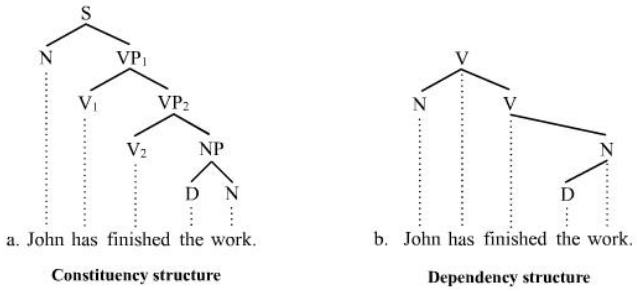


*Figure 1*

Now imagine that we were able to "see directly" the syntactical structure of a sentence, abstracting away from its specific words. Supposing we were looking *only* at syntactic frame, we would be at sea in determining which language we were apprehending. Any English (or French, or Spanish…) sentence will share the very same syntactic frame with arbitrarily many other sentences of English, and with arbitrarily many sentences of other languages. Syntactic structure-wise, 'John finished the work' and 'Jane ran the race' are exactly alike. Matters are not much improved if we combine semantics and syntactic structure. Often enough, a sentence of one language will have *both* the same semantics and syntactic structure as a sentence of another language. For instance, 'John has finished the work' and 'John a terminé le travail', share the same semantics and syntax. Yet any bilingual speaker of English and French has no trouble at all distinguishing whether she has just heard the English or French version of the statement. She must do so by noticing something other than the sentence's semantics and syntactic structure.

This point may seem trivial in the case of the sentences we see on the page or hear in conversation, as no one thinks we in fact "see directly" through to the syntax and semantics of such sentences. To have any awareness of the sentence at all, we first have to either see the graphic features of words—their distinctive shapes and colors—or hear their phonetic features. But matters are different in the case of inner speech. Most theories in psycholinguistics break speech production into discrete stages. The most influential theory holds that language production begins with a *message* to be conveyed, where the message has a certain semantic content; then, a syntactic frame appropriate to linguistically encoding that message is selected; this leads, in turn, to selection of the language-specific phonemes relevant to each syntactic slot; with those choices in place, articulation of the phonemes—and the words they constitute—can begin (Bock & Levelt, 1994; Levelt, 1989; Postma, 2000). Supposing each stage of this processing is neurologically discrete, it is at least possible that a person would have introspective access to the representations at each stage. This would be the sort of "direct" access to syntax and semantics that

we lack when we see or hear a sentence. But the fact that such features massively underdetermine the languages that they partly characterize shows that direct awareness of those features will not tell us which language we are apprehending.[5] To the extent that our inner speech nevertheless seems to be keyed to a specific language—incorporating specific words of that language—that appearance cannot be explained by appeal to our awareness of semantic and syntactic stages of language processing alone.

A response at this point might be that inner speech typically occurs in more than single sentence or single word bursts and that, in longer sequences, there will rarely be a complete overlap in the semantics and syntactic structures of the sentences that would translate one language into the other. This is true, however the point remains that we are usually able to distinguish the language to which our inner speech is keyed one sentence at a time, and even one word at a time. This cannot be explained by appeal to language-specific semantic and syntactical properties that emerge only in larger groups of sentences.

It seems we are left with phonemes and graphemes as the only features of expressions we could reliably exploit to determine their language. I have already described phonemes as the most basic, repeatable units of sound that make up the spoken words of a language (where changes in the sounds correspond to changes in word meaning). In a moment I will consider whether phonemes must always be sounds.[6] But first let's consider graphemes. Graphemes are clusters of written characters, or single written characters, that, when spoken, generate a single phoneme (Berndt, D'autrechy, & Reggia, 1994; Shallice, 1988). So, for instance, the word 'school' has four phonemes: /s/, /k/, /ew/, and /l/. Finding the graphemes in 'school' amounts to finding the written characters in the word that correspond to each phoneme. Thus, the graphemes for 'school' are <s>, <ch>, <oo>, and <l>. Note that there is no phoneme-independent characterization of graphemes, so understood, because graphemes are simply the

[5] One might object that, if a person's comparison class of languages is limited to one or two, such underdetermination will rarely occur. I address this worry in Section 3.1 below. For now, note that a person fluent in five or six languages will have no more trouble determining the language to which her inner speech is keyed than a person who knows only one; yet, she obviously should have much more trouble if she is relying upon semantics and syntax alone.

[6] In linguistics, phonemes are typically distinguished from *phones.* Phones are sometimes described as "concrete" sounds events—having a duration, spatial location, and so on—whereas phonemes are described as "abstractions" over phones, such that a phoneme is a set of phones that tend to be heard as the same linguistic building-block. Whereas changing a phoneme in a word changes the word's meaning, altering a phone may not. (Two phones heard as the same phoneme are called *allophones.*) This might make it seem as though phonemes are not, in fact, sounds, but are something more abstract—something imperceptible. But that would be a confusion. Phonemes are no more "abstract", or non-concrete, than are, say, *mammals,* where the set of mammals includes many different species. Mammals are things we see and hear, even if the set of mammals is not. The same goes for phonemes: their instances are concrete and perceptible, but the type itself involves an abstraction. This is true of phones as well, insofar as we can speak of different *types* of phones. Indeed, the convention of using square brackets [ ] to symbolize phones requires that phones can be typed into reoccurring kinds that abstract away from potential differences among their concrete instances. Thus, phonemes, in their instances, are just as concrete as phones are in theirs. There is, therefore, no confusion in describing phonemes as events we aurally represent and, in so doing, perceive.

characters in any written word that correspond to the word's phonemes. (By contrast, the notion of a *letter* or *character* is phoneme-independent, as a single letter—e.g. 'c'—can be associated with multiple phonemes.) There is evidence that graphemes, and not letters, are the primary perceptual units during reading (Rey, Ziegler, & Jacobs, 2000); hence my focus on them here. But the points I make will hold equally for letters.

Unlike semantics and syntactic structure, the phonemic and graphemic properties of sentences are highly language-specific. It is very rare for sentences of different languages to have the same phonemes or graphemes, in the same order.[7] This means that the phonemic and graphemic structures of sentences are the right sorts of things to serve as "signatures" for that language—features by which a sentence can reliably be discriminated as occurring in the language it does. As you read this sentence, you are able to discern it as English thanks to your ability to discern grapheme combinations characteristic of English—and, most likely, to translate those graphemes into phoneme sequences distinctive of English. The question is whether we judge our own *inner speech* to occur in a specific language by discerning graphemes.

We can quickly rule out the hypothesis that we *always* make use of graphemes to determine the language to which our inner speech is keyed. We perceive graphemes visually, by noticing the shapes of written characters and character combinations. "Sounding out" words when learning to read amounts to visually discerning graphemes and translating them to familiar phonemes. If inner speech always required the representation and discrimination of graphemes, people incapable of making those discriminations—such as the blind, and the illiterate—would be incapable of inner speech. Yet the blind and the illiterate no more lack inner speech than they lack outer speech. Consider also that five year old children, whose reading abilities are, on average, minimal, have been shown to employ "covert" linguistic rehearsal during working memory tasks (Johnston, Johnson, & Gray, 1987). As with adults, their performance was subject to the phonemic similarity effect, which strongly suggests that they were using inner speech as part of the memory task (Johnston & Conning, 1990). This is good evidence that something we are normally prepared to count as inner speech—namely, the resource used during verbal working memory tasks—*precedes* the acquisition of reading abilities, and so does not rely upon the sort of graphemic discrimination that reading requires.

