AUTHOR QUERY FORM

AIP Publishing	**Journal: J. Acoust. Soc. Am.** **Article Number: JASA-12954**	**Please provide your responses and any corrections by annotating this PDF and uploading it according to the instructions provided in the proof notification email.**

Dear Author,

Below are the queries associated with your article; please answer all of these queries before sending the proof back to AIP.

Article checklist: In order to ensure greater accuracy, please check the following and make all necessary corrections before returning your proof.

1. Is the title of your article accurate and spelled correctly?
2. Please check affiliations including spelling, completeness, and correct linking to authors.
3. Did you remember to include acknowledgment of funding, if required, and is it accurate?

Location in article	Query / Remark: click on the Q link to navigate to the appropriate spot in the proof. There, insert your comments as a PDF annotation.
AQ1	Please check that the author names are in the proper order and spelled correctly. Also, please ensure that each author's given and surnames have been correctly identified (given names are highlighted in red and surnames appear in blue).
AQ2	Please verify the single-page article in Ref(s). Tajadura-Jimenez *et al.* (2017); otherwise, please provide the full page rangef.

Thank you for your assistance.

J_ID: JASMAN DOI: 10.1121/10.0039927 Date: 11-November-25 Stage: Page: 1 Total Pages: 4
PROOF COPY [JASA-12954]

JASA LETTER TO THE EDITOR

Beyond acoustics: Self-relevance as a key to voice naturalness (L)

AQ1 Ana P. Pinheiro[a)]
Faculdade de Psicologia, CICPSI, Universidade de Lisboa, Lisboa, Portugal

ABSTRACT:
Synthetic voices can now achieve remarkable acoustic accuracy, yet often fail to sound "natural," especially when designed to reproduce one's own voice. Existing frameworks define naturalness along two dimensions: deviation from acoustic norms and human-likeness. Yet these dimensions overlook the self-voice, which can feel natural or unnatural for reasons beyond the signal itself. Here, self-relevance is proposed as a complementary dimension, capturing the subjective alignment between a voice and the listener's self-representation. Evidence shows that self-relevance modulates perceived naturalness independently of acoustic match. A full understanding of voice naturalness, therefore, requires integrating physical speech properties with the listener's self-representational framework.
 https://doi.org/10.1121/10.0039927
(Received 28 August 2025; revised 30 October 2025; accepted 3 November 2025; published online xx xx xxxx)
[Editor: Jody Kreiman] Pages: 1–3

What makes a voice sound *natural*? For decades, acoustics has provided the primary explanation: naturalness is tied to how closely a signal conforms to expected patterns in the speech spectrum. Today, synthetic voices can achieve a near-perfect acoustic match to natural speech, yet still sound subtly "off" to listeners, often judged less favorably (Gessinger *et al.*, 2023; Herrmann, 2023) and, in particular, less socially appealing (Bruder *et al.*, 2025) than human voices. This paradox reveals that naturalness cannot be reduced to acoustics alone.

A recent framework (Nussbaum *et al.*, 2025) defines naturalness along two main acoustic-perceptual dimensions: *deviation-based naturalness* captures how much a voice diverges from typical acoustic patterns, whereas *human-likeness-based naturalness* reflects the extent to which a voice sounds recognizably human. In speech synthesis research, for example, naturalness is typically evaluated through perceptual methods such as the absolute category rating (ACR) and its outcome, the mean opinion score (MOS), which reflects the perceived closeness of synthetic to natural speech (Le Maguer *et al.*, 2024; Perrotin *et al.*, 2025). However, while the two dimensions explain why many artificial or degraded voices sound strange or unnatural, they overlook the self-voice, which can feel natural or unnatural for reasons extending beyond the signal itself. Here, I propose a complementary dimension of naturalness: self-relevance.

Whereas the *self-voice* refers to the acoustic stimulus associated with one's own vocal identity, *self-relevance* captures the subjective alignment between that voice and the listener's internal model of the self (Crow *et al.*, 2021). It is shaped by sensorimotor contingencies (Khalilian-Gourtani *et al.*, 2024), memory-based self-representations (Iannotti *et al.*, 2022), and multisensory integration (Orepic *et al.*, 2023). Crucially, it is not reducible to *self-recognition*, i.e., the perceptual identification of a voice as one's own (Candini *et al.*, 2014), or self-attribution, i.e., the inference that one is the source of the voice (Ohata *et al.*, 2022). Rather, it concerns whether a voice feels *natural for me*, even when acoustically atypical or externally generated. For instance, a hoarse version of one's own voice during illness may sound more natural than a flawless synthetic clone.

