Commentary

Psychological Science
1–3

DOI: 10.1177/0956797615599341
pss.sagepub.com

Auditory Feedback Is Used for Self-Comprehension: When We Hear Ourselves Saying Something Other Than What We Said, We Believe We Said What We Hear

Andreas Lind[1,2], Lars Hall[1], Björn Breidegard[3], Christian Balkenius[1], and Petter Johansson[1,4]

[1]Lund University Cognitive Science, Lund University; [2]IRCAM (Institut de Recherche et Coordination Acoustique/Musique), STMS CNRS UMR9912, Paris, France; [3]Certec, Division of Rehabilitation Engineering Research, Department of Design Sciences, Faculty of Engineering, Lund University; and [4]Swedish Collegium for Advanced Study, Linneanum, Uppsala University

Received 12/11/14; Revision accepted 7/16/15

What is the nature of speech intentions? In a previous *Psychological Science* article (Lind, Hall, Breidegard, Balkenius, & Johansson, 2014b), we introduced *real-time speech exchange* (RSE), a technique that allows researchers to covertly manipulate auditory feedback so that participants say one thing but hear themselves saying something else. We used this technique to test contrasting predictions from *comparator* models of speech production, which assume clear preverbal intentions as benchmarks for feedback monitoring, and *inferential* models, according to which speakers use auditory feedback more actively to help determine the meaning of their utterances (e.g., Dennett, 1991; Hickok, 2012, 2014; Levelt, 1989; Linell, 2009; see also Lind, Hall, Breidegard, Balkenius, & Johansson, 2014a). Meekings et al. (2015) think our study is a timely attempt to tackle the core issue of speech intentions, but they present three critiques of our conclusions. We address each in order.

First, Meekings et al. disregarded our exclusion criteria and recalculated the frequency of detection of the manipulation, arriving at a figure of 73%. On the basis of that number, they conclude that auditory feedback is unlikely to be a prime mechanism for self-comprehension. As our study was the first exploration of a new phenomenon, there was no objective standard regarding what to report, and we welcome any discussion about the criteria that are relevant to apply when describing RSE data. Unfortunately, however, Meekings et al. present no arguments for ignoring our calculations.

We excluded data from trials following detection of the manipulation because following such detection, participants actively searched for manipulations, and our aim was to investigate the *everyday use* of, not the *maximum capacity* for, self-monitoring. As we put it in the article, "the critical question for our investigation is the extent to which speakers rely on auditory feedback . . . in natural speech, when no helpful experimenters hang around to inform them about the exact need for self-monitoring" (p. 1203). If we had believed that it would be nearly impossible to detect the manipulations, then we would have modified the task to examine explicit error detection, telling participants about the exchanges and asking them to report what they actually said. But this would have measured maximum capacity under optimal conditions, and our results would not have supported any interesting generalizations to everyday language use. It makes little sense to analyze data from trials that came after participants had been alerted to the manipulation, as both comparator and inferential models would predict detection of the manipulation under such circumstances.

Similarly, we emphasized data from those trials in which the exchange fell within a specified timing window. Again, we fail to see why Meekings et al. ignored this timing criterion, as neither the comparator nor the inferential perspective would predict acceptance of the manipulated feedback when it arrives considerably out

Corresponding Author:
Andreas Lind, Lund University Cognitive Science, Lund University, LUX, Helgonavägen 3, Hus B, 223 62 Lund, Sweden
E-mail: andreas.lind@lucs.lu.se

of sync with actual speech. As we highlighted in Figure 2 of our article, the assumption that mistimed exchanges would lead to an artificial increase in detection was confirmed by the data: Exchanges outside the timing window were nearly twice as likely to be detected by the participants compared with exchanges inside the timing window. Thus, when the predefined conditions of the experiment were met, no more than 32% of the exchanges were detected.

Naturally, there are other possible ways of reporting the data. To avoid the issue of exclusion altogether, we could have reported the data from the first manipulated trial only, in which case we would have had a detection frequency of 38% (with 50% of these exchanges being mistimed). The reason we did not focus exclusively on these trials was that doing so would have removed interesting data about repeated nondetections without substantially altering detection frequency. We contend that what is important in the context of our experiment is not the exact percentages, but the basic finding that a large percentage of the participants failed to detect that their speech had been replaced by something they had said earlier. In future studies, we will aim for larger sample sizes so that we can analyze both the effects of individual differences in monitoring capacity and contextual factors that might contribute to detection levels.