Nevertheless, it still could be that we sometimes visualize words and that this is "thinking in words," in a sense. On what grounds can we say this is not a kind of inner speech—one where we exploit

[7] Though such mirroring occurs on occasion. An example is the phoneme sequence corresponding to 'Empedocles leaped'. The German 'Empedocles liebt' uses the same phoneme sequence but means that Empedocles loved (whether or not he leaped in the process). Thanks to Daniel Gregory for the pointer.

graphemes in order to judge the language to which our inner speech is keyed? If it is, and if it also lacks an auditory-phonological component, this would falsify the claim that *all* inner speech has an auditory-phonological component. My response is that such (partly) visual episodes don't lack an auditory-phonological component. We know this because silent reading has a strong auditory-phonological component, as evidenced by well-known phonological effects on lexical decision tasks. In these tasks participants are shown a string of letters and asked if it corresponds to a real word. Responses to letter strings (such as *shrood*) that sound like a real word but are not spelled like one, are slower and more error-prone than responses to strings of letters (such as *slint*) that neither look nor sound like a real word (Besner & Davelaar, 1983; Coltheart, Davelaar, Jonasson, & Besner, 1977). The effect appears due to a phonological stage in the comprehension of written text. And there are complementary findings, such as that when skilled readers are given information about the accent and speaking rate of a text's author, this influences the rate at which they read the text (Alexander & Nygaard, 2008). This all fits with the developmental fact that learning to read involves learning to translate unfamiliar graphemes to familiar phonemes. Further, a number of imaging studies have shown that neural areas underlying auditory speech perception are differentially activated during silent reading (Kell et al., 2017; Marcela Perrone-Bertolotti et al., 2012; Yao et al., 2011). Taken together, the data on silent reading strongly suggest that reading involves a stage where visually perceived graphemes are translated to states with an auditory-phonological component. It therefore stands to reason that the visualization of words does as well, whenever such words are processed as having a semantic content (and not merely as colors and shapes). Thus, even if we accept cases where we visualize words as episodes of inner speech, we can infer that those episodes have an auditory-phonological component in addition to their visual component. The strong conclusion that all inner speech has an auditory phonological component remains viable.

Of course, there is little empirical or introspective reason to think that most inner speech involves a visuo-graphic element. Still before us, then, is the question of which feature we *normally* exploit to determine the language to which our inner speech is keyed. Here phonemes are the only plausible answer. In contrast to graphemes, there is strong *prima facie* evidence that inner speech often has an auditory-phonological component. Many find it intuitive to describe inner speech in auditory terms. For instance, Cassam thinks that "auditory metaphors are virtually inescapable. The sense in which one is aware of inwardly saying to oneself that *P* is that one 'hears' oneself saying to oneself that *P*. This hearing is with the mind's ear rather than with the ears attached to one's skull" (Cassam, 2011, p. 10). Further, experimental tasks that, intuitively, draw upon inner speech—such as remembering lists of words, silently reading text, or silently judging rhymes and homophones—show interference and performance effects related to the phonemes of the words involved. (See Loevenbruck *et al.* (this volume) and Geva (this volume) for reviews). The difficult question is whether inner speech *invariably*

has an auditory-phonological component. My answer is that, *to the extent that we are able to recognize our own inner speech as occurring in a particular natural language*, then it must have an auditory-phonological component. For this is the only feature of our inner speech that we could plausibly exploit in making that determination. The other salient features of sentences are either too ambiguous between and within languages (as was the case with semantic and syntactic features), or not plausibly instantiated by most episodes of inner speech (as was the case with graphemes).

Seeing as we already have empirical and intuitive reason to think that much inner speech has an auditory-phonological component, this component remains only reasonable candidate for what it is that makes our inner speech seem to us to occur in a specific language (*modulo* the discussion of graphemes and silent reading, above).

*2.1. From phenomenology to essence*

The argument just given has an important limitation. It makes a positive claim about inner speech episodes that seem, to the person having them, to be keyed to a specific language. It explains why such episodes seem that way to the person, by appeal to their auditory-phonological components. But the argument says nothing about episodes of inner speech that do not seem, to the person having them, to occur in a natural language. And it is safe to assume that *unconscious* inner speech, should it occur, will seem no way at all to the person having it. So my argument does not directly show that unconscious inner speech has an auditory phonological component. However, a few more steps can take us there.

First, it is reasonable to think that an episode of unconscious inner speech will be just the sort of state that, when conscious, is picked out in the commonsense ways of identifying inner speech. For it will be the type of state that, when conscious, seems to be keyed to a specific natural language; the best explanation of why it seems that way, when conscious, is that it has an auditory phonological component, even when unconscious. To say that unconscious inner speech only gains its auditory phonological component once it is made conscious begs the question of why, in its pre-conscious version, it is proper to consider it an episode of inner speech. For it assumes that the very feature which leads people to describe inner speech as being keyed to a language is not, in fact, one of its essential features. It is possible that it

is not. But we would need a good reason to think it is not; and introspection cannot offer one, as all the inner speech introspection detects appears keyed to a natural language.[8]

A better rationale may appear to lie in the cognitive scientific study of language-processing. From the perspective of psycholinguistics and cognitive neuroscience, it may seem propoer to call the activation of *any* neural areas or cognitive systems involved in language production "inner speech", regardless of how they seem to the person having them. However, linguistic processing in this broad sense goes on during both inner and overt speech. We could, perhaps, specify that such processing constitutes inner speech *only* when overt speech does not also occur. But in adopting that kind of negative criterion for inner speech, we risk having changed the topic. Inner speech is a specific, salient element of our conscious lives that we want to investigate. We cannot back our way into studying it by identifying it with all the internal aspects of speech production that take place whenever overt speech does not also occur. For there's no reason to think that this negative proviso—viz, *whenever overt speech does not also occur*—lands us at the positive phenomenon we initially wanted to study.

[8] This argument may seem to fall flat from the perspective of Christopher Gauker's view of inner speech (Gauker, 2011, pp. 257-260; this volume). As mentioned, Gauker holds that auditory verbal imagery serves to make us aware of our inner speech by *representing* it (just as auditory-verbal perceptual states serve to make us aware of the overt speech of others), and that inner speech itself is a separate entity from that auditory verbal imagery (just as the speech of others is separate from our representation of that speech). So Gauker might agree that it is the auditory phonological component of our auditory verbal imagery that explains why our inner speech seems to us to occur in a specific natural language, while objecting that we cannot conclude from this that inner speech *itself* has an auditory-phonological component. Yet this leaves Gauker with at least two awkward consequences, both of which he confronts in his chapter for this volume. The first is that, on his view, our sole means for becoming aware of our own inner speech—auditory verbal imagery—*always misrepresents* that speech as being composed of phonemes. (Assuming that phonemes are sounds and that Gaukerian inner speech, *qua* language-like neural process, is silent). For while there is no tension in the idea that auditory verbal imagery might *occasionally* mispresent a neural event as composed of sounds, the idea that it *always* does so leaves us with no account of how such a misrepresentational faculty could provide useful information about what it continually (mis)represents. How, we might ask, are we able to recover accurate information about the semantics of such inner speech, if it is continually misrepresented? An alternative for Gauker would be to follow the motor theory of speech perception (discussed below, Section 3.3) in holding that the perception of phonemes during ordinary speech perception is in fact the perception of neural states of a kind (*viz.*, the speaker's articulatory intentions). This might suggest that inner speech, as he conceives it, has phonological features after all—and that the phonemes are simply neural events of a kind. At that point we would be in agreement that inner speech always has a phonological component (though perhaps not an *auditory* one), despite disagreeing over whether auditory verbal imagery ought to be considered a proper component of inner speech. Gauker is still, however, left with the question of how we become aware of the semantics of our own inner speech, if our auditory verbal imagery only makes us aware of its phonological features.

