From Speech to Voice: On the Content of Inner Speech

Abstract: Theorists have found it difficult to reconcile the unity of inner speech as a mental state kind with the diversity of its manifestations. I argue that existing views concerning the content of inner speech fail to accommodate both of these features because they mistakenly assume that its content is to be found in the 'speech processing hierarchy', which includes semantic, syntactic, phonemic, phonetic, and articulatory levels. Upon rejecting this assumption, I offer a position on which the content of inner speech is determined by voice processing, of which speech processing is but one component. The resulting view does justice to the idea that inner speech is a motley assortment of episodes that nevertheless form a kind.

Keywords: inner speech; speech processing; voice processing; occurrent thought

Word count: 11,382 words

1. Introduction

What goes on when we think? James Joyce comes pretty close to capturing it in his *Ulysses*. Take a peek inside 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)

Like Molly, many of us often think in words, or engage in what philosophers and psychologists have come to call "inner speech".[1] Given his medium, Joyce is forced to present inner speech as a uniform phenomenon: Molly is presented as thinking in strings of words pronounced in her head. However, as I suspect Joyce would attest, our inner speech is not nearly so neat. As I am typing this sentence, I silently move my mouth in unison; but a few words in, my lips now pressed together, I find whole words popping into my head without any corresponding motor sensation; soon enough, I have an auditory experience, as I return to a word to hear it in my head just as I would have heard

[1] Although see Hurlburt and Heavey (2018) for skepticism about reported frequencies of inner speech.

it aloud; and finally, stepping back to consider the whole of what I have just written, I affirm the bare thought to some generalized other, but without an auditory or linguistic garb. Inner speech seems to be more a shape-shifting menagerie, taking on different forms as it unfolds, than a consistent march of words pronounced in the head.

I will argue that this diversity tells against a standard picture of the content of inner speech. The standard picture starts with a widely supported view of speech production as a hierarchical process (see, e.g., Levelt, 1993). On this view, 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. To this widely supported view of speech production, the standard picture adds the idea that inner speech is just truncated speech production: we start with the selection of a proposition and move down the hierarchy, but at some point processing is cut short. The idea that inner speech is truncated speech production is implicit in many philosophical and psychological discussions of inner speech:

> 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)

Thus, according to the standard picture, the contents of inner speech are those which are implicated in the lead-up to speech production, minus the contents that have been truncated.

Here I will argue against this standard picture of inner speech. All of the dominant positions regarding the content of inner speech – *concretism*, *abstractionism*, and *pluralism* – presuppose the standard picture. In Section 2, I characterize these three views. In Section 3, I provide theoretical challenges to Peter Langland-Hassan's argument for concretism and Christopher Gauker's version of abstractionism. This leaves pluralism as the most plausible view about the content of inner speech. However, in Section 4, I argue against Agustín Vicente and Fernando Martínez-Manrique's pluralist model by showing that it fails to treat inner speech as a kind of mental state. In Section 5, I show that speech processing is only one component of a number of processes targeting voice. In light of this broader view, in Section 6, I present an alternative position, *vocalism*, according to which the content of inner speech is vocal: during an inner speech episode, one represents a voice communicating information. In Section 7, I close by summarizing how vocalism captures at once the diversity and the unity of inner speech.

2. A Menu of Views

There are three assumptions that will ground my discussion.

First, my discussion of the nature of inner speech can be framed either in terms of the representational content of inner speech or the representational vehicle of inner speech. I adopt the former framing here, in part because a number of previous discussions have been pitched in those terms (e.g., Langland-Hassan, 2014). However, one can instead focus on representational vehicles without loss to the central argument of the paper. Second, my discussion will focus on a particular kind of diversity observed in inner speech: diversity regarding the representational content of inner speech. Inner speech is also diverse in regards to its functions, which include reading comprehension, planning, and rehearsal, among others. Moreover, it has been suggested that differences in the representational content of inner speech interact with differences in function (Alderson-Day and Fernyhough, 2015). However, nothing will be lost by focusing on representational content alone, and in fact, we can gain a clearer understanding of interaction effects by first getting a clear understanding of the representational content of inner speech. Third, my discussion will focus on models of speech production as found in modern computational psycholinguistics (e.g., models stemming from Fromkin, 1971). Although the Vygotskyean tradition presents a complementary research program, emphasizing the development of inner speech from childhood into adulthood (Jones, 2009), I focus on modern computational models because the processes represented by these models are what most immediately endow inner speech with representational content.

With these assumptions in hand, let us turn now to the speech processing hierarchy and the three views of the content of inner speech that map onto it. 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] The nature of the speech processing hierarchy remains contested. Psycholinguists disagree about how information flows through the hierarchy – serially or in parallel, feedforward or feedback – and about the exact operations and sub-operations within the hierarchy (e.g., Fromkin, 1971; Dell, 1986). Despite these differences, psycholinguists tend to agree on the organization presented in Figure 1.

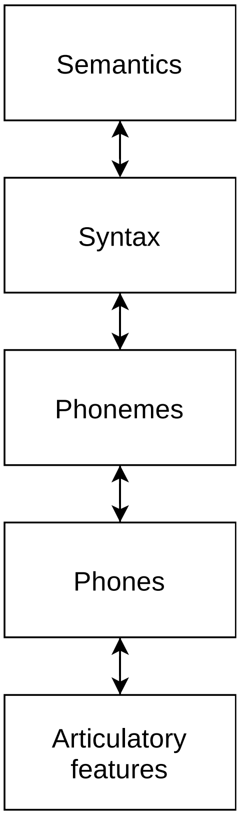

Fig. 1: The speech processing hierarchy

I will understand the topmost level of the hierarchy – 'Semantics' – as selecting 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. Phonemes are 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 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.