Second, Meekings et al. argue that the feedback we used in the experiment differed from the usual experience of hearing one's own voice, and that this invalidates our conclusions. Our RSE software did not simulate the perceived spatial location of self-speech, and we made no attempt to eliminate information from somatosensory and bone-conduction pathways. But this objection misses the point of the experiment. Our aim was not to distinguish between possible effects of self-produced and other-produced speech, but rather to test the comparator and inferential perspectives' contrasting predictions as to whether the auditory feedback would be accepted as self-produced. The fact that other-produced speech sometimes influences self-perception, as noted by Meekings et al., is not a problem for the inferential hypothesis, but rather is part of the backdrop that inspires and supports it.

In everyday life, people do not indiscriminately assimilate all surrounding speech as their own, and both the inferential and the comparator perspectives acknowledge that adaptive error correction exists. Thus, had we used voice distortion that made the exchanged words sound like they were produced by a growling monster, and had the participants nevertheless accepted these wildly improbable insertions as being self-produced, this would have been a critical problem for *all* speech-production theories of which we are aware, the inferential model included. But within the restricted context of our experiment, the inferential model makes no principled distinction between a perfect simulation of a participant's own voice and a reasonable estimate such as we used (Shuster & Durrant, 2003; see also Reinfeldt, Östli, Håkansson, & Stenfelt, 2010). Had we managed to include bone conduction and 3-D localization, then presumably the feedback would have been even more convincing, and the manipulation would have been detected on even fewer trials. But as our results show, these features were not needed to create a plausible manipulation. The identity of the voice used in the feedback is an interesting dimension to explore, but fully controlling this identity is not critical to our conclusions. We are happy to conclude that speakers listen to their own speech, or any other person's speech if it is of sufficient contextual plausibility, to help specify the meaning of what they say.

Third, Meekings et al. suspect that, because of the special executive demands of the Stroop task, our results will not generalize to natural interactions. They write: "Stroop interference results from competition between the color of the text and the distractor. . . . Both are automatically processed and prepared for response production, and executive-control systems are required for the final response selection" (p. XXX). But whereas Meekings et al. see the Stroop task as a critical anomaly, we see it as capturing the underlying structure of speech production. Generativity in speech has to be accounted for by competition and interference, and the beauty of the RSE methodology is that it can test how selection is accomplished by the black-box executive-control systems that are taken for granted by the dominant comparator model (Dennett, 1991).

We agree with Meekings et al. that the Stroop task is problematic, but for the exactly opposite reason: It is unnaturally easy to self-monitor in this setting. This is so because (a) the task induces participants to be more vigilant regarding their performance than they are in everyday discourse; (b) there is an objective standard of correctness in the task, which is seldom the case for everyday speech; (c) participants have a visual short-term memory of this standard, as it has been displayed on the computer screen; and (d) participants learn after a long series of correct answers that it is highly unlikely for them to err. In our experiment, to satisfy the technical demands of RSE, we had to trade off naturalness of speech for the predictability and regularity of the response, but in future studies, in order to satisfy both ourselves and Meekings et al., we will aim to find more ecologically valid tasks to which we can apply RSE (see Lind et al., 2014a).

Author Contributions

A. Lind, L. Hall, and P. Johansson drafted the manuscript, and B. Breidegard and C. Balkenius provided critical revisions. All authors approved the final version of the manuscript for submission.

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 Bank of Sweden Tercentenary Foundation (Grant P13-1059:1), the Swedish Research Council (Grants 2011-1795 and 2014-1371), the Pufendorf Institute, and the European Union Goal-Leaders project (Grant FP7 270108).

References

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

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

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

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014a). 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/article/10.3389/fnhum.2014.00166/full

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014b). Speakers' acceptance of real-time speech exchanges indicates that we use auditory feedback to specify the meaning of what we say. *Psychological Science*, *25*, 1198–1205. doi:10.1177/0956797614529797

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

Meekings, S., Boebinger, D., Evans, S., Lima, C. F., Chen, S., Ostarek, M., & Scott, S. K. (2015). Do we know what we're saying? The roles of attention and sensory information during speech production. *Psychological Science*, *26*, XXX–XXX.

Reinfeldt, S., Östli, P., Håkansson, B., & Stenfelt, S. (2010). Hearing one's own voice during phoneme vocalization—transmission by air and bone conduction. *Journal of the Acoustical Society of America*, *128*, 751–762. doi:10.1121/1.3458855

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. doi:10.1016/S0021-9924(02)00132-6