The second awkward consequence for Gauker lies in the fact that most other forms of imagery—including visual imagery, and, by Gauker's own account, some episodes of auditory verbal imagery—do not represent actual internal mental events, but rather non-present worldly objects and scenarios. For instance, a visual image of a yellow toaster represents a (non-present) yellow toaster, and not the brain state that gave rise to it; likewise, as Gauker notes, the auditory verbal imagery used in imagining the sound of one's own voice as one greets a friend represents a (non-existent) outward utterance, and not a prior brain state that was its cause. Gauker is left with the question of how auditory imagery is able to "go both ways", representing both internal and external objects, when other forms of sensory imagery do not.

Such an approach would be akin to identifying visual imagery with the endogenous activation of *any* neural areas involved visual perception. But this is not, in fact, how visual imagery is identified by those studying its neural bases. Instead, researchers begin with a task that, intuitively, requires visual imagery and then search for related neural activation (Ganis, Thompson, & Kosslyn, 2004; Kosslyn et al., 1999; Slotnick, Thompson, & Kosslyn, 2005). Applying the same standard to inner speech, the experimental tasks taken to elicit inner speech—and to reveal its underlying bases—should only be those where the subject reports experiencing a mental phenomenon keyed to a particular natural language.

This is not to say that studying all of the many cognitive components—conscious and unconscious—of language production and perception is not an important project. Of course it is. It is only to insist that *inner speech* be treated as a particular subset of those phenomena—one closely tethered to the subjective experience of language. To maintain this connection is not to wall-off inner speech from empirical study. It is, instead, to maintain the possibility that meaningful links will emerge between a central element of our conscious lives—the little voice in the head—and its cognitive and neurological bases. Short of a strong theoretical reason for dissociating inner speech from the feature by which it is introspectively identified, we should accept that all inner speech—conscious or not—has an auditory phonological component.

## 3. **Some objections considered**

The argument of the previous section went quickly in places and passed over some controversial issues. Here I want to slow down to consider some reasonable objections.

### *3.1. Objection: I usually speak English; that's why my inner speech always seems to be in English*

The first objection is that there are simple probabilistic heuristics that could be used when judging whether a sentence—internal or external—occurred in a particular language, and that such heuristics don't require us to exploit the auditory-phonological features of sentences. For instance, suppose I know that my neighbor Barry is a monolingual English speaker. I see him in his front yard, talking with the mail carrier. Even though they are too far away for me to hear them, I can reliably judge Barry to be speaking English, just because I know that he is monolingual. By the same token, because most of my inner speech is keyed to my native language of English, I could reliably assume that any of my inner speech is keyed to English, without noticing any auditory-phonological component it may or may not have.

For those who mainly speak a single language, reliable use of such a heuristic is certainly possible. Could this be in fact what goes on when a person judges her inner speech to be keyed to a certain language? I don't think so. Go back to my neighbor, Barry. If, unbeknownst to me, Barry has acquired some Spanish and is engaging the mail carrier *en Español*, I won't realize it. I'll still assume he's speaking English. Given the nature of my heuristic and the fact that I can't hear him from my distance, I'm in no position to make any surprising discoveries about the language he's speaking. But when we notice ourselves to be "thinking in words" that are keyed to a certain language, we *are* in a position to notice changes in the language. We are in a position to notice sudden switches in the language to which our inner speech is keyed, just as we are in a position to notice a change in the language Barry is speaking whenever we can *hear* him. The fact that our inner speech episodes invariably have an auditory-phonological component can explain our being in such a position; a probabilistic heuristic cannot.

This is not to say that we will always take note of such a switch when it occurs. It is just to say that we are always in a good position to notice such a shift, should we attend to the matter. By the same token, it seems absurd to hold that a bilingual who speaks two languages equally often has only a 50% chance of correctly determining the language to which her inner speech is keyed; yet her chances should be no better than that if she is relying upon a probabilistic heuristic.[9] Her far greater chances are again best explained by her inner speech's having language-specific auditory phonological components that allow for the discrimination.

With this in mind, we can consider a closely related challenge. There are exceptional cases where two languages overlap in semantics, syntax, and phonology. Fernando Martínez-Manrique mentioned to me the example of Spanish and Galician (Galician is a romance language in northwestern Spain). While the two languages typically differ at the sentence level, they share number of expressions that are the same (e.g. *gato negro*). Suppose Marta is fluent in both languages and speaks both regularly, with the same accent. How will she determine the language to which her fragmentary inner utterance of "gato negro" is keyed? (Supposing there are no special contextual cues she can exploit, such as that she has just been conversing in Galician). My answer is that she will not be able to—or, at least, that we have no good reason to think she will.

If this response seems question-begging, we need only ask ourselves how else she might know. Assuming she couldn't tell the difference if she simply heard someone else speak the two words (again, assuming they spoke Spanish and Galician with the same accent), what might she have access to in her own case that would enable her to make the discrimination? There remains a possibility highlighted by

[9] Thanks to Jordan Ochs for suggesting this point.

this case: Marta might know the language by *knowing her own intentions* to speak either Spanish or Galician. This is the next objection I will consider.

### *3.2. Objection: My intentions reveal to me the language to which my inner speech is keyed*[10]

When we speak to others, we typically have an intention to do so. On a Gricean (1957) picture of linguistic communication, successful communication relies in part upon an audience recognizing that intention in the speaker. If intentions to speak usually accompany our overt utterances, it might seem reasonable to think they accompany our inner utterances as well. In that case, one might think that our inner speech episodes seem to be keyed to a certain language just because we are aware of intentions to generate speech in that language.

Yet, even granting the (questionable[11]) view that we typically form intentions to generate our inner speech episodes, there is no reason to think that those intentions carry with them an explicit specification of the language to be generated. I might intend to say "Let's go home," without intending to say something in English—even if I do in fact say something in English. So, awareness of my intention will not settle the question of the language to which the utterance is keyed. Nor does all of our inner speech plausibly result from an intention to generate it. At times we may intend to stop carrying on an imagined conversation, yet fail. The resulting inner speech still seems keyed to a language, despite the absence of a relevant intention.

However, there remains another sense in which our cognitive systems may already know which language is being generated, as it is being generated. This is because—on a number of leading models of speech production (Postma, 2000)—there are monitoring mechanisms that check whether proper syntactic structures—known as *syntactic frames*—are being generated. A particular syntactic frame might specify that the utterance to come will have a subject, verb, and object, in that order. These mechanisms are held to operate prior to the selection of phonetic features of the words that will fill the spaces in the frame, yet must also be sensitive to the particular grammatical rules of the language. For instance, one language may place the object of a verb before the verb, and another after it. To properly monitor the generation of these syntactic frames, the system must, in a sense, "know" which language is being generated at any

[10] Thanks to Fernando Martìnez-Manrique for pressing me to be clearer on the possible role of intentions and syntactic-frame monitoring in inner speech.