Each level of the speech processing hierarchy is thus associated with a particular type of content (or vehicle; see above). Perhaps the most intuitive view about the content of inner speech is that it is phonetic or involves speech sounds. According to this view, which I will label *concretism*, inner speech episodes represent speech sounds (e.g., Langland-Hassan, 2018). Though naïve introspection seems to reveal that inner speech is phonetic, others 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 do not represent speech sounds (e.g., Gauker, 2018). In contrast to both of these positions, some authors have been open to the possibility that the contents of inner speech are variable: during default contexts, inner speech does not represent phonetic content, but during "stress and cognitive challenge," inner speech does represent such content (Alderson-Day and Fernyhough, 2015, p. 933). According to *pluralism*, inner speech episodes have speech sound content in some contexts, but not in others (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 frames the debate over the content of inner speech. The debate concerns whether inner speech engages

[3] Although see Langland-Hassan (2018) for a contrasting position on phonemes, according to which phonemes are auditory.

phonetic contents (concretism versus abstractionism) and whether the views are exclusive (pluralism). The aim of this paper is not to adjudicate the debate between concretism, abstractionism, and pluralism. Rather, my aim is to reject an assumption that serves as common ground for the debate: that the content of inner speech is exhaustively derivable from the speech processing hierarchy.

3. Assessing Concretism and Abstractionism

Empirical evidence tends to be equivocal regarding support for concretism and abstractionism. For this reason, the most promising arguments for these positions tend to be philosophical in character. I first assess Peter Langland-Hassan's argument for concretism, and then turn to Christopher Gauker's defense of an abstractionist view of inner speech.[4]

3.1 Against Langland-Hassan's Concretism

Peter Langland-Hassan (2018) has argued that inner speech always has an "auditory-phonological" or speech sound component.[5] Although this view is often taken as a "truism, a platitude of common sense", Langland-Hassan also seeks to provide an argument in its favor (p.

[4] Langland-Hassan (2018) and Gauker (2018) differ in their framing of concretism and abstractionism. Langland-Hassan seems to assume that inner speech always represents phonetic *content*, while Gauker seems to assume that inner speech never has a phonetic *vehicle*. This difference will not matter in my discussion of the views. For this reason, I will use 'phonetic/auditory/speech sound component' with the understanding that it translates as 'phonetic/auditory/speech sound content *or* vehicle'.

[5] Recall that Langland-Hassan (2018) believes that phonemes are auditory (see footnote 3). Although I have denied this (see Section 2), for the sake of the present argument, I will use 'phonological' in the sense that Langland-Hassan intends.

78). 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. 82).

Langland-Hassan runs through four possible 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. In this context, 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. 82). 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, French, Spanish, 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 on 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 a visual phenomenon. This leaves only one possibility: there must be an auditory component of inner speech that accounts for our knowledge of the language of our inner speech. In effect, I know that my inner speech is in English, according to Langland-Hassan, because I am representing the speech sounds of English (i.e., it sounds like English).

Although Langland-Hassan presents a plausible case against semantics and graphemes, he is mistaken in thinking that syntax cannot ground our knowledge of the language of inner speech. Langland-Hassan's discussion of syntax seems to stem from W.J.M. Levelt's classic model of word retrieval, which has been echoed in a number of more recent models (Levelt, 1993). Levelt distinguishes between two types of representation of a word: a lemma and a lexeme. A 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 lexemes, and so are pre-phonological/auditory. Although Langland-Hassan fails to mention the lemma/lexeme distinction, for his argument to go through, he would need to argue that lemmas represent only the semantic and syntactic structure of a word, but not the *word* whose semantic and syntactic structure it is. For example, the lemma for *cake*, according to Langland-Hassan, represents its referent (semantics) and that it is a noun (syntax), but does not represent the identity of the word whose semantic and syntactic properties are in question – the word *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 for Langland-Hassan is that this view of lemmas is contradicted by existing psycholinguistic work (for evidence see Roelofs, Meyer, and Levelt (1998) and Jescheniak, Meyer, and Levelt (2003)). 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. Thus, even if the lemmas for *give* (English) and *dar* (Spanish) are identical in terms of their semantic and syntactic features, the lemmas will differ with regard to the non-phonological word (*give* vs. *dar*) they contain. 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, it seems that pre-phonological/auditory words provide a possible ground for knowing the language of one's inner speech.

A second reason to doubt Langland-Hassan's inference to the best explanation is that there is nothing in his 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. The conclusion is misguided because I can know that I am speaking an English sentence even if my ears are completely plugged, my facial bones are unable to conduct energy, and my vocal apparatus is numbed. (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 one.) In this context, I would know the language of my outer speech independent of observation of kinesthetic or auditory properties associated with my outer speech. If Langland-Hassan's argument seems suspect when applied to outer speech, I see no reason it should be compelling for inner speech. The lesson is that I know the language of my inner speech independent of observation of kinesthetic or auditory properties associated with inner speech.[6] I therefore conclude, on both psycholinguistic and philosophical grounds, that Langland-Hassan's argument fails to show that an auditory component is always present in inner speech.

3.2 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 a phonetic component. 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. The view is motivated by an analogy between inner and outer speech: just as we should distinguish between outer speech and the auditory experience of outer speech, so too, according to Gauker, we should distinguish between inner speech and the auditory imagery of inner speech. On Gauker's view, the auditory imagery that represents inner speech possesses an auditory component, but inner speech itself never possesses an auditory component.

[6] This line of argument puts into relief a plausible alternative explanation of how I know 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 (2000)). On this account, I know that my inner speech is in English because I *use* English words in my inner speech, where this knowledge is not grounded in observation. Although Langland-Hassan seems to assume that the knowledge of the language of our inner speech is gained by introspection, the alternative I have mentioned rejects this restriction.

There are three important parts of Gauker's "perception theory of the auditory imagery of inner speech".

