This article is the OnlineFirst version of:

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014). Speakers' acceptance of Real-Time Speech Exchange indicates that we use auditory feedback to specify the meaning of what we say. *Psychological Science*. DOI: 10.1177/0956797614529797 (including Supplemental Online Material)

Abstract: Speech is usually assumed to start with a clearly defined preverbal message, which provides a benchmark for self-monitoring and a robust sense of agency for one's utterances. However, an alternative hypothesis states that speakers often have no detailed preview of what they are about to say, and that they instead use auditory feedback to infer the meaning of their words. In the experiment reported here, participants performed a Stroop color-naming task while we covertly manipulated their auditory feedback in real time so that they said one thing but heard themselves saying something else. Under ideal timing conditions, two thirds of these semantic exchanges went undetected by the participants, and in 85% of all nondetected exchanges, the inserted words were experienced as self-produced. These findings indicate that the sense of agency for speech has a strong inferential component, and that auditory feedback of one's own voice acts as a pathway for semantic monitoring, potentially overriding other feedback loops.

Keywords: Speech Production, Self-Monitoring, Voice Manipulation, Sense of Agency, Error Monitoring, Real-time Speech Exchange, Auditory Feedback

Psychological Science OnlineFirst, published on April 28, 2014 as doi:10.1177/0956797614529797

Research Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
1–8
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797614529797
pss.sagepub.com
SAGE

Speakers' Acceptance of Real-Time Speech Exchange Indicates That We Use Auditory Feedback to Specify the Meaning of What We Say

Andreas Lind[1], Lars Hall[1], Björn Breidegard[2], Christian Balkenius[1], and Petter Johansson[1,3]
[1]Lund University Cognitive Science, Lund University; [2]Certec, Division of Rehabilitation Engineering Research, Department of Design Sciences, Faculty of Engineering, Lund University; and [3]Swedish Collegium for Advanced Study, Linneanum, Uppsala University

Abstract
Speech is usually assumed to start with a clearly defined preverbal message, which provides a benchmark for self-monitoring and a robust sense of agency for one's utterances. However, an alternative hypothesis states that speakers often have no detailed preview of what they are about to say, and that they instead use auditory feedback to infer the meaning of their words. In the experiment reported here, participants performed a Stroop color-naming task while we covertly manipulated their auditory feedback in real time so that they said one thing but heard themselves saying something else. Under ideal timing conditions, two thirds of these semantic exchanges went undetected by the participants, and in 85% of all nondetected exchanges, the inserted words were experienced as self-produced. These findings indicate that the sense of agency for speech has a strong inferential component, and that auditory feedback of one's own voice acts as a pathway for semantic monitoring, potentially overriding other feedback loops.

Keywords
speech production, self-monitoring, voice manipulation, sense of agency

Received 7/16/13; Revision accepted 3/3/14

As adults with intimate experience of our own minds, we feel it is self-evident that we always know the meaning of what we are going to say, before we actually say it. But what would it be like if we said one thing and heard ourselves saying something else? Would we experience this as an alien voice in our heads, a strange form of auditory hallucination? Or would we perhaps trust our ears over our mouths, and believe we actually said the thing we heard?

Current theories of speech production assume that speech starts with a clear preverbal conception of what to say, which is then translated into an utterance through successive levels of linguistic and articulatory encoding. A cascade of internal monitoring loops—from conceptual, to lexical, to syntactic, to articulatory, to efference, to proprioceptive monitoring, and finally out to auditory feedback—serves to guarantee agreement between intention and outcome (e.g., Hickok, 2012, 2014; Levelt, 1989; Pickering & Garrod, 2013; Postma, 2000). Thus, according to this dominant view, the intended meaning always precedes the ultimate shape of the utterance.

According to an alternative model, however, speech is not just the dutiful translation of a well-defined preverbal message. Rather, through rapid, on-line interaction between the speaker and the conversational context, competing and approximate speech goals arise and become increasingly specific during the articulation process (Dennett, 1991; Linell, 1982, 2009; see also Lind, Hall, Breidegard, Balkenius, & Johansson, 2014). From

Corresponding Author:
Andreas Lind, Lund University Cognitive Science, Lund University, Kungshuset, Lundagård, 222 22 Lund, Sweden
E-mail: andreas.lind@lucs.lu.se