[11] After all, some inner speech episodes might themselves *be* occurrent intentions (or decisions). And in response to the objection that inner-speaking is something we *do*, and must therefore be *intentional*, inner speech is only invariably an intentional action insofar as thinking, in general, is an action; and, on pain of regress, not all thinking can be generated by an intention, so long as intentions themselves are thoughts of a kind.

given time, so as to apply the proper norms. And one might think that this knowledge, whatever its form, is what the subject accesses when judging her inner speech to occur in a specific language.

However, in contrast to intentions, it is unlikely that we have introspective access to this sort of linguistic know-how, such that it could be what makes our inner speech appear keyed to a particular language. Suppose, for instance, that you have a song lyric stuck running through your head. We know there is no intention to continue repeating the lyric. Are we aware of something else—a kind of motor command—that is its cause? There must *be* something of that sort, in virtue of which the lyric continues repeating. But we are not aware of it. The same points apply to distracting inner speech that we wish to stop generating, and which is unintentional. Whatever motor system commands are involved in generating those episodes—and in ensuring that the proper syntactic norms are applied to their monitoring—they are not the kinds of states that we are in a position to introspectively report, and that could be what we are introspectively aware of when our inner speech appears keyed to a particular language.

### *3.3 Objection: Inner speech could have a phonological component without being auditory*

A third objection takes issue with the assumption that there is something inherently auditory about phonemes. My claim, recall, is that inner speech always has an *auditory*-phonological component. But if there is nothing auditory about phonemes, then there need not be anything auditory about inner speech—*even if* we were committed to its always having a phonological component. As earlier noted, my thesis is limited to the inner speech of adults with ordinary language abilities. Even so, reflection on the language—and inner language—of the deaf is relevant here.

Theorists studying American Sign Language (ASL) are in agreement that it has a phonology (Sandler, 2012; Stokoe, 2005). But ASL does not involve the production or perception of any sounds. How, then, could phonemes be auditory in nature? To answer, we need to look at what theorists have in mind when they say that ASL has a phonology. According to Sandler, the discovery that ASL has a phonology followed from Stokoe's (1960) demonstration that ASL signs "are created from a finite list of meaningless elements that combine and recombine" (2012, p. 162). This showed that signs "are not holistic pictorial gestures, as previously believed" (*ibid.*).

As Sandler sees it, and following Stokoe (1960), the question of whether ASL has a phonology, and associated phonemes, is at bottom a question of whether ASL makes use of "a finite list of meaningless elements that combine and recombine" that "are not holistic pictorial gestures" like icons or images. Research into this question returns a positive answer. The meaning of any arbitrary ASL sign is

a function of several simultaneous movements of the hands, arms, head, and upper body; these movements—like individual letters of a word—are meaningless on their own and individually reoccur as proper parts of many different signs that *are* meaningful, obeying specific rules governing their recombination (Sandler, 2012; Stokoe, 2005). The signs of ASL are not signs of English. But they are signs of a language nonetheless.

Assuming that ASL has a phonology in the sense just described, what are the phonemes? They are specific hand, arm, and bodily *movements*; the exact number used by ASL is a matter of debate. (See Volger and Metaxas (2004) for discussion). Comprehending someone's signing will, in the normal case, require *seeing* their movements; the phonemes of ASL will be visually represented. By the same token, it remains correct to say that the phonemes of English are sounds—where those sounds are certain vibrations in the air—and that ordinary speech perception involves *aurally representing* (or, simply, *hearing*) those sounds. Thus, the general notion of a phoneme may be modality-neutral, in the sense that it can be used to describe the most basic, meaningless units of language that combine and recombine to create the smallest meaningful units of that language. But that sort of neutrality does not apply to the phonemes of specific language themselves, which will always be physical events of a sort that can be perceived by an audience through the use of some sense modality or other. The upshot is that, even if we can speak of other languages as having phonemes that are not sounds, this is quite consistent with the phonemes of English being sounds, and with the inner speech of adults with ordinary speech and hearing having an auditory-phonological component.

However, an influential theory of speech perception—the motor theory of speech perception (Liberman & Mattingly, 1985)—might seem to cast doubt on the claim that the phonemes discriminated during ordinary speech perception are auditory in nature. According to the motor theory of speech perception, the proper objects of speech perception are not acoustic signals in the air but rather phonetic *gestures*, or *intentions to produce* phonetic gestures. Phonetic gestures are the bodily (lip, tongue, throat) movements needed to generate vocal utterances; and the intentions to produce such are the neural motor commands needed to generate the movements. These views of course accept that an acoustic signal is received and processed during speech perception; their point is that *what is perceived* is not the acoustic signal itself but rather the gestures that produce it. By analogy, one might hold that what is perceived in vision are ordinary objects like tables and chairs, and not photons—even if tables and chares are perceived *by means of* perceptually receiving photons. Streams of photons carry information about the tables and chairs we are able to see, just as (on the motor theory) vibrations of air carry information about the phonetic gestures (or intentions to produce such gestures) we are able to hear. In support of this idea they note that one and the same sound—measured in terms of its duration and frequency—can be heard as

distinct phonemes in different contexts, depending on the phonemes proceeding or following it. For instance, a short signal at 1440 Hz sounds like /p/ before the vowels /i/ ("ee") and /u/ ("oo"), but like /k/ before /a/ ('ah') (Liberman, Delattre, & Cooper, 1952). And two distinct sounds—measured in terms of frequency—can be heard as the same phoneme, depending on the sort of vowel following it.

This might lead one to think that it is the bodily movements that produce speech that are perceived when we perceive different phonemes, and not the raw sounds in the air from which information about phonemes is (perhaps) extracted. However, ambiguity follows to the level of the bodily movements themselves, as the very same mouth and throat motions can be involved in articulating distinct phonemes in different contexts. This is because multiple phonetic gestures influence a speaker's mouth and throat positions at any instant (a phenomenon known as "coarticulation") (Liberman & Mattingly, 1985). For instance, the /s/ at the beginning of 'sue' is generated through different lip movements than the /s/ at the start of 'see', as the lips anticipate the different vowels that follow each /s/ (Scott, McGettigan, & Eisner, 2009b). In this way, the vowel phonemes following each /s/ influence the bodily movements used to generate the /s/ itself. Thus, at any instant, the shape of the bodily articulators can be seen as expressing multiple phonemes simultaneously. This prevents any specific bodily movement or shape from being identified with a particular phoneme. As a result, Liberman & Mattingly (1985) propose that the actual objects of speech perception are the *intended* articulatory gestures—where this might be an "'upstream' neural command for the gesture from which the peripheral articulatory movements unfold" (1985, p. 9).

Yet granting, just for argument, that the objects of speech perception are not sounds—and are possibly even *neural* events—this does not touch the claim that inner speech (together with ordinary speech perception) has an auditory-phonological component. We simply have to distinguish between what is represented—the *object* of perception—and the manner in which it is represented, which is the *format* of representation. For instance, we can perceptually discriminate shapes both visually and tactually: a single cube might be the object of both a visual and tactual perceptual experience. Here the difference is a difference in perceptual format. (We can remain neutral on whether the difference in format can be reduced to a difference in the finer-grained properties of the cube that are represented in each case). *Even if* the objects of speech perception are not sounds, those objects may still be represented in an auditory format, just as a cube may be represented in either a visual or tactual format. For instance, we can represent relative spatial location both aurally—by the sound of someone's footsteps—or visually—by how far away they look to be. The objects of speech perception are represented in an auditory format, as it is (primarily) through our ears that we perceive speech. This is all we need to hold that inner speech has an auditory phonological component: it represents phonemes (whatever their nature) in an auditory format.