First, Gauker thinks that there is a representational relation between auditory imagery and inner speech. According to Gauker, just as speech perception represents speech sounds, so too auditory imagery represents inner speech. The difference, on Gauker's view, is that speech perception accurately represents speech sounds, whereas auditory imagery systematically misrepresents inner speech as involving the occurrence of sounds. This misrepresentation is responsible for our impression that inner speech itself is auditory. Second, according to Gauker, just as speech perception is caused by the actual presence of speech sounds, so too the auditory imagery that systematically misrepresents inner speech is caused by the actual presence of inner speech. Finally, Gauker believes that there is continuity between the perception of outer speech and the perception of inner speech: the same type of experience – the experience of speech sounds – that is caused by the presence of outer speech is also caused by the presence of inner speech. As Gauker claims:

> the experience of sound can be caused either by sound waves striking the ear drums or by an episode of inner speech. In fact, the only reason to say that the experience of inner speech misrepresents inner speech as sounds is that we regard outer speech as the paradigmatic cause of that kind of experience. (p. 59)

Though the experience of speech sounds is paradigmatically caused by outer speech, the experience is also caused by inner speech, according to Gauker. When the experience is caused by inner speech, it is a misrepresentation of inner speech as involving speech sounds.

There are both empirical and theoretical problems with Gauker's brand of abstractionism.

Gauker's fundamental claim is that inner speech itself does not possess a phonetic component. However, if it can be shown that there are cases in which phonetic inner speech persists despite a subject's inability to perceive inner speech, then this will suggest, *contra* Gauker, that inner speech itself possesses a phonetic component. People with pure word deafness or auditory agnosia show that there are such cases. People with pure word deafness are able to perceive environmental sounds, but unable to perceive speech; to them, speech sounds like mumbling, noise, or a foreign language. Despite this deficit in speech perception, a number of case studies show that people with pure word deafness seem to have intact phonetic inner speech. Marshall et al. (1985) tested for auditory inner speech in a patient with auditory agnosia, and found that the patient was able to silently judge which of a set of written words rhymed with a target written word. Success at a rhyme task is plausibly taken to require the use of inner speech with a phonetic component: the subject must employ speech sound representations and compare them in working memory to figure out if they rhyme (Langland-Hassan, 2014). This coupling of a deficit in speech perception with intact phonetic inner speech has been replicated in a number of other case studies (see Denes and Semenza, 1975; Buchtel and Stewart 1989; Papathanasiou et al., 1998). This evidence suggests that the phonetic component associated with inner speech cannot, *contra* Gauker, stem from the experience of speech sounds.

How might one account for auditory inner speech in people with pure word deafness and auditory agnosia? The key is to see that there are at least *two* token representations of auditory features: one present in a *perception* stream, which is employed for the perception of speech sounds and which is impaired in those with pure word deafness, and the other present in a *production* stream, which is intact in those with pure word deafness. This structure would account for the presence of intact phonetic inner speech despite deficits in speech perception: inner speech exploits representations of auditory features in the production stream, not in the perception stream. In fact, Levelt's model of speech control reflects a structure of just this kind. According to Levelt, in the course of speech production a phonetic plan is generated by the speech production module, which serves as input to the speech perception module, which assesses the phonetic plan for errors (p.

470).[7] Levelt identifies inner speech with the phonetic plan as it arises in the *speech production module* – independent of perceptual, error-detection processes. The structure of Levelt's speech control model helps explain why people with pure word deafness are able to test for rhymes in inner speech: their inner speech possesses a phonetic component that can be used in working memory independent of breakdowns in perceptual processes.

Hence, there are two parts to my criticism of Gauker's brand of abstractionism. First, the existence of patients with pure word deafness but intact phonetic inner speech presents a counter-example to his view. Second, there is an explanation of the dissociation that is independently motivated – motivated by considerations involving speech control – that claims that the phonetic component associated with inner speech is part of a production process, not a perceptual process. Even if the reader disagrees with my explanation of the dissociation, the dissociation itself is sufficient to undermine Gauker's version of abstractionism. The phonetic component associated with inner speech is part of inner speech itself, not a perceptual representation of inner speech.

4. Against Pluralist Accounts

Given the challenges facing monist views of the content of inner speech, a number of "pluralist" positions have been offered. According to these theories, in some contexts, inner speech represents speech sounds, while in others, it does not. For example, in their recent review of the literature, Alderson-Day and Fernyhough (2015) conclude that "the core of inner speech is

[7] Gauker might reject Levelt's speech control model, since Gauker may take it to depend on the doubtful existence of a language of thought. However, Levelt's speech control model does not depend on the existence of a language of thought, since the control model is a model of *peripheral* processes of speech production, e.g., motor control processes, and not *core* processes, e.g., concept selection.

[an] abstract code containing a combination of semantic, syntactic, and phonological information" that can sometimes be "unpacked" in terms of representations of speech sounds and articulation (p. 950). Oppenheim and Dell (2010) hold a similar position, arguing that 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).

Although most versions of pluralism emphasize phonemic, phonetic, and motoric forms of inner speech, there seem to be propositional and syntactic forms of inner speech as well. Using Descriptive Experience Sampling, Hurlburt et al. (2013) report that at times inner speech seems to be "missing all of its words" (p. 1483). According to Hurlburt and Heavey (2018):

> We have also seen (infrequent) instances where the experience is of speaking yet no words at all are experienced…Hurlburt, Heavey, and Kelsey (2013) call that unworded inner speaking: there is the experience of speaking, but there are no experienced words. Unworded inner speaking is distinct from unsymbolized thinking: unsymbolized thinking does not involve any experiences of speaking…" (p. 182)

This so-called 'unworded' inner speech seems to be propositional, not word-like. In addition to 'unworded' inner speech, at 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. 58)

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 also seems to 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). Thus, alongside the often-reported auditory and articulatory forms of inner speech, theorists have also suggested that there are forms of inner speech that are propositional or syntactic.

Vicente and Martínez-Manrique (2016) have offered a pluralist model that attempts to account for these diverse forms of inner speech (see also Vicente and Jorba (2017) and Grandchamp, et al. (2019)). Recall that, on the standard picture, inner speech is just truncated speech production: we start with the selection of a proposition and move down the hierarchy, but at some point processing is cut short. Where versions of monism claim that there is a specific place where processing is cut short – for Langland-Hassan, it is after the activation of a phonetic component, while for Gauker it is before – Vicente and Martínez-Manrique suggest that processing can be aborted at any level in the speech processing hierarchy, with the result being one or another form of inner speech. On their view, then, different forms of inner speech are explained in terms of the different points at which speech production processing might be cut short.