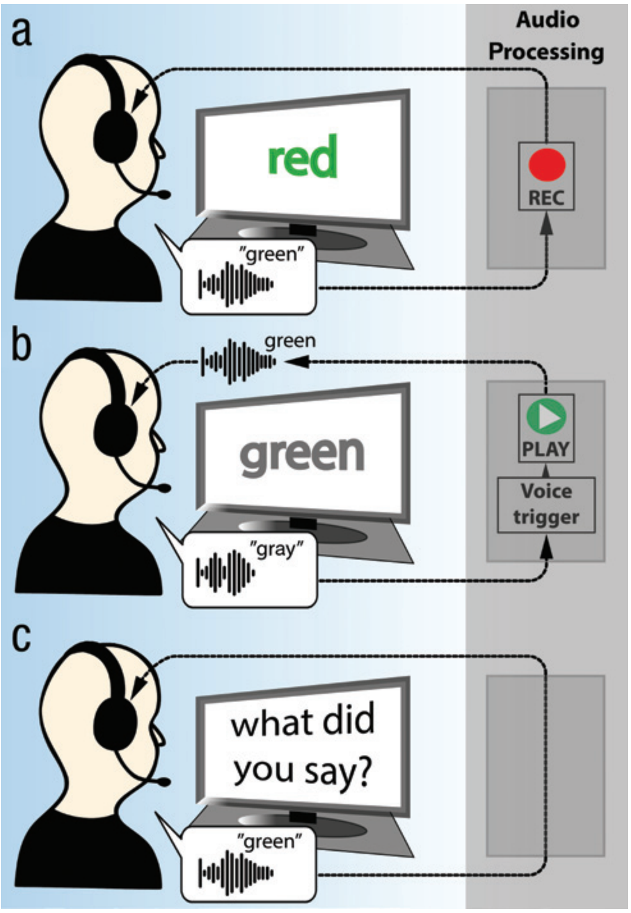

Fig. 1. Illustration of the experimental procedure. Participants performed a Stroop test, in which they were asked to name the font color of each word presented on the screen. They heard their own voice through a noise-canceling headset, while the experimenter surreptitiously recorded the words they said (a). During manipulated trials (b), the experimenter activated a voice trigger, and when the microphone signal exceeded a preset amplitude, the previously recorded word was substituted for the uttered word in the auditory feedback; the sound of the participant's actual utterance was blocked out. The inserted recording was the color word named by the letters, and was thus an incorrect response in the Stroop test. Directly following each manipulated trial (c), the question "What did you say?" appeared on the screen and remained until the participant verbalized an answer.

this perspective, auditory feedback takes on a much more active and interpretive role, and speakers listen to their own utterances to help specify the meaning of what they just said. Depending on timing and contextual demands, they might rely more or less on auditory feedback, but they always use that channel as a source of evidence in interpreting their utterances.

This interesting opposition of *comparator* and *inferential* models (or *predictive* and *reconstructive* models, as they sometimes are called; see Haggard & Clark, 2003; Kühn, Brass, & Haggard, 2012; Synofzik, Vosgerau, & Newen, 2008) is similarly found, and has been widely discussed, in the domain of manual action. According to the first perspective, comparator processes anchor people's fundamental sense of agency, and allow them to discriminate between actions generated by themselves and actions generated by others (Blakemore, Wolpert, & Frith, 2002; David, 2012; Gallagher, 2000; Kühn et al., 2012). Furthermore, these processes enable error correction by separating deliberate from accidental outcomes (Frith, 2013) and what one has done from what one planned to do (Sugimori, Asai, & Tanno, 2013). In contrast, inferential models see attribution of agency as a more situated and fluent process, and maintain that it often can be confused in both natural and experimental conditions (Moore, Wegner, & Haggard, 2009; Wegner & Wheatley, 1999).

However, surprisingly, there have been very few attempts to directly test the relative adequacy of these opposing views in the speech domain (Dennett, 1991). A conceptually simple but technically challenging way to engineer such a test would be to create the hypothetical scenario mentioned in our introduction: A person says one thing but hears him- or herself saying something else. If the dominant comparator view of speech production is correct, whole-word substitutions created at the auditory-feedback stage should be readily detected. But if auditory feedback is a critical factor in an inferential process of agency attribution, then such mismatches might go undetected and influence speakers' beliefs about what they have said, making them act as if the inserted statements were self-produced.

In the experiment reported here, we performed such a direct test. To create the convincing speech exchange that was required, we had to fulfill three conditions: First, we needed to be able to predict what participants would say in response to an experimental stimulus, and when they would say it, in order to record the appropriate words and subsequently insert them into the feedback loop. Second, to prevent the substituted words from being immediately discounted as too improbable, we needed to create a context in which more than one response to the experimental stimuli was possible. Third, the word insertions had to be made with great temporal precision, or else mismatches could be detected on the basis of timing discrepancies alone. To meet these demands, we used the classic Stroop test (naming the font color of a presented color word) to provide structure and predictability, and we created a voice-triggered playback platform that achieved speech exchange with very high timing accuracy. During the experiment, we recorded some single color-word utterances and then covertly played them back on later trials (see Fig. 1). Thus, participants said one thing, but heard themselves through headsets saying something else. Directly following the manipulation, an on-screen prompt asked participants, "What did you say?" which allowed us to measure whether they believed that they had uttered the inserted word.