*3.4 Objection: Motor imagery allows us to judge the language to which our inner speech is keyed*

Speaking is a complex bodily action, requiring coordinated movements of the mouth, lips, tongue, throat and vocal chords. The movements needed to generate the specific phonemic sequences of a language are just as specific to the language as the phonemes they serve to produce (*modulo* the earlier points about coarticulation). If, during inner speech, we were somehow aware of those movements themselves, they would provide a means for distinguishing the language generated. Of course, for the most part, such movements don't actually occur during inner speech.[12] But then, neither are any sounds or phonemes generated. In the case of phonemes, we can say they are *represented by* the auditory-verbal imagery that partly comprises inner speech (or, alternatively, that inner speech has auditory-phonological phenomenal character, even if no phonemes are in fact instantiated during inner speech). Likewise, it could be that inner speech involves our generating *motor imagery* of the bodily movements needed to produce the utterances in question. Such imagery has in any case been thought to play a role in how overt speech is monitored (Postma, 2000; Postma & Noordanus, 1996; Tian & Poeppel, 2010, 2012). The question I want to consider is whether we are able to exploit motor imagery to determine the language to which our inner speech is keyed. If we are, then we needn't conclude that inner speech has an auditory phonological component that facilitates the discrimination.

We know that we have the ability to *aurally* discriminate a signal as being an utterance of a familiar language. That's what we do when we understand someone else's speech: we move from an aurally represented sound signal to a judgment about the words that were said. And we know why we have this ability: learning to understand speech just is gaining the capacity to distinguish which auditory signals are speech and, from among those, how they encode specific sequences of words. By contrast, there is no reason to think that we are able to discriminate languages through the use of proprioceptive or kinesthetic perception. In understanding overt speech, we are not faced with the task of distinguishing which of our own bodily movements are speech episodes and which are not. After all, it is one thing to use motor imagery in the prediction and monitoring of one's speech-motor actions, quite another to judge, on the basis of such imagery, the specific langue that is being generated. There is no cultural or evolutionary pressure on us to do the latter.

However, at this point we must consider a second important claim of the motor theory of speech perception. For the motor theory holds not only that the primary objects of speech perception are motoric events, but also that the speech production system is itself recruited for speech perception (e.g. Liberman

[12] To the extent that articulatory movements do occur during inner speech, the arguments below apply equally to them as to motor imagery of such movements.

*et al.*, 1967).[13] If the very same cognitive abilities required for speech production are exploited during speech perception, this would lend credence to the idea that we can and do exploit motor imagery in the service of perceiving and discriminating languages. For it suggests that speech perception involves translating an auditory input into a set of motor commands, or motor images, needed to produce the utterance. If this is indeed what speech perception amounts to, there would be very *good* reason to think that we *can* discriminate languages on the basis of the distinctive motor imagery needed to generate them.

However, there is little support for this aspect of the motor theory of speech perception and plenty of evidence against it (Galantucci, Fowler, & Turvey, 2006; Scott, McGettigan, & Eisner, 2009a). The evidence in favor consists primarily of the co-activation of motor cortex during speech perception and production (Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Schomers & Pulvermuller, 2016). Little can be concluded from such data, given that the same areas of motor cortex are activated during the perception of non-linguistic sounds as well (Hauk et al., 2006; Etzel et al, 2008). Further, neural responses in motor cortex during speech perception may simply reflect one's automatic preparation for responding to perceived speech, or play a role in regulating the timing of turn-taking during conversation (Scott et al., 2009a).

At the same time, dissociations observed in neuropsychology strongly suggest that the role that speech production areas play in speech perception is peripheral at best. For instance, people with productive aphasias, resulting from lesions in brain areas underlying speech production (such as Broca's area), can perform normally in tests of auditory speech comprehension, yet have severely impaired language production (Blank, Bird, Turkheimer, & Wise, 2003; Crinion, Warburton, Lambon-Ralph, Howard, & Wise, 2005). The opposite dissociation is observed in the fluent aphasias, where speech production is preserved—albeit with garbled syntax and semantics—while speech comprehension is impaired (Bogen & Bogen, 1976). Similar dissociations occur in development as well. Children with severely impoverished speech—a condition known as *dysarthia* or *anarthia*—can have entirely normal speech comprehension (Bishop, 2014; Bishop, Brown, & Robson, 1990). So, it appears unlikely that the activation of speech-motor production areas is necessary or sufficient for the comprehension of language. There is, then, little reason to think that we recognize the language to which our own inner speech is keyed by attending to motor commands or motor imagery.

[13] The nature of this recruitment varies depending on the iteration of the view one consults. In later work, Liberman & Mattingly hold that the speech production is "required for" speech perception only insofar as "adaptations of the motor system for controlling the organs of the vocal tract *took precedence* in the evolution of speech" (emphasis added). "A perceiving system, specialized to take account of the complex acoustic consequences, developed concomitantly" (1985, p. 7). This (mere) ontogenetic priority of speech production over perception does not implicate the very same neural or cognitive states in language production and comprehension.

In response to the question, "How do we distinguish the language to which our inner speech is keyed?" aurally-represented phonemes remain our only plausible answer.

## 4. Inserted thoughts, and the language in which they occur

In this final section, I want to discuss a consequence of the conclusion just reached for our understanding of two signature symptoms of schizophrenia: auditory verbal hallucinations (AVHs) and thought insertion. AVHs are typically defined as "sensory experience(s) which occur in the absence of corresponding external stimulation of the relevant sensory organ" (David, 2004), whereas thought insertion is diagnosed when "the subject experiences thoughts which are not his own intruding into his mind" (Wing *et al.*, 1983). Unlike AVHs, the diagnosis of thought insertion makes no explicit reference to states with sensory character. Nevertheless, my argument will be that, to the extent that an "inserted thought" seems to the patient to occur in a natural language, we can infer that it has an auditory-phonological component—even if the patient is unsure whether to describe it in those terms. This allows for some—and perhaps even most—inserted thoughts to be seen as a subset of AVHs, opening the door to a single strategy for explaining both symptoms.

### *4.1. AVHs, inserted thoughts, and patient reports*

While there is no consensus among theorists on how to understand the relationship between AVHs and thought insertion (Mullins & Spence, 2003), the distinction appears to have traction among patients. In a survey of 100 patients prone to AVHs, Nayani & David (1996) found that 46 reported experiencing a *distinct* phenomenon of thought insertion. At the same time, close examination of patient reports reveals a thick fog at the border of the two symptoms. First, many patients report *hearing voices* that lack any strong auditory or sensory component (Graham & Stephens, 2000; Junginger & Frame, 1985). Some AVHs are even described by patients as "soundless" or "inaudible" (Humpston & Broome, 2016; Larøi et al., 2012). More commonly, the reported sensory features of AVHs are pale in comparison to those associated with the ordinary perception of voices. For instance, in one survey, 37% of voice-hearers admitted that their voices "did not appear very real," and 52% said that the voices were "less loud than real voices" (Moritz & Larøi, 2008). In another study of 50 participants prone to AVHs, 70% reported that the voices they heard were not louder than their own "verbal thought"; only 26.5% reported that the voices always seemed to come from outside the head; and over 30% reported that their AVHs sometimes seemed to occur in their own tone of voice (Hoffman, Varanko, Gilmore, & Mishara, 2008). These episodes were all registered by experimenters as being *AVHs* and not cases of thought insertion. This means that *many* of the phenomena travelling under the name "Auditory Verbal Hallucination" are not all that much like hearing another person speak. They are more akin to one's own "verbal thought."