Evidence supports treating self-relevance as separate from acoustic accuracy. Voices that deviate from typical acoustic features or patterns can still be judged as "natural" or "self-like" when they align with memory-based self-representations or multisensory predictions (Tajadura-Jiménez *et al.*, 2017), sometimes even altering self-perception (Tajadura-Jiménez *et al.*, 2017), mood (Aucouturier *et al.*, 2016), or social attitudes (Arakawa *et al.*, 2021). Conversely, acoustically unaltered voice feedback can feel "alien" or externally generated if it violates sensorimotor contingencies (Franken *et al.*, 2018) or lacks expected bone-conducted vibrotactile cues (Orepic *et al.*, 2023). Recent work (Rosi *et al.*, 2025b) further illustrates this complexity: participants sometimes rated cloned voices of others more favourably than their own, despite equivalent acoustic deviation. Such findings underscore that acoustics is necessary but not sufficient for judging a voice as *naturally "mine."*

Neural evidence further supports the role of self-relevance. Like deviation and human-likeness, it modulates early auditory processing within ~200 ms after voice onset, consistent with rapid sensory gain for self-relevant input (Pinheiro *et al.*, 2016; Pinheiro *et al.*, 2023). These effects suggest that the brain treats the self-voice as a special category beyond simple acoustic analysis (e.g., Conde *et al.*, 2016, 2018). Its influence extends to later processing stages,

[a)]Email: appinheiro@psicologia.ulisboa.pt

when voice input is evaluated against memory-based self-representations (Iannotti *et al.*, 2022; Pinheiro *et al.*, 2023). Therefore, self-relevance bridges (early) low-level auditory and (later) higher-level cognitive analyses of voice information. It may act as a contextual modulation of voice naturalness, explaining why some voice distortions feel authentic or why even acoustically matched voice clones can fail to sound "natural."

Recognizing the limits of traditional evaluation criteria (e.g., MOS; Le Maguer *et al.*, 2024), recent developments in speech synthesis have reconceptualized naturalness as *appropriateness*—the idea that synthetic speech should be judged within its communicative context rather than in isolation (Pandey *et al.*, 2025). This perspective aligns with evidence showing that hearing is dynamically shaped by perceptual, cognitive, social, and emotional context (Kreiman, 2024). Within this framework, *self-voice evaluation* represents a specific form of contextual appropriateness, particularly relevant when synthesized voices aim to express or preserve the user's own vocal identity. Such cases include personalized text-to-speech systems, clinical voice restoration (e.g., for individuals with amyotrophic lateral sclerosis or after laryngectomy), and personalized human-AI interfaces (e.g., McGettigan *et al.*, 2024; Rosi *et al.*, 2025a). Conversely, in contexts such as virtual assistants or public service announcements—where the listener is not the voice's owner—self-relevance becomes secondary to task-related or social appropriateness.

Understanding when and how self-relevance contributes to perceived naturalness has direct implications for both clinical and technological domains. Altered self-voice feedback has been linked to auditory verbal hallucinations (Pinheiro *et al.*, 2020), where acoustically realistic voices are experienced as alien, not because they sound unnatural, but because they lack the perceptual and neural signatures of self-generation (e.g., Pinheiro *et al.*, 2020). In technological contexts, while users often express positive attitudes toward digital voice transformations (Guerouaou *et al.*, 2024), these technologies can have unintended psychological consequences, revealing the limits of acoustics as a predictor of perceived naturalness. For example, voice cloning may destabilize a speaker's own sense of vocal identity, consistent with evidence linking how we sound to who we believe ourselves to be (Stern *et al.*, 2021).

To conclude, acoustics is essential but not sufficient for understanding voice naturalness. Voice perception—including that of synthetic voices—depends not only on the physical signal but also on cognitive and contextual factors that shape its interpretation (Kreiman, 2024). Naturalness, though still a loosely defined, multifaceted perceptual construct (Pandey *et al.*, 2025), must therefore be understood as emerging from the interaction between the acoustic properties of speech and the listener's representational framework. Introducing self-relevance as a third dimension highlights the perceptual uniqueness of the self-voice and opens a novel research agenda at the intersection of acoustics, self-representation, and artificial intelligence—one that can inform both scientific understanding and responsible technological design. For speech synthesis, this means expanding beyond signal fidelity and human-likeness toward perceptual congruence with the listener's internal model of the self, while developing evaluation methods that explicitly capture this self-relevant dimension. In a world where the line between genuine and synthetic speech grows even thinner, accounting for self-relevance will be essential for understanding and preserving what makes a voice truly natural.

ACKNOWLEDGMENTS

This work was supported by Fundação para a Ciência e Tecnologia and BIAL Foundation (Grant Nos. 2023.00041. RESTART and BIAL 146/2020). The author would like to thank the VoicES Lab members (www.voicesneurolab.com) for the fruitful discussions about voice naturalness and its underlying neural and functional mechanisms.

AUTHOR DECLARATIONS

Conflict of Interest

The author has no conflicts to disclose.

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Arakawa, R., Kashino, Z., Takamichi, S., Verhulst, A., and Inami, M. (**2021**). "Digital speech makeup: Voice conversion based altered auditory feedback for transforming self-representation," in *Proceedings of the 2021 International Conference on Multimodal Interact*, pp. 159–167.

Aucouturier, J.-J., Johansson, P., Hall, L., Segnini, R., Mercadié, L., and Watanabe, K. (**2016**). "Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction," PNAS **113**, 948–953.

Bruder, C., Breda, P., and Larrouy