Method

Participants

Eighty-three participants (44 female, 39 male; mean age = 23.7 years, *SD* = 4.1), most of whom were students, were recruited at Lund University. All participants spoke Swedish as their first language, and none had any auditory or visual impairments. Participants were fully debriefed after the experiment, before giving informed consent for their data to be used. The data from 5 participants were removed from further analysis because of technical problems, which left 78 participants. The study was approved by the Lund University ethics board (Reference No. 2008–2435).

Materials

We constructed a semiautomated auditory-feedback control system that allowed us to covertly record and trim a specific word and, using a voice trigger, play this word back to the participants through headphones at the exact time that they uttered another word (see Fig. 1).[1] Participants wore a specially constructed sound-isolated circumaural headset, characterized by high sound quality combined with considerable passive sound attenuation of the air-conducted auditory signal (see the Supplemental Material available online). Very high timing accuracy was achieved for the majority of the trials with the speech exchange (*manipulated trials*). However, sometimes the trimming failed or smacking noises triggered the playback, so that the timing of the manipulated segment did not match the timing of the participant's speech.

Procedure

The participants performed a 250-word Stroop test in Swedish, with the instruction to name the color each presented word was written in. Twenty-five different word-color combinations were used; each appeared 10 times in the experiment. The order of the words was randomized, and the same order was used for all participants. The words were presented for 200 ms, and the interstimulus interval was 2,000 ms.

Participants were seated in front of a computer screen and were given verbal instructions about how to perform the Stroop test. They were told that the test would occasionally stop and that the question "What did you say?" would be displayed on the screen. Once they had answered the question, the test would resume. The experiment took approximately 10 min to complete.

During the experiment, two color-word combinations were used in the manipulated trials: Either the previously recorded word "green" ("grön") was inserted when participants uttered "gray" ("grå") or vice versa. In effect, we inserted the incorrect answer in the current Stroop trial. In Swedish, "gray" ("grå") is pronounced [gɹoː], and "green" ("grön") is pronounced [gɹøːn]. Thus, these words are phonologically similar but semantically distinct. In total, four manipulated trials were included in the experiment (two of each kind, in alternating order). Participants were asked, "What did you say?" at the end of the manipulated trials and also, as a control, at the end of four nonmanipulated trials distributed among the manipulated trials (for additional details on the procedure, see the Supplemental Material).

Detection criteria

To determine if the participants had become aware of the manipulations, we conducted a structured posttest interview, asking increasingly specific questions about the participants' experience of the experiment, before finally revealing the manipulation and asking if they had suspected any substitutions (see the Supplemental Material). If participants indicated that they had detected any of the manipulations, we asked follow-up questions to capture their experiences of the manipulated feedback as fully as possible. Combined with listening to the participants' behavior on each individual trial, this procedure allowed us to establish a trial-by-trial detection rate.

The certainty with which participants expressed potential detections varied widely. To capture this variation, we classified detections into three levels of epistemic certainty. If participants explicitly described how they had received false feedback, we categorized the trial as a "certain detection." If they had a suspicion but did not identify what had happened, we categorized the trial as an "uncertain detection." Finally, if they expressed vague confusion about the utterance, or if they claimed to have noticed something strange about the feedback only after we revealed the full procedure to them, the trial was considered a "possible detection" (Johansson, Hall, Sikström, & Olsson, 2005).

Results

Sixteen of the manipulated trials were aborted because of difficulty in securing a prior recording of the target word, and an additional 12 trials were removed from analysis because the participants made errors in the Stroop test. If participants detected an exchange, they were alerted to the external manipulation and the purpose of the experiment. The test then changed to an explicit mismatch-detection task, and given the low baseline error rate on

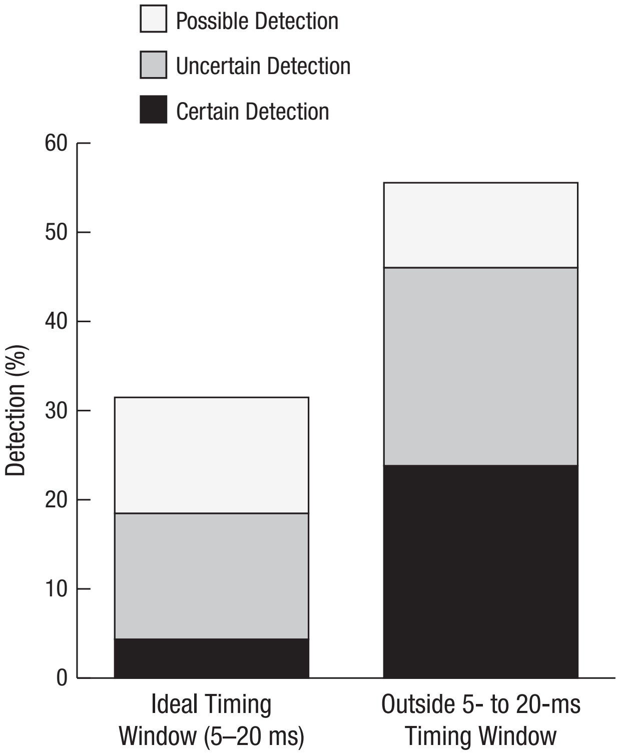

Fig. 2. Percentage of manipulated trials that were detected for trials within (n = 92) and outside (n = 63) the a priori estimated ideal timing window.