And if they are similar to that kind of *thought*, then we can expect them to be reported as thoughts—perhaps as *inserted thoughts*—by some.

Analogously, many cases of thought insertion are described by patients as taking the form of an alien or "inserted" *voice* (Gunn, 2016; Nayani & David, 1996). Clearly, the "internal voices" reported as inserted thoughts could be the same phenomenon that others describe as AVHs that are lacking in strong sensory features. A voice-like inserted thought and a verbal-thought-like AVH could be two sides of the same mountain. For that reason, it's natural to ask whether AVHs and inserted thoughts might indeed be the same phenomenon at bottom, reported in different ways. I've pursued that idea in earlier work (Langland-Hassan, 2008, 2016); and others have recently made the same kind of case (Badcock, 2016; Humpston & Broome, 2016). Here I want to reinforce—yet also partly temper—those arguments by suggesting a new way of adjudicating the question. But first a few words on why we should care if such an assimilation is possible.

*4.2 Sensorimotor accounts of agency*

In the many cases where an AVH or inserted thought is not all that much like hearing someone else speak, what could make it seem to be someone else's voice or thought? A first proposal might be that the episodes seem to result from another's agency just because they occur without one's intending them, or without one's being aware of any such intention. However, this approach founders when we recognize that ordinary experience is filled with such episodes—such as songs stuck in the head, or the many thoughts and images that enter our mind when we are trying to focus on something else. We are not aware of any intention to generate those episodes; yet, neither do they seem to result from *another's* agency.

A more subtle means has thus been proposed for explaining the disrupted sense of agency responsible for reports of inserted thoughts and AVHs. The general idea behind this approach is that such episodes result from the malfunctioning of cognitive mechanisms that normally aid us in distinguishing self-generated sensory changes from changes we perceive in our environment. We can call these *sensorimotor* views of agency, with an eye toward generality. The most common such views invoke a cognitive architecture involving *forward models* that serve to generate predictions of the sensory input that will result from carrying out certain motor commands (Miall, Weir, Wolpert, & Stein, 1993; M. Perrone-Bertolotti et al., 2014; Wolpert, Miall, & Kawato, 1998). Yet much the same function of distinguishing self from other-caused changes in sensation is served by alternative mechanisms within the *Predictive Processing Framework* (Hohwy, 2013; Wilkinson, 2014) and *auditory processing stream frameworks* (Badcock, 2016), each of which has been extended to explain the unusual phenomenology of both AVHs and inserted thoughts.

I won't review the evidence in favor these approaches here, nor respond to challenges one might raise as they are applied to inner speech. (See Swiney (this volume) and Loevenbruk *et al.* (this volume) for extended discussion on those scores.) I want instead to focus on an assumption of all sensorimotor accounts of agency as they are applied to AVHs and inserted thoughts, which is that the aberrant mental episode in question—typically thought to be episode of inner speech—has sensory character, and is therefore a plausible result of malfunctioning sensorimotor mechanisms. For while sensorimotor approaches don't assume that AVHs and inserted thoughts always have strong sensory features, they do require that the episodes arise out of general mechanisms involved in sensorimotor control. And some of the phenomenological descriptions reviewed above—where inserted thoughts are described as "soundless" or "inaudible"—appear to call that assumption into question. Moved by reports of seemingly "amodal" voices and inserted thoughts, a number of theorists have argued that we will only properly face up to the challenge of explaining these phenomena when we admit that they lack sensory character and therefore cannot result from aberrant sensorimotor processes (Graham & Stephens, 2000; Vosgerau & Newen, 2007).

What can be said in defense of sensorimotor accounts of inserted thoughts and "soundless" voices? It is here that inner speech's auditory-phonological component—and our means for knowing about it—become relevant. My question in earlier sections was: how is it that we are able to distinguish the particular language to which our inner speech is keyed? I argued that it must be by exploiting inner speech's auditory-phonological component. This raises an interesting question with respect to voices with diminished, or no reported sensory features. Presumably these voices seem to their subjects to occur in a spoken language. Why else describe them as voices? But then, if the voices seem to be speaking a specific language, it must be that they have auditory-phonological features distinctive of that language. This follows from the arguments of Sections 2 and 3, above. And the episodes must have those sensory features *whether or not patients find it proper to describe them in auditory or sensory terms*. In short, if a "soundless" voice seems to be speaking English, and not French, we can infer that it has an auditory phonological component. The upshot for sensorimotor theories of AVHs and inserted thoughts is that they can be indeed be extended to explain episodes with weak, or even no explicitly reported sensory phenomenology, so long as the patients are confident that the episodes seem to occur in a specific language.

### *4.3 A proposal for new diagnostic questions*

This raises the question: *do* the episodes reported as inserted thoughts seem, to their reporters, to occur in a specific language? We know that many do, where the "inserted" thought is described as a kind

of message-conveying voice. Yet not all reports of inserted thoughts describe them as voices. For instance, some of Nayani & David's participants—those who described their AVHs as emanating from *within* the head—distinguished those AVH experiences from inserted thoughts, which they characterized as "bad impulses or unpleasant visual images" (*ibid.*). And in a famous report from Mellor (1970), a patient describes his inserted thoughts as *picture*-like: "The thoughts of Eamonn Andrews come into my mind…There are no other thoughts there, only his…He treats my mind like a screen and flashes thoughts onto it like you flash a picture."

My conclusion, then, takes the form of a conditional statement and a recommendation for future diagnostic practice. *If* a patient reports that an AVH or inserted thought seemed to occur in a specific natural language, then we can be confident that the episode had an auditory-phonological component—*even if* the patient does not describe it in those terms.[14] It will then be the sort of thing that sensorimotor accounts of agency are well-placed to explain (provided that the accounts are otherwise well-founded). On the other hand, if the episode does not seem to the subject to occur in a specific language, more caution is required. It is possible that it still has sensory character—perhaps related to vision, or another modality—and so remains within the purview of sensorimotor accounts. Nevertheless, it is also possible that the state altogether lacks sensory character and has no essential tie to sensation or perception. In that case, we would indeed be in the situation envisioned by Graham & Stephens (2000), and Vosgerau & Newen (2007); we would need an account of disrupted agency that did not appeal sensorimotor mechanisms or states with sensory character. (Though see Vicente & Jorba (*under review*) for an articulation of how amodal "inserted thoughts" might still be explained within a comparator framework).

A limitation of current psychiatric questionnaires is that they don't help us to distinguish among these possibilities. Patients are diagnosed with thought insertion when they affirmatively answer questions such as: "Do there ever seem to be thoughts in your mind which are not your own, which seem to come from elsewhere?" (SCAN[15], 18.006), and "Do you ever feel as though the thoughts in your head are not your own?" (CAPE[16], #26). Positive answers here don't by themselves allow inferences as to whether the episode seemed to occur in a natural language. But it would be easy enough to add more precise questions to such assessments. If patients answer "yes" to the questions just listed, a follow-up

[14] While I have limited my conclusions throughout to those with ordinary language production and comprehension, it bears noting that a corresponding distinction between AVH and thought insertion diagnostics can be proposed for deaf signers as well. Insofar as deaf signers report that the "voices" they see (or hear, or feel) occur in ASL (or some other sign language), we can be confident that the episodes have a sensory-phonological component that warrants distinguishing them from inserted thoughts. Existing research on the phenomenological features of the hallucinations experienced by congenitally and pre-linguistically deaf people with schizophrenia paints a complicated and at times conflicting picture of their sensory characteristics (Atkinson, 2006; DuFeu & McKenna, 1999; Schonauer, Achtergarde, Gotthardt, & Folkerts, 1998).