In developing their view, Vicente and Martínez-Manrique draw on predictive control models of speech production, an architecture that is supposed to enable the control, guidance, and correction of speech. Central to such models is the idea that predictions serve as a standard against which a given speech output can be assessed as correct or incorrect, allowing for the detection and subsequent correction of speech errors. According to these models, an intention to produce speech sounds is converted into a set of speech motor commands, which are executed at the vocal tract. In addition to this feedforward process, there is also a lateral process whereby a copy of the speech motor commands – an "efference copy" – is produced and used to generate a prediction of the speech sounds that would be present were the motor commands successfully executed. If there is a match between the predicted speech sounds and the actual speech sounds, then the vocalization is considered a success; if not, then error is fed back into the system and the vocalization is reattempted.

According to Vicente and Martínez-Manrique, inner speech is identical to a prediction of speech in the absence of the actual production of speech. On this account, speech motor commands are suppressed at the vocal tract, thereby blocking the production of speech. Nevertheless, the lateral process is still engaged, and a prediction of speech sounds is generated. Phonetic inner speech, on their view, just is the experience of the prediction of speech sounds in the absence of actual speech production. While traditional predictive control models posit a prediction only after the selection of motor commands, Vicente and Martínez-Manrique posit predictions corresponding to *each* level of the speech processing hierarchy. Drawing on Pickering and Garrod (2013), Vicente and Martínez-Manrique posit not only a prediction of speech sounds, but also a prediction of propositional and syntactic contents. They focus on the propositional case:

> In inner speech we form the non-conscious intention to express a certain thought, recruit semantic, syntactic, phonological representations, and issue a motor command to produce overt speech that is subsequently aborted…. Suppose, however, that we abort the process earlier, in particular, before the message is ready for emission. In that case, only one kind of prediction will be issued, namely, a prediction about the meaning of the message, which will be experienced as such meaning – and only a meaning. (2016, p.181)

On their view, propositional inner speech involves aborting speech production processing once a prediction "about the meaning of the message…and only [the] meaning" is on the scene. In contrast to Langland-Hassan and Gauker, Vicente and Martínez-Manrique suggest that there are multiple levels at which the speech production process may be aborted to produce inner speech, each level involving a different sort of prediction. Aborting speech production at different levels therefore produces inner speech with different contents.

A virtue of Vicente and Martínez-Manrique's position is that it recognizes a plurality of forms of inner speech and offers a general framework to account for them. However, I will now argue that pluralist accounts of the sort put forward by Vicente and Martínez-Manrique face two problems: the scope problem and the kindhood problem. These problems arise because the contents of the speech processing hierarchy – upon which predictive control models are built – fail to explain how inner speech can both have a plurality of manifestations and nevertheless form a kind. Although I will discuss these problems in relation to Vicente and Martínez-Manrique's model, they apply to pluralist models generally.

4.1 The Scope Problem

Vicente and Martínez-Manrique attempt to account for propositional and syntactic inner speech using the speech processing hierarchy situated within a predictive control architecture. However, looking only to the speech processing hierarchy leaves one unable to account for why representations of propositional contents and representations of syntactic contents count as states of inner speech as opposed to some other kind of mental state. Notice first that propositional contents are also the contents of beliefs, hopes, and fears, and syntactic contents are also the contents of inner writing. As a result, there is nothing about the *representational content* of such states that marks them as cases of inner speech as opposed to some other mental state kind. Moreover, since the speech processing hierarchy does not involve such speech-centric attitudes as assertion, saying, etc., there is nothing about the *manner of representing* propositional and syntactic contents that could mark off these states as states of inner speech either.

This problem of scope is made vivid in Vicente and Martínez-Manrique's model of propositional inner speech. On their view, propositional inner speech is identical to a prediction "about the meaning of the message…and only [the] meaning". The problem is that a prediction of a meaning or proposition need not be a case of inner speech. Of course, Vicente and Martínez-

Manrique might try to provide a functional characterization of what it is that makes a prediction of a meaning a case of inner speech. Given their predictive control model, they might claim that the activation of a prediction counts as propositional inner speech just in case it is caused by an efference copy of syntactic structure and causes an error signal if the prediction fails to match the actual proposition generated by the syntactic structure, and does not cause an error signal if it matches. But this at most provides a functional characterization of *a prediction of meaning*. Nothing about the proposed functional role indicates that what satisfies it is a case of *inner speech*. By adopting the standard picture, Vicente and Martínez-Manrique fail to show that propositional and syntactic inner speech count as cases of inner speech.

4.2 The Kindhood Problem

The problem of kindhood generalizes the conclusion of the problem of scope: whereas the previous section argued that the resources of the speech processing hierarchy are not enough to explain why propositional and syntactic forms of inner speech count as inner speech, the current section will argue that the hierarchy fails to account for inner speech as a mental state kind in the first place. A version of the kindhood problem was first raised by Langland-Hassan (2014).[8] Langland-Hassan considers and rejects three potential ways of accounting for inner speech as a

[8] Langland-Hassan (2014) discusses two problems. One problem – call it 'the kindhood problem' – concerns how distinct inner speech episodes with different combinations of contents from the speech processing hierarchy all count as being cases of inner speech (pp. 519-520). Another problem – call it 'the binding problem' – concerns how it is that a single inner speech episode possesses a combination of different contents from the speech processing hierarchy (pp. 520-529). Although these problems are related, a solution to the one does not entail a solution to the other. In particular, an account that states what it is in virtue of which inner speech episodes with various combinations of contents count as being cases of inner speech does not thereby state how any one of those episodes possesses the combination of contents it does. This paper is intended to provide a solution to the kindhood problem, not the binding problem.

kind of mental state: inner speech can be typed as a mental state kind in terms of its representational content, its functional role, or its neurobiological realizer.