the Stroop test (2%), it was easy to self-monitor on the basis of the objective criterion of correctness in the task (i.e., participants could remember the correct answer by recalling the visual representation on the screen). To avoid any such confounds, we removed all trials following a first detection (total of 129 trials). Thus, our analyses included 155 manipulated trials (54.6% of all manipulated trials). There were no differences in detection rate between manipulated trials in which “gray” was replaced by “green” and in which “green” was replaced by “gray,” $\chi^2(1, N = 155) = 0.002$, $p = .96$, so we present results for a combined measure.

As there were no prior studies of real-time speech exchange, we explored the impact of timing accuracy by dividing the trials into two categories based on the timing of the auditory exchange relative to the uttered word. A timing mismatch of no more than 5 to 20 ms was considered the ideal (see the Supplemental Material). Under these ideal timing conditions, we found a low detection rate (total of 32% for the three detection categories), with only 4% of these detections falling in the “certain” category (Fig. 2). This means that when near-simultaneous timing conditions were met, very few participants had more than a vague hunch that what they heard themselves say was not what they actually said. As Figure 2 shows, significantly more manipulations were detected when the timing mismatches fell outside the 5- to 20-ms window, $\chi^2(1, N = 155) = 7.9$, $p = .005$, even though a considerable percentage of the manipulations remained undetected (for additional analyses involving detection rates, see the Supplemental Material).

But regardless of timing, how did participants respond to the question “What did you say?” when they did not detect the manipulation? Virtually every time they were asked this question following a nonmanipulated trial (99.4%), they simply repeated what they had said, showing that they were focused and attentive during the test, and had no trouble answering this question. However, looking at the manipulated trials, we found a variety of responses indicating that participants accepted the exchanged word as being self-produced.

We classified these responses into four categories (Table 1). On a large number of trials, participants answered the question according to what they had heard themselves say, in effect acknowledging that they had made an error on the test. On other trials, participants spontaneously corrected themselves, thereby indicating that they believed they had uttered the inserted word. These corrections took the form of either repeating what they had actually said (before the question was shown) or clarifying the correct Stroop response when answering the question (e.g., “I mean gray”). There was also a class of trials in which participants answered the question by repeating what they had actually said, but (as revealed in the posttest interview) believed they had made a mistake and were correcting what they had said. That is, they accepted the inserted word as self-produced, but they answered the question according to what they thought the correct answer to the Stroop trial was. Finally, in a few cases, participants similarly repeated the correct answer, but their responses and interviews provided inconclusive evidence as to whether they believed they had uttered the inserted words. Some participants’ responses following the manipulated trials fell into more than one category (e.g., one type of response was elicited for the first manipulation and another type for the second). Summing the frequencies of the first three categories of responses, we found that in a full 85% of the nondetected manipulated trials, participants accepted the manipulated feedback as having been self-produced.

Discussion

Our paradigm created a mismatch between what participants said and the auditory feedback they received, thereby allowing us to investigate semantic aspects of the real-time interaction between feed-forward and feedback mechanisms in speech production. Participants had strong evidence about what they had actually said, from proprioceptive and bone-conducted feedback, as well as

Table 1. Classification of Trials in Which the Manipulation Was Not Detected According to the Evidence Indicating Whether Participants Believed They Had Uttered the Inserted Word

	Example[a]			
Participants' behavior	Response to Stroop stimulus	Inserted word	Answer to "What did you say?"	*n*
Reported saying the inserted word[b]	"gray"	"green"	"green"	35 (38.5%)
Corrected themselves spontaneously[b]				15 (16.5%)
Repeated the Stroop response before being asked what they said	"gray . . . no, gray"	"green"	—	
Clarified the Stroop response	"gray"	"green"	"I mean gray"	
Admitted (in the posttest interview) that their report of what they had said was a correction[b]	"gray"	"green"	"gray"	27 (29.7%)
Inconclusive	"gray"	"green"	"gray"	14 (15.4%)

[a]These examples are taken from trials in which participants correctly said "gray" in response to the Stroop stimulus but the word "green" was substituted in the auditory feedback. [b]These responses indicate that participants believed they had uttered the inserted words. This was the case on 85% of the trials in which the manipulation was not detected.

from their visual memory of the experimental stimulus and their long history of correct answers in the test. But despite this, they often accepted the inserted words as self-produced. This indicates that speakers listen to their own voices to help specify to themselves the meaning of what they are saying, rather than just to make sure they have said what they intended to say. Specifically, it suggests that auditory feedback is a pathway for high-level semantic monitoring that is powerful enough to override other monitoring channels.