[15] Schedules for Clinical Assessment in Neuropsychiatry (Wing et al., 1990).

[16] The Community Assessment of Psychic Experience (Stefanis et al., 2002).

could be: "Do the thoughts seem to occur in words? If so, are they in a particular language?" Positive answers to *these* questions would provide some justification for assimilating the episode to the phenomenon of "hearing voices" in general, and open the door to sensorimotor approaches for understanding their etiology. Negative answers would warrant consideration of other possibilities.

Once we have assimilated "inserted thoughts" that seem to occur in a natural language to the class of AVHs, the term 'thought insertion' can be usefully repurposed to apply only to episodes do *not* seem to occur in a language. For then we can have confidence that the phenomenon will not admit of the same style of explanation as AVHs. Some reports of this nature were discussed above. Yet it remains an open question whether thought insertion, *so redefined,* is a robust symptom in schizophrenia. Once reports of inserted thoughts that seem, to those reporting them, to occur in a language are subtracted from the complete set of thought insertion reports, there may not remain a widespread phenomenon of thought insertion to be explained. Researchers currently lack the data requisite to saying whether that is the case. We do, however, know how to go about gathering it.

## References


Atkinson, J. R. (2006). The Perceptual Characteristics of Voice-Hallucinations in Deaf People: Insights into the Nature of Subvocal Thought and Sensory Feedback Loops. *Schizophrenia Bulletin, 32*(4), 701-708. doi:10.1093/schbul/sbj063

Badcock, J. C. (2016). A Neuropsychological Approach to Auditory Verbal Hallucinations and Thought Insertion - Grounded in Normal Voice Perception. *Review of Philosophy and Psychology, 7*(3), 631-652. doi:10.1007/s13164-015-0270-3

Berndt, R. S., D'autrechy, C. L., & Reggia, J. A. (1994). Functional pronunciation units in English words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(4), 977.

Besner, D., & Davelaar, E. (1983). Suedohomofoan effects in visual word recognition: Evidence for phonological processing. *Canadian Journal of Psychology/Revue canadienne de psychologie, 37*(2), 300-305. doi:10.1037/h0080719

Bishop, D. (2014). *Uncommon Understanding (Classic Edition): Development and disorders of language comprehension in children*: Psychology Press.

Bishop, D., Brown, B. B., & Robson, J. (1990). The Relationship Between Phoneme Discrimination, Speech Production, and Language Comprehension in Cerebral-Palsied Individuals. *Journal of Speech, Language, and Hearing Research, 33*(2), 210-219. doi:10.1044/jshr.3302.210

Blank, S. C., Bird, H., Turkheimer, F., & Wise, R. J. (2003). Speech production after stroke: the role of the right pars opercularis. *Ann Neurol, 54*(3), 310-320. doi:10.1002/ana.10656
Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. Handbook of psycholinguistics, ed. by Morton Ann Gernsbacher, 945-84: San Diego: Academic Press.
Bogen, J. E., & Bogen, G. M. (1976). Wernicke's region--Where is it? *Ann N Y Acad Sci, 280*, 834-843.
Carruthers, P. (this volume). The Causes and Contents of Inner Speech. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: Nature and Functions*. Oxford: Oxford University Press.
Cassam, Q. (2011). Knowing What You Believe. *Proceedings of the Aristotelian Society, 111*(1pt1), 1-23.
Clark, A. (1998). Magic words: how language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and Thought: Interdisciplinary Themes* (pp. 162-183). Cambridge: Cambridge University Press.
Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon.
Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology, 55*, 429-432.
Crinion, J. T., Warburton, E. A., Lambon-Ralph, M. A., Howard, D., & Wise, R. J. S. (2005). Listening to Narrative Speech after Aphasic Stroke: the Role of the Left Anterior Temporal Lobe. *Cerebral Cortex, 16*(8), 1116-1125. doi:10.1093/cercor/bhj053
DuFeu, M., & McKenna, P. J. (1999). Prelingually profoundly deaf schizophrenic patients who hear voices: a phenomenological analysis. *Acta Psychiatrica Scandinavica, 99*(6), 453-459. doi:10.1111/j.1600-0447.1999.tb00992.x
Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *Eur J Neurosci, 15*(2), 399-402.
Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review, 13*(3), 361-377.
Ganis, G., Thompson, W. L., & Kosslyn, S. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cognitive Brain Research, 20*, 226-241.
Gauker, C. (2011). *Words and Images: An Essay on the Origin of Ideas*. Oxford: Oxford University Press.
Gauker, C. (this volume). Inner Speech as the Internalization of Outer Speech. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: Nature and Functions*. Oxford: Oxford University Press.
Geva, S., Bennett, S., Warburton, E. A., & Patterson, K. (2011). Discrepancy between inner and overt speech: Implications for post-stroke aphasia and normal language processing. *Aphasiology, 25*(3), 323-343. doi:10.1080/02687038.2010.511236
Graham, G., & Stephens, G. L. (2000). *When Self-Consciousness Breaks*. Cambridge, MA: MIT Press.
Grice, H. P. (1957). Meaning. *The Philosophical Review*, 377-388.
Gunn, R. (2016). On Thought Insertion. *Review of Philosophy and Psychology, 7*(3), 559-575. doi:10.1007/s13164-015-0271-2
Hoffman, R., Varanko, M., Gilmore, J., & Mishara, A. L. (2008). Experiential features used by patients with schizophrenia to differentiate 'voices' from ordinary verbal thought. *Psychological Medicine, 38*, 1167-1176.
Hohwy, J. (2013). The Predictive Mind. Oxford: Oxford University Press.
Humpston, C. S., & Broome, M. R. (2016). The Spectra of Soundless Voices and Audible Thoughts: Towards an Integrative Model of Auditory Verbal Hallucinations and Thought Insertion. *Review of Philosophy and Psychology, 7*(3), 611-629. doi:10.1007/s13164-015-0232-9
Hurlburt, R. T., Heavey, C. L., & Kelsey, J. M. (2013). Toward a phenomenology of inner speaking. *Consciousness and Cognition, 22*(4), 1477-1494. doi:10.1016/J.CONCOG.2013.10.003
Hurlburt, R. T., & Schwitzgebel, E. (2007). *Describing Inner Experience? Proponent meets Skeptic*. Cambridge, MA: MIT Press.
Jackendoff, R. (1996). How language helps us think. *Pragmatics and Cognition, 4*(1), 1-34.