According to Langland-Hassan, the representational content of inner speech fails to account for inner speech as a mental state kind because each level of the speech processing hierarchy concerns a distinct kind of subject matter. Propositional contents typically concern physical objects in the world, e.g., laundry and tables; syntactic contents concern words, e.g., *Sam*; phonetic contents concern speech sounds, e.g., clear l; and motor contents concern articulatory features, e.g., frication. It is unclear how the collection of states with such diverse contents would mark out a mental state kind. The situation would be analogous to grouping together a representation of olfactory information about pie with a representation of visual information about sand and calling the resulting collection a kind.

Appeal to functional role also fails to account for inner speech as a mental state kind. In order to succeed, according to Langland-Hassan, there needs to be a single functional role associated with inner speech. However, representations of propositions, representations of syntax, representations of phonemes, representations of speech sounds, and representations of articulatory features are not associated with a single functional role. For example, the functional role associated with representations of propositions – e.g., JOHN IS AT THE MEETING – will be vastly different than that of representations of articulatory features – e.g., frication.

Finally, appeal to neurobiological vehicles fails as well. Neurobiological states are typed by their causes and effects. But no single neurobiological state type will have a causal profile that makes it appropriate for it to serve as, for example, the vehicle of propositional content, the vehicle of syntactic content, the vehicle of phonetic content, and the vehicle of articulatory content. The upshot of the kindhood problem is that inner speech cannot be typed as a mental state either in terms of content, functional role, or neurobiological realizer.

Vicente and Martínez-Manrique's pluralist model of inner speech faces its own version of the kindhood problem. Consider trying to type inner speech by its functional role or neurobiological realizer given their pluralist model. Since predictive control models are hierarchical, predictions of meaning and predictions of speech sounds have different causes: whereas predictions of meaning are caused by an efference copy of syntax, predictions of speech sounds are caused by an efference copy of motor commands. As a result, the predictions will have different functional roles, and so cannot be lumped together as inner speech in virtue of their functional role. A similar argument can be given against typing inner speech in terms of its neurobiological realizer. We expect predictions of meaning and predictions of speech sounds to be subserved by neural states with distinct causal profiles. Indeed, in Grandchamp et al.'s (2019) pluralist model, propositional inner speech ("condensed inner speech") is associated with activation of the posterior middle temporal gyrus, while phonetic inner speech ("semi-expanded inner speech") is realized by the inferior temporal lobule. Although these are broad brain regions, they nevertheless have different causal profiles, suggesting that inner speech cannot be typed as a mental state kind using neurobiological realizers.[9]

The kindhood problem seems to present us with a dilemma: either we affirm the intuition that inner speech is a mental state kind, but give up the idea that there are distinct forms of inner speech, or else embrace the idea that there are distinct forms of inner speech, but give up on the

[9] A leftover possibility is that a state counts as inner speech in virtue of being a part of an aborted speech production process. The problem with this proposal, however, is that being a part of a mental process does not individuate mental state kinds. For example, states corresponding to prediction, prediction error, precision, and data are assumed to be parts of the mental process of prediction error minimization (Clark, 2015). But these states do not belong to some further mental state kind in virtue of being a part of the process of prediction error minimization. Or again the mere fact that motor, sensory, and cognitive states are part of the speech perception process does not entail that they are themselves states of speech perception, nor that there is some further kind to which they belong.

hope of accounting for inner speech as a mental state kind. However, as I shall argue, this dilemma is a false one. It arises from a background allegiance to the standard picture of inner speech. Because the contents of the speech processing hierarchy are just too diverse to demarcate a kind of mental state, allegiance to the standard picture forces us to identify inner speech with some single level of the hierarchy or else reject the idea that inner speech is a mental state kind. But what if we look elsewhere for the content of inner speech? Can a different type of content provide inner speech with unity to ground its kindhood? I will argue that the contents of inner speech stem, in part, from resources devoted to *voice* processing. Appeal to voice processing provides us with a uniform representational content that shows how inner speech can be a mental state kind despite the plurality of its manifestations.

5. Toward Vocal Content

A more promising view of the content of inner speech comes into focus once we 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 – speech, affect, and identity – are captured by Belin and colleagues’ model of voice processing. According to their model, a signal communicated by a voice is independently processed for speech, affect, and identity (see Figure 2).

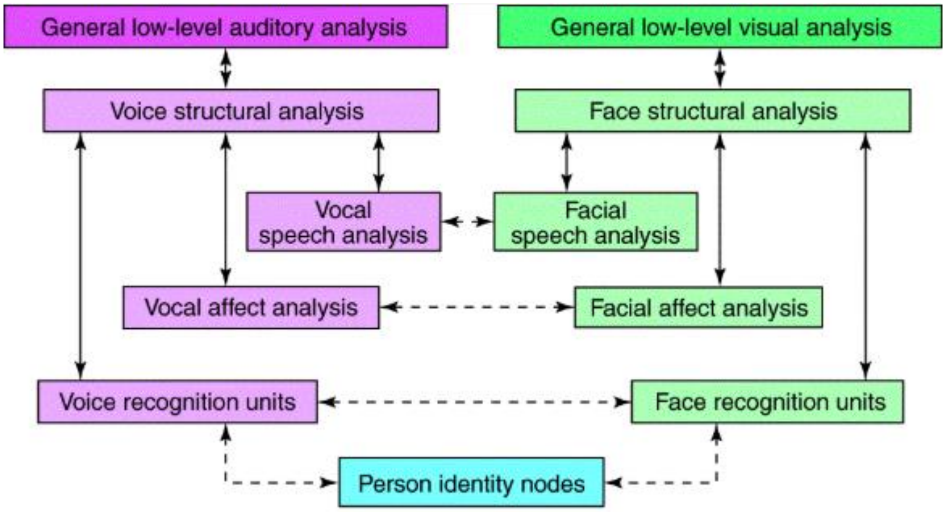

Fig. 2: Models of face and voice recognition. Reproduced from Belin et al. (2004).