Prior studies of auditory-feedback perturbation have established that speakers react to frequency shifts of the fundamental frequency (F0) and the first two formants (F1 and F2) of the vowels in their speech by shifting their production in the opposite direction to achieve the target frequencies set for them by the experimenters (e.g., Burnett, Senner, & Larson, 1997; Houde & Jordan, 1998; Jones & Munhall, 2000). It has also been shown that the auditory cortex is selectively suppressed when speakers receive unaltered auditory feedback of their own voice, as opposed to when the feedback is distorted (*speaking-induced suppression*; e.g., Chang, Niziolek, Knight, Nagarajan, & Houde, 2013; Heinks-Maldonado, Nagarajan, & Houde, 2006). This finding suggests that the auditory cortex anticipates the effects of self-produced speech. Based on evidence from these studies, a case has been made for the existence of internal feed-forward models that predict and simulate auditory and somatosensory outcomes before speech execution, and trigger behavioral adaptation when feedback does not meet target expectations. However, in the present experiment, the mismatch alarm from these low-level mechanisms was ignored in favor of the contextual semantic-level inferences made by our participants. This highlights the problem of generalizing the architecture proposed by well-established models of motor loops to the level that concerns what speakers intend and decide to say (Hickok, 2012, 2014; Pickering & Garrod, 2013).

Our results are similarly problematic for the assumption that the articulatory speech plan can be monitored prior to the actual utterance through an internal channel (Levelt, 1989). The reliance on auditory feedback shown in our experiment suggests that either this postulated internal channel is unavailable during overt speech (as Huettig & Hartsuiker, 2010, and Nozari, Dell, & Schwartz, 2011, have speculated) or auditory feedback can override it.

Instead, the current results better fit an account of speech production in which speech intentions are seen as properties of the system as a whole, rather than originating from a dedicated black box "conceptualizer" buried at the heart of the model (Dennett, 1991). In this account, the meaning of an utterance is not fully internal to the speaker, but instead is partly determined by feedback from and inferences about the conversational context (Dennett, 1987, 1991; Linell, 2009). So, even though at some point in the speech process a particular word needs to be selected, and specific motor commands need to be issued to articulate this word, intentions can be underspecified with respect to the understanding of the speakers themselves. In our previous research on the phenomenon of choice blindness, we contributed evidence to the effect that knowing one's own attitudes is an inferential process, and that people cannot simply rely on introspection to determine why they choose to act the way they do (e.g., Hall, Johansson, & Strandberg, 2012; Hall et al., 2013; Johansson et al., 2005; Johansson, Hall, Tärning, Sikström, & Chater, 2013). The current study indicates that speech intentions can be regarded in a similar vein. Thus, our findings can be seen as a particularly striking demonstration of reconstructive rather than predictive authorship processing (e.g., Wegner & Wheatley, 1999; see also Lind et al., 2014, for further discussion of this issue).

Note that this alternative, inferential model does not deny that people can mentally rehearse actions (linguistic or otherwise) before execution, or that speakers sometimes might formulate very clear and detailed accounts of what to do next. Similarly, it does not deny that error correction exists. Many studies have detailed the different forms of self-correction that speakers engage in (e.g., Blackmer & Mitton, 1991; Seyfeddinipur, Kita, & Indefrey, 2008), and this is a phenomenon that any theory of speech production must explain. The dominant model emphasizes the internal criteria for error correction provided by the message formulated in the conceptualizer. Therefore, it predicts that participants will immediately detect words that are externally inserted, as in the current experiment. The alternative model instead puts the emphasis on external criteria for error correction: Taking into account their prior state and history, as well as the wider conversational context, speakers use general inferential processes to ensure that their utterances are successful, plausible, and error free. In the context of our experiment, this means that different sources of evidence regarding the meaning of each utterance were weighed in order to arrive at a conclusion about whether the inserted word was self-produced or not. In relation to the broader agency literature, this position is similar to a Bayesian, or cue-integration, account (Moore & Fletcher, 2012).

This is a backdrop to consider when evaluating the generalizability of the current study. If the experimental situation had not afforded at least minimal plausibility for different candidate utterances, then the two models' predictions regarding monitoring would have been the same. For example, if we had asked participants to name the object in an unambiguous picture of a cat, and we had replaced their answer with something phonologically similar but completely unrelated semantically (e.g., "mat"), then the alternative model would have predicted that participants would distrust the inserted word simply because it makes no sense whatsoever to say "mat" when asked to name a cat. However, we nevertheless have reason to assume that word insertions would be accepted in spontaneous speech as well, because in that context, there is no imposed standard of correctness, which creates far greater ambiguity and plausibility for different alternative utterances. To see this, compare the favorable conditions for monitoring when you are explicitly instructed to name the font color of a word displayed on a screen with the uncertainty you experience at a dinner party when trying to make a pithy interjection in a fluid discourse. The critical question for our investigation is the extent to which speakers rely on auditory feedback to specify the meaning of what they say in natural speech, when no helpful experimenters hang around to inform them about the exact need for self-monitoring, and when their speech acts are not accompanied by simultaneous forced-choice questions and reaction time measures that exhaustively probe their conscious knowledge.