Johnston, R. S., & Conning, A. (1990). The effects of overt and covert rehearsal on the emergence of the phonological similarity effect in 5-year-old children. *British Journal of Developmental Psychology, 8*(4), 411-418.
Johnston, R. S., Johnson, C., & Gray, C. (1987). The emergence of the word length effect in young children: The effects of overt and covert rehearsal. *British Journal of Developmental Psychology, 5*(3), 243-248.
Junginger, J., & Frame, C. L. (1985). Self-Report of the Frequency and Phenomenology of Verbal Hallucinations. *The Journal of Nervous and Mental Disease, 173*(3), 149-155.
Kell, C. A., Darquea, M., Behrens, M., Cordani, L., Keller, C., & Fuchs, S. (2017). Phonetic detail and lateralization of reading-related inner speech and of auditory and somatosensory feedback processing during overt reading. *Hum Brain Mapp, 38*(1), 493-508. doi:10.1002/hbm.23398
Kosslyn, S., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J. P., Thompson, W. L., . . . Alpert, N. M. (1999). The role of area 17 in visual imagery: Convergent evidence from PET and rTMS. *Science, 2*(5411), 167-170.
Langland-Hassan, P. (2008). Fractured Phenomenologies: Thought Insertion, Inner Speech, and the Puzzle of Extraneity. *Mind & Language, 23*(4), 369-401.
Langland-Hassan, P. (2014). Inner Speech and Metacognition: In Search of a Connection. *Mind and Language, 29*(5), 511-533.
Langland-Hassan, P. (2016). Hearing a Voice as one's own: Two Views of Inner Speech Self-Monitoring Deficits in Schizophrenia. *Review of Philosophy and Psychology, 7*(3), 675-699. doi:10.1007/s13164-015-0250-7
Langland-Hassan, P., Faries, F., Richardson, M., & Dietz, A. (2015). Inner Speech Deficits in People with Aphasia. *Frontiers in Psychology, 6*, 528.
Larøi, F., Sommer, I. E., Blom, J. D., Fernyhough, C., ffytche, D. H., Hugdahl, K., . . . Waters, F. (2012). The Characteristic Features of Auditory Verbal Hallucinations in Clinical and Nonclinical Groups: State-of-the-Art Overview and Future Directions. *Schizophrenia Bulletin, 38*(4), 724-733. doi:10.1093/schbul/sbs061
Levelt, W. J. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *Am J Psychol, 65*(4), 497-516.
Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1-36.
Mellor, C. H. (1970). First rank symptoms of schizophrenia. *British Journal of Psychology, 117*, 15-23.
Miall, R. C., Weir, D. J., Wolpert, D. M., & Stein, R. C. (1993). Is the cerebellum a Smith Predictor? *Journal of Motor Behavior, 25*, 203-216.
Moritz, S., & Larøi, F. (2008). Differences and similarities in the sensory and cognitive signatures of voice-hearing, intrusions and thoughts. *Schizophrenia Research, 102*(1–3), 96-107. doi:http://dx.doi.org/10.1016/j.schres.2008.04.007
Mullins, S., & Spence, S. A. (2003). Re-examining thought insertion. *The British Journal of Psychiatry, 182*(4), 293.
Nayani, T. H., & David, A. (1996). The auditory hallucination: a phenomenological survey. *Psychological Medicine, 26*, 177-189.
Perrone-Bertolotti, M., Kujala, J., Vidal, J. R., Hamame, C. M., Ossandon, T., Bertrand, O., . . . Lachaux, J.-P. (2012). How Silent Is Silent Reading? Intracerebral Evidence for Top-Down Activation of Temporal Voice Areas during Reading. *The Journal of Neuroscience, 32*(49), 17554-17562. doi:10.1523/jneurosci.2982-12.2012
Perrone-Bertolotti, M., Rapin, L., Lachaux, J. P., Baciu, M., & Loevenbruck, H. (2014). What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behav Brain Res, 261*, 220-239. doi:10.1016/j.bbr.2013.12.034
Postma, A. (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition, 77*(2), 97-132.

Postma, A., & Noordanus, C. (1996). Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Language and Speech, 39*, 375-392.
Rey, A., Ziegler, J. C., & Jacobs, A. M. (2000). Graphemes are perceptual reading units. *Cognition, 75*(1), B1-12.
Sandler, W. (2012). The phonological organization of sign languages. *Language and linguistics compass, 6*(3), 162-182.
Schomers, M. R., & Pulvermuller, F. (2016). Is the Sensorimotor Cortex Relevant for Speech Perception and Understanding? An Integrative Review. *Front Hum Neurosci, 10*, 435. doi:10.3389/fnhum.2016.00435
Schonauer, K., Achtergarde, D., Gotthardt, U., & Folkerts, H. (1998). Hallucinatory modalities in prelingually deaf schizophrenic patients: a retrospective analysis of 67 cases. *Acta Psychiatrica Scandinavica, 98*(5), 377-383.
Scott, S. K., McGettigan, C., & Eisner, F. (2009a). A little more conversation, a little less action--candidate roles for the motor cortex in speech perception. *Nat Rev Neurosci, 10*(4), 295-302. doi:10.1038/nrn2603
Scott, S. K., McGettigan, C., & Eisner, F. (2009b). A little more conversation, a little less action [mdash] candidate roles for the motor cortex in speech perception. *Nat Rev Neurosci, 10*(4), 295-302.
Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge, England: Cambridge University Press.
Slotnick, S., Thompson, W., & Kosslyn, S. (2005). Visual mental imagery induces retinotopically organized activation of early visual areas. *Cerebral Cortex, 15*, 1570-1583.
Stefanis, N., Hanssen, M., Smirnis, N., Avramopoulos, D., Evdokimidis, I., Stefanis, C., . . . Van Os, J. (2002). Evidence that three dimensions of psychosis have a distribution in the general population. *Psychological Medicine, 32*(02), 347-358.
Stokoe, W. C. (1960). *Sign language structure: An outline of the visula communication systems of the American deaf*. Buffalo, NY: University of Buffalo.
Stokoe, W. C. (2005). Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *The Journal of Deaf Studies and Deaf Education, 10*(1), 3-37. doi:10.1093/deafed/eni001
Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology, 1*(166). doi:10.3389/fpsyg.2010.00166
Tian, X., & Poeppel, D. (2012). Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Front Hum Neurosci, 6*, 314. doi:10.3389/fnhum.2012.00314
Vogler, C., & Metaxas, D. (2004). Handshapes and Movements: Multiple-Channel American Sign Language Recognition. In A. Camurri & G. Volpe (Eds.), *Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop, GW 2003, Genova, Italy, April 15-17, 2003, Selected Revised Papers* (pp. 247-258). Berlin, Heidelberg: Springer Berlin Heidelberg.
Vosgerau, G., & Newen, A. (2007). Thoughts, motor actions, and the self. *Mind and Language, 22*, 22-43.
Wilkinson, S. (2014). Accounting for the phenomenology and varieties of auditory verbal hallucination within a predictive processing framework. *Consciousness and Cognition, 30*, 142-155.
Wing, J., Babor, T., Brugha, T., Burke, J., Cooper, J., Giel, R., . . . Sartorius, N. (1990). SCAN. Schedules for Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry, 47*(6), 589.
Wolpert, D. M., Miall, R. C., & Kawato, M. (1998). Internal Models in the cerebellum. *TRENDS in Cognitive Science, 2*, 338-347.
Wu, W. (2012). Explaining Schizophrenia: Auditory Verbal Hallucination and Self-Monitoring. *Mind and Language, 27*(1), 86-107.
Yao, B., Belin, P., & Scheepers, C. (2011). Silent Reading of Direct versus Indirect Speech Activates Voice-selective Areas in the Auditory Cortex. *Journal of Cognitive Neuroscience, 23*(10), 3146-3152. doi:10.1162/jocn_a_00022