In the first stage, the 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 signal for information relating to affect and other socially relevant properties, including weight, gender, competence, personality, and the like. Identity processing extracts from the signal the identity of the speaker, e.g., that it is Sam's voice. Finally, on the basis 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. 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 (e.g., angrily, sadly, etc.) and *whose voice* made it (e.g., 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. Thus, when we hear someone's utterance, we do not simply represent speech/language information – propositions, words, phonemes, etc. – but also a *voice communicating* such information.

The view of inner speech I will be presenting in Section 6 builds on Belin and colleagues' work to argue that the content of inner speech is *vocal content*. As I will characterize it, vocal content is representational content that specifies a two-place relation: a voice communicating information. There is a slot for a specification of voice, a 'voice slot', and a slot for the specification of information, an 'information slot'. In the rest of this Section, I explain all three components of vocal content, starting with the relation of *communication*.

There is some difficulty in determining the exact relation that voice bears to information. One option is that voice *carries* or *expresses* information. However, as it is typically used, *expression* 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.

I favor a strategy on which voice is thought of as a *channel* of information. 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: motoric, phonetic, phonemic, syntactic, 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.[10]

Turning next to the information slot in vocal content, the vocal channel can communicate any of the contents of the speech processing hierarchy as well as contents relating to affect and identity. My voice (a channel) might communicate secrecy or fear (affect) as it also communicates the words *don't tell anyone* (speech), all the while giving away who I am to the receiver (identity). In normal cases, there is a dependency between the types of information communicated by voice. Vocal channels communicate a proposition typically by communicating words, and they communicate words typically by communicating speech sounds, and so on. Thus, in normal

[10] One might object that my use of the concept of voice is equivocal. On the one hand, voice refers to *information* that is extracted from a signal, while on the other hand, voice refers also to the *medium* within which a signal is communicated. The charge of equivocation stems from the fact that I can come to represent someone's voice – as a medium – only by extracting information from a signal. This can make it seem as if voice is also *represented as* information extracted from a signal (e.g., on a par with speech sound information). But this does not follow. Although I learn facts about a medium *via* the signal it carries, I nevertheless continue to represent the medium as a medium. The same holds for voice: I learn how a voice sounds or whose voice it is by extracting information from a signal, but in so doing I continue to represent the voice as a medium through which the signal is communicated.

contexts, there is no such thing as a vocal channel directly communicating a proposition. However, as we shall see in Section 6, 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. As I noted, voices are channels of communication, not sources or signals. 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. Representing a voice as belonging to Bob requires not only that I distinguish it from voices belonging to other subjects, but also that I represent variations of Bob's voice as being instances of 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. 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 values are nearby those of the prototype are taken to be produced by the same voice. 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. 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. 9). 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 at least 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 that derive from the speech processing hierarchy. While the latter include propositional, syntactic, phonemic, phonetic, and articulatory contents, 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 the plurality of ways of filling in the voice and information slots. Voices can be characterized either auditorily or amodally, while the information communicated by voices includes not only the contents of the speech processing hierarchy, but also information relating to affect and identity.

6. 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 implicates vocal processing.

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 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. 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. The takeaway from this evidence is that inner speech implicates not just speech processing, but voice processing more broadly.[11]

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.

[11] One might object that this evidence bears on the phenomenon of *imagining speech* and not on the phenomenon of *inner speech*. According to this objection, imagined speech involves the representation of another’s voice, while inner speech necessarily involves the representation of one’s own voice. However, it is not clear that the ‘own voice-other voice’ distinction marks a difference in mental state kinds; the objector must show that it does.

Vocalism can be used to reconstruct the forms of inner speech observed in Section 4 by specifying both the voice slot and the information slot (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 communicating *the meeting is cancelled*.

Phonemic inner speech is the sort of inner speech at the heart of Oppenheim and Dell’s (2010) “flexible abstraction hypothesis”, according to which “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 communicating /ðə mitɪŋ ɪz kænsəld/.

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	Articulatory features

Phonetic inner speech is the sort of paradigmatic “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 communicating [ðə ˈmiːɾɪŋ ɪz ˈk^{h}ænsəld].

Articulatory inner speech is the sort of inner speech exhibited in minute movements of the vocal tract. McGuigan and Dollins (1989) have shown that when subjects produce a phone in inner speech their vocal tract moves (minutely) in a manner consistent with the phone produced. Such articulatory inner speech is marked by a representation of a motoric specification of voice communicating vocal tract movements. For example, I may represent my own voice in a motor register communicate {…}, where '…' indicates the complex array of motor gestures needed for the voice to communicate [ðə ˈmiːɾɪŋ ɪz ˈk^{h}ænsəld].

There are two important points to make about this lineup of forms of inner speech. First, propositional, syntactic, and phonemic inner speech each implicate the same substitution for the voice slot: an amodal specification of voice. In general, I assume that if one represents information that is amodal – non-auditory and non-motoric – then one must also represent as amodal the *voice* communicating that information. What it is to represent an amodal specification of a voice is to represent the vocal identity of the speaker, e.g., Sam's voice. The representation of vocal identity is a representation of voice that is abstracted from auditory and articulatory features (see Section 5). Propositional, syntactic, and phonemic contents are thus represented as being communicated by a voice that is amodally specified. Second, 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 5, it can be thought of (roughly) as a specification of voice in terms of those motor features that would, if executed, give rise to an auditory characterization of voice.

In addition to accounting for the diverse manifestations of inner speech, vocalism avoids the two problems for the standard picture of the contents of inner speech: the problem of scope and the problem of kindhood.