In summary, the results from our real-time speech-exchange experiment indicate that speakers listen to their own voices to help specify the meaning of what they are saying. Thus, our results suggest that the sense of agency for speech has a strong inferential component, and that the meaning of spoken words is to be found in an interaction among the speaker, the listener, and the conversational context (see, e.g., Linell, 2009). In addition, our real-time speech-exchange method could be used to study cases in which aberrant feedback processing has been implicated, such as in aphasia or stuttering (Cai et al., 2012; Oomen, Postma, & Kolk, 2005), and to simulate auditory hallucinations in mentally ill and healthy individuals (Badcock & Hugdahl, 2012; Cahill, Silbersweig, & Frith, 1996). More generally, we believe that our results raise interesting questions about philosophical and psychological theories positing that the fundamental sense of self arises from comparator processes, and that people are perfectly aware of what they mean by their words before actually uttering them.

Author Contributions

A. Lind, L. Hall, and P. Johansson developed the study concept and wrote the manuscript. All the authors contributed to the study design. A. Lind performed testing and data collection. A. Lind, L. Hall, P. Johansson, and C. Balkenius performed the data analysis. B. Breidegard designed and implemented the real-time speech-exchange algorithm. All the authors approved the final version of the manuscript for submission.

Acknowledgments

We thank Anders Hulteen for constructing and building the headsets; Sverker Sikström for programming a previous version of the speech-exchange software; Johan Blomberg, Philip Pärnamets, Peter Gärdenfors, and Peter Kitzing for commenting on the manuscript; and Danilo Stankovic for assisting with Figure 1.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This work was supported by Uno Otterstedt's Foundation (Grant EKDO2010/54), the Crafoord Foundation (Grant 20101020), the Swedish Research Council, the Bank of Sweden Tercentenary Foundation, the Pufendorf Institute, and the European Union Goal-Leaders project (Grant FP7 270108).

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Note

1. The microphone-to-speaker system had a very low, and constant, latency of 8 ms.

References

Badcock, J. C., & Hugdahl, K. (2012). Cognitive mechanisms of auditory verbal hallucinations in psychotic and non-psychotic groups. *Neuroscience & Biobehavioral Reviews*, *36*, 431–438. doi:10.1016/j.neubiorev.2011.07.010

Blackmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, *39*, 173–194. doi:10.1016/0010-0277(91)90052-6

Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences*, *6*, 237–242. doi:10.1016/s1364-6613(02)01907-1

Burnett, T. A., Senner, J. E., & Larson, C. R. (1997). Voice F0 responses to pitch-shifted auditory feedback: A preliminary study. *Journal of Voice*, *11*, 202–211. doi:10.1016/s0892-1997(97)80079-3

Cahill, C., Silbersweig, D., & Frith, C. (1996). Psychotic experiences induced in deluded patients using distorted auditory feedback. *Cognitive Neuropsychiatry*, *1*, 201–211. doi:10.1080/135468096396505

Cai, S., Beal, D. S., Ghosh, S. S., Tiede, M. K., Guenther, F. H., & Perkell, J. S. (2012). Weak responses to auditory feedback perturbation during articulation in persons who stutter: Evidence for abnormal auditory-motor transformation. *PLoS ONE*, *7*(7), Article e41830. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0041830

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S., & Houde, J. F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences, USA*, *110*, 2653–2658. doi:10.1073/pnas.1216827110

David, N. (2012). New frontiers in the neuroscience of the sense of agency. *Frontiers in Human Neuroscience*, *6*, Article 161. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fnhum.2012.00161/full

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness explained*. Boston, MA: Little, Brown.

Frith, C. D. (2013). Action, agency and responsibility. *Neuropsychologia*, *55*, 137–142. doi:10.1016/j.neuropsychologia.2013.09.007

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, *4*, 14–21. doi:10.1016/s1364-6613(99)01417-5

Haggard, P., & Clark, S. (2003). Intentional action: Conscious experience and neural prediction. *Consciousness and Cognition*, *12*, 695–707. doi:10.1016/s1053-8100(03)00052-7

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLoS ONE*, *7*(9), Article e45457. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0045457

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLoS ONE*, *8*(4), Article e60554. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0060554

Heinks-Maldonado, T. H., Nagarajan, S. S., & Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *NeuroReport*, *17*, 1375–1379. doi:10.1097/01.wnr.0000233102.43526.e9

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13*, 135–145. doi:10.1038/nrn3158

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language and Cognitive Processes*, *29*, 2–20. doi:10.1080/01690965.2013.834370