Recall that appeal to the contents of the speech processing hierarchy is unable to account for propositional and syntactic forms of inner speech. Vocalism solves this problem by claiming that cases of inner speech count as inner speech in virtue of possessing vocal content. In the case of propositional inner speech, one represents a voice communicating a proposition, while in the case of syntactic inner speech, one represents a voice communicating a syntactic ordering of words. Given that propositional and syntactic inner speech involve vocal content, these more abstract forms of inner speech also thereby count as cases of inner speech. The immediate result is that the same feature that makes phonetic inner speech a case of inner speech also makes propositional and syntactic inner speech cases of inner speech. On the picture presented by vocalism, phonetic inner speech does not have more of a claim to being inner speech because it also possesses speech-centric content. There is no longer reason to view phonetic content as the center of gravity when it comes to inner speech. According to vocalism, different forms of inner speech are all on a par in virtue of possessing vocal content.

The problem of kindhood arose because the standard picture made it hard to see how inner speech could be a mental state kind. Vocalism solves this problem by claiming that inner speech is a mental state kind in virtue of a *structural feature* of its content. As I type this sentence, I may engage in motoric inner speech, phonetic inner speech, phonemic inner speech, syntactic inner speech, or propositional inner speech, or some combination thereof. Each of these manifestations of inner speech possesses a different type of content: motoric contents, phonetic contents, phonemic contents, syntactic contents, and propositional contents. These content types mark part of the difference between these manifestations. The error of the standard picture is the assumption that this is all there is to the content of inner speech. In contrast, according to vocalism, each of these content types itself figures as one part of a more complex, embracing content. When I engage in motor inner speech, I represent a *voice communicating* motor gestures; when I engage in phonetic inner speech, I represent a *voice communicating* speech sounds; and so on. Each vocal content thus possesses a common structure: *a voice type communicating an information type*. And it is this common structure, shared across different vocal contents, that qualifies an episode of inner speech as being one of inner speech. What the problem of kindhood gets right is that if there is nothing in common across the contents of different types of inner speech, then we are forced to conclude that inner speech is not a mental state kind. What vocalism points out is that there *is* something in common once we look past the speech processing hierarchy.[12]

One might object that the standard picture can be enlisted to do all the explanatory work that vocalism has been made to do. According to this strategy, we can use the speech processing hierarchy to construct a view analogous to vocalism. On this view, the relevant contents do not have the form, *voice type communicating information type*, but instead have the form, *speech communicating information type*. Thus, on this picture, in inner speech we may represent speech communicating a proposition, speech communicating syntax, speech communicating phonemes, and so on. The scope and kindhood problems would be avoided by this view since, as is the case with vocalism, different forms of inner speech involve contents with the same structure: *speech communicating an information type*. If this speech processing-based account is as plausible as

[12] This paper has been concerned with the nature of the contents of inner speech, and not with the further, important question concerning the mechanism by which inner speech possesses its content (see Carruthers, 2010; Langland-Hassan, 2014; Vicente and Martínez-Manrique, 2016; Knappik, 2017). I believe that it is fruitful to first provide an account of the content of inner speech before fixing on one or another specific mechanism by which it is endowed with such content.

vocalism, then there is nothing essential about appealing to vocal content in resolving the scope and kindhood problems.

The problem with this view is that although the contents of the speech processing hierarchy play a role in the generation of speech, in representing propositional, syntactic, phonemic, phonetic, and articulatory contents one does not thereby represent these contents *as* communicated by speech. For this reason, representing the contents of the speech processing hierarchy does not entail that one represents contents of the form, *speech communicates information type*. In order to emulate vocalism, the contents of inner speech must arise from a confluence of distinct processes. Vocal processing is composed of speech, affect, and identity processes, which, when they come together in inner speech generate complex contents with a common structure: *a voice type communicating an information type*. But the speech processing-based view on offer implicates only one of these processes – the speech process – and thereby fails to generate complex contents analogous to those provided by vocalism. The speech processing hierarchy provides for representations of propositional, syntactic, phonemic, phonetic, and articulatory content – but nothing more.

7. Conclusion: Unity and Plurality in Inner Speech

This paper has attempted to perform a balancing act. On the one hand, inner speech is diverse. There are propositional and syntactic forms of inner speech in addition to paradigmatic sensorimotor forms. On the other hand, inner speech is unified. It is a mental state kind and not an arbitrary collection of propositional, syntactic, phonemic, phonetic, or articulatory states. I have sought to do justice to both the diversity and unity of inner speech. The unity of inner speech is explained by the common structural element within the contents of inner speech, while its plurality stems from the different ways that structure can be filled in.[13]

[13] Many thanks to Edouard Machery, Peter Langland-Hassan, Wayne Wu, James Shaw, Mark Wilson, and Zina Ward for discussion and feedback on this paper.

References

Alderson-Day, B., & Fernyhough, C. (2015). Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology. *Psychological Bulletin*, *141*(5), 931–965. https://doi.org/10.1037/bul0000021

Anscombe, G. E. M. (2000). *Intention* (2nd edition). Harvard University Press.

Awwad Shiekh Hasan, B., Valdes-Sosa, M., Gross, J., & Belin, P. (2016). "Hearing faces and seeing voices": Amodal coding of person identity in the human brain. *Scientific Reports*, *6*(1), 37494. https://doi.org/10.1038/srep37494

Belin, P. (2019). The "Vocal Brain": Core and extended cerebral networks for voice processing. In *The Oxford Handbook of Voice Perception*. Oxford University Press.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312. https://doi.org/10.1038/35002078

Buchtel, H. A., & Stewart, J. D. (1989). Auditory agnosia: Apperceptive or associative disorder? *Brain and Language*, *37*(1), 12–25. https://doi.org/10.1016/0093-934x(89)90098-9

Carruthers, P. (2010). Introspection: Divided and Partly Eliminated. *Philosophy and Phenomenological Research*, *80*(1), 76–111. https://doi.org/10.1111/j.1933-1592.2009.00311.x

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Denes, G., & Semenza, C. (1975). Auditory modality-specific anomia: Evidence from a case of pure word deafness. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *11*(4), 401–411. https://doi.org/10.1016/S0010-9452(75)80032-3

Fromkin, V. (1971). The Non-Anomalous Nature of Anomalous Utterances. *Language*, *47*. https://doi.org/10.2307/412187

Fruhholz, S., & Belin, P. (2019). *The Oxford Handbook of Voice Perception*. Oxford University Press.