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*, 1213–1216. doi:10.1126/science.279.5354.1213

Huettig, F., & Hartsuiker, R. J. (2010). Listening to yourself is like listening to others: External, but not internal, verbal self-monitoring is based on speech perception. *Language and Cognitive Processes*, *25*, 347–374. doi:10.1080/01690960903046926

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*, 116–119. doi:10.1126/science.1111709

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2013). Choice blindness and preference change: You will like this paper better if you (believe you) chose to read it! *Journal of Behavioural Decision Making*. Advance online publication. doi:10.1002/bdm.1807

Jones, J. J., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *Journal of the Acoustical Society of America*, *108*, 1246–1251. doi:10.1121/1.1288414

Kühn, S., Brass, M., & Haggard, P. (2012). Feeling in control: Neural correlates of experience of agency. *Cortex*, *49*, 1935–1942. doi:10.1016/j.cortex.2012.09.002

Levelt, W. J. M. (1989). *Speaking*. Cambridge, MA: MIT Press.

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014). Auditory feedback of one's own voice is used for high-level, semantic monitoring: The "self-comprehension" hypothesis. *Frontiers in Human Neuroscience*, *8*, Article 166. Retrieved from http://journal.frontiersin.org/Journal/10.3389/fnhum.2014.00166/full

Linell, P. (1982). The concept of phonological form and the activities of speech production and speech perception. *Journal of Phonetics*, *10*, 37–72.

Linell, P. (2009). *Rethinking language, mind and world dialogically*. Charlotte, NC: Information Age.

Moore, J. W., & Fletcher, P. C. (2012). Sense of agency in health and disease: A review of cue integration approaches. *Consciousness and Cognition*, *21*, 59–68. doi:10.1016/j.concog.2011.08.010

Moore, J. W., Wegner, D. M., & Haggard, P. (2009). Modulating the sense of agency with external cues. *Consciousness and Cognition*, *18*, 1056–1064. doi:10.1016/j.concog.2009.05.004

Nozari, N., Dell, G. S., & Schwartz, M. F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology*, *63*, 1–33. doi:10.1016/j.cogpsych.2011.05.001

Oomen, C. C. E., Postma, A., & Kolk, H. H. J. (2005). Speech monitoring in aphasia: Error detection and repair behaviour in a patient with Broca's aphasia. In R. J. Hartsuiker, R. Bastiaanse, A. Postma, & F. Wijnen (Eds.), *Phonological encoding and monitoring in normal and pathological speech* (pp. 209–225). Hove, England: Psychology Press.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension [Target article and commentaries]. *Behavioral & Brain Sciences*, *36*, 329–392. doi:10.1017/S0140525X12001495

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, *77*, 97–131. doi:10.1016/s0010-0277(00)00090-1

Seyfeddinipur, M., Kita, S., & Indefrey, P. (2008). How speakers interrupt themselves in managing problems in speaking: Evidence from self-repairs. *Cognition*, *108*, 837–842. doi:10.1016/j.cognition.2008.05.004

Sugimori, E., Asai, T., & Tanno, Y. (2013). The potential link between sense of agency and output monitoring over speech. *Consciousness and Cognition*, *22*, 360–374. doi:10.1016/j.concog.2012.07.010

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, *17*, 219–239. doi:10.1016/j.concog.2007.03.010

Wegner, D., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, *54*, 1–13. doi:10.1037//0003-066x.54.7.480

DOI: 10.1177/0956797614529797

Supplemental Online Material
Speakers' Acceptance of Real-Time Speech Exchange Indicates That We Use Auditory Feedback to Specify the Meaning of What We Say

Andreas Lind, Lars Hall, Björn Breidegard, Christian Balkenius & Petter Johansson

Methods

Technical setup

The specially constructed headset was customized by installing the transducers from Philips SHP 8900 into Howard Leight Leightning L1 hearing protection earmuffs, and by mounting a König omnidirectional microphone on the headset. The headset was connected to a Behringer AMP 800 headphone amplifier and to a Sound Blaster X-Fi Titanium sound card installed in a PC computer (Windows 7 Professional, Pentium Dual-core E6800, 3.33 GHz, from year 2011) which executed the specially designed control application (including the voice-exchange algorithms). The same sound level was used for all participants, 8-10 dB above normal speaking level. This is somewhat louder than the feedback we regularly receive from our own voice, but without sounding unnaturally loud to the participants. The increase in loudness served the function of masking the air-conducted sound of the participant's own voice that may leak through the earmuffs. The sound signal was also low-pass filtered to make the voice feedback more natural (e.g. Shuster & Durrant, 2003; Heinks-Maldonado, Nagarajan & Houde, 2006).

Procedure

The experimenter left the experiment room before the test started. Before he left, the participants were told they could stop the test whenever they wished to, in case they felt any discomfort. They were then asked to read through a sheet of instructions for the test that were identical to the verbal explanation already given, and, when feeling ready, to start the test by pressing a push-button in front of them.