Gauker, C. (2018). Inner Speech as the Internalization of Outer Speech. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: New Voices* (pp. 53–77). Oxford: Oxford University Press.

Grandchamp, R., Rapin, L., Perrone-Bertolotti, M., Pichat, C., Haldin, C., Cousin, E., Lachaux, J.-P., Dohen, M., Perrier, P., Garnier, M., Baciu, M., & Lœvenbruck, H. (2019). The ConDialInt Model: Condensation, Dialogality, and Intentionality Dimensions of Inner Speech Within a Hierarchical Predictive Control Framework. *Frontiers in Psychology*, *10*, 2019. https://doi.org/10.3389/fpsyg.2019.02019

Hurlburt, R. T., & Heavey, C. L. (2018). Inner Speech as Pristine Inner Experience. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: New Voices*. Oxford: Oxford University Press.

Hurlburt, R. T., Heavey, C. L., & Kelsey, J. M. (2013). Toward a phenomenology of inner speaking. *Consciousness and Cognition*, *22*(4), 1477–1494. https://doi.org/10.1016/j.concog.2013.10.003

Jescheniak, J. D., Meyer, A. S., & Levelt, W. J. M. (2003). Specific-word frequency is not all that counts in speech production: Comments on Caramazza, Costa, et al. (2001) and new experimental data. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(3), 432–438. https://doi.org/10.1037/0278-7393.29.3.432

Jones, P. E. (2009). From ‘external speech’ to ‘inner speech’ in Vygotsky: A critical appraisal and fresh perspectives. *Language & Communication*, *29*(2), 166–181. https://doi.org/10.1016/j.langcom.2008.12.003

Joyce, J. (2015). *Ulysses* (D. Kiberd, Ed.; New Edition). Penguin.

Knappik, F. (2018). Bayes and the first person: Consciousness of thoughts, inner speech and probabilistic inference. *Synthese*, *195*(5), 2113–2140. https://doi.org/10.1007/s11229-017-1321-3

Kurby, C. A., Magliano, J. P., & Rapp, D. N. (2009). Those voices in your head: Activation of auditory images during reading. *Cognition*, *112*(3), 457–461. https://doi.org/10.1016/j.cognition.2009.05.007

Langland-Hassan, P. (2014). Inner Speech and Metacognition: In Search of a Connection. *Mind & Language*, *29*(5), 511–533. https://doi.org/10.1111/mila.12064

Langland-Hassan, P. (2018). From Introspection to Essence: The Auditory Nature of Inner Speech. In P. Langland-Hassan & A. Vicente (Eds.), *Inner Speech: New Voices*. Oxford University Press.

Langland-Hassan, P., & Vicente, A. (Eds.). (2018). *Inner Speech: New Voices*. Oxford University Press.

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-Based Coding of Voice Identity in Human Auditory Cortex. *Current Biology*, *23*(12), 1075–1080. https://doi.org/10.1016/j.cub.2013.04.055

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications*, *10*(1), 2404. https://doi.org/10.1038/s41467-019-10295-w

Levelt, W. J. M. (1993). *Speaking: From Intention to Articulation*. MIT Press.

MacKay, D. G. (1992). Constraints on theories of inner speech. In *Auditory imagery* (pp. 121–149). Lawrence Erlbaum Associates, Inc.

Marshall, R. C., Rappaport, B. Z., & Garcia-Bunuel, L. (1985). Self-monitoring behavior in a case of severe auditory agnosia with aphasia. *Brain and Language*, *24*(2), 297–313. https://doi.org/10.1016/0093-934X(85)90137-3

McGuigan, F. J., & Dollins, A. B. (1989). Patterns of covert speech behavior and phonetic coding. *The Pavlovian Journal of Biological Science*, *24*(1), 19–26.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Oppenheim, G. M., & Dell, G. S. (2010). Motor movement matters: The flexible abstractness of inner speech. *Memory & Cognition*, *38*(8), 1147–1160. https://doi.org/10.3758/MC.38.8.1147

Papathanasiou, I., Macfarlane, S., & Heron, C. (1998). A Case of Verbal Auditory Agnosia: Missing the Word…missing the Sound…. *International Journal of Language & Communication Disorders*, *33*(S1), 214–218. https://doi.org/10.3109/13682829809179425

Perrone-Bertolotti, M., Rapin, L., Lachaux, J. P., Baciu, M., & Lœvenbruck, H. (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*, 220–239. https://doi.org/10.1016/j.bbr.2013.12.034

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *The Behavioral and Brain Sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Roelofs, A., Meyer, A. S., & Levelt, W. J. (1998). A case for the lemma/lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition*, *69*(2), 219–230. https://doi.org/10.1016/s0010-0277(98)00056-0

Tsantani, M., Kriegeskorte, N., McGettigan, C., & Garrido, L. (2019). Faces and voices in the brain: A modality-general person-identity representation in superior temporal sulcus. *NeuroImage*, *201*. https://doi.org/10.1016/j.neuroimage.2019.07.017

Vicente, A., & Jorba, M. (2019). The Linguistic Determination of Conscious Thought Contents. *Noûs*, *53*(3), 737–759. https://doi.org/10.1111/nous.12239

Vicente, A., & Martínez-Manrique, F. (2016). The nature of unsymbolized thinking. *Philosophical Explorations*, *19*(2), 173–187. https://doi.org/10.1080/13869795.2016.1176234

Wu, W. (2012). Explaining Schizophrenia: Auditory Verbal Hallucination and Self-Monitoring. *Mind & Language*, *27*(1), 86–107. https://doi.org/10.1111/j.1468-0017.2011.01436.x

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. https://doi.org/10.1162/jocn_a_00022