During the test, the experimenter was seated in a hidden control room adjacent to the recording studio. Using a Logitech Precision gamepad, he controlled the auditory feedback of the participants. Recorded words were trimmed automatically so that the onset of the recording matched the onset of the word. Judgments of the precision of the trimming were made on a waveform display and by listening to the word, with the possibility of further manual trimming. Using the gamepad, the experimenter could then enable the voice trigger just before the participant vocalized at the manipulated trial, and once the participant produced a sound above a certain level, the readied manipulation word was inserted and feedback of the actually uttered word was blocked.

Twenty-five different word-color combinations of the words/colors "blue," "red," "grey," "brown" and "green" were used, each combination appearing 10 times in the experiment. The words that were used during manipulations were recorded in as close proximity to each manipulated trial as possible, to maximize similarity in pronunciation. The manipulated trials were inserted on trials 18, 35, 95 and 136. The four non-manipulated stop trials were included at trials 61, 104, 159 and 189. In total, there were 312 planned manipulations across all participants.

The experiment was recorded to a two channel wav-format sound file. We used Praat (Boersma & Weenink, 2011) to make precise retrospective timing measurements, comparing the timing of the inserted words with that of the word the participant was actually uttering.

Posttest interview

The questions asked during the posttest interview were: 1) “what did you think of the test in general?,” 2) “did you make many mistakes?,” 3) “what was it like to answer the questions that popped up now and then?,” 4) “did you notice anything strange?” After this, we 5) revealed that we had manipulated their feedback and asked them if they had noticed this. If the participants at any point during the posttest interview indicated that they had detected a switch, this trial was registered as a detection.

Data Analysis

The timing window

We have no exact benchmark measures for how much timing mismatch can be allowed before the participants readily detect the word substitution solely on the basis of the mismatch. Previous studies on delayed auditory feedback using a different paradigm than whole word substitution have shown that delays of 30 ms or more can disrupt speech (Lee, 1950; Takaso, Einer, Wise & Scott, 2010). Research on echo location thresholds has shown that in an explicit detection task, people are capable of detecting a delay as low as 5-10 ms (Krumbholz & Nobbe, 2002; Backer, Hill, Shahin & Miller, 2010). As both of these paradigms show some similarities to our setup, but neither being fully comparable to it, we therefore decided a priori to focus on trials falling within a timing window of 5-20 ms, and to compare the detection rate of trials within this window with trials with larger timing mismatch. The chosen window of 5-20 ms in the current experiment includes the 8 ms delay in the system, which means that the mismatch between the delay the participants have habituated to (8 ms) and the maximum delay of inserted segments is 12 ms. The window thus includes all trials down to -3 ms, in order to allow for measurement inaccuracies. Of a total 284 successful manipulations, 175 (61.6%) were within the timing window.

While there were significantly fewer detected manipulations among trials within the timing window (Fig. 2), we did not find any differences when comparing trials with a negative timing mismatch (below 5 ms) and trials with a positive timing mismatch (above 20 ms), neither for detection, $\chi^2(1, N = 63) = 0.0072$, $p = .9326$, nor for trials accepted as self-produced, $\chi^2(1, N = 63) = 2e\text{-}04$, $p = .9897$. The classification of the non-detected trials that were within the timing window in relation to the participants’ responses in the Stroop test was 26 trials in category A; 11 trials in category B; 20 trials in category C; and 6 trials in category D (see table 1).

There were no differences in detection rate in relation to gender, $\chi^2(1, N = 155) = 0.004$, $p = .95$).

References

Backer, K. C., Hill, K. T., Shahin, A. J. & Miller, L. M. (2010). Neural time course of echo suppression in humans. *Journal of Neuroscience, 30***,** 1905-1913.

Boersma, P. & Weenink, D. (2011). Praat: Doing Phonetics by Computer (University of Amsterdam), Version 5.3.02.

Heinks-Maldonado, T. H., Nagarajan, S. S. & Houde, J. F. (2006). Magnetoencephalographic evidence for a precise forward model in speech production. *Neuroreport, 17,* 1375-1379.

Krumbholz, K. & Nobbe, A. (2002). Buildup and breakdown of echo suppression for stimuli presented over headphones-the effects of interaural time and level differences. *Journal of the Acoustical Society of America, 112,* 654-663.

Lee, B. S. (1950). Effects of delayed speech feedback. *Journal of the Acoustical Society of America, 22,* 824-826.

Shuster, L.I. & Durrant, J.D. (2003). Toward a better understanding of the perception of self-produced speech. *Journal of Communication Disorders, 36,* 1-11.

Takaso, H., Eisner, F., Wise, R. J. S. & Scott, S. K. (2010). The effect of delayed auditory feedback on activity in the temporal lobe while speaking: a positron emission tomography study. *Journal of Speech, Language and Hearing Research, 53,* 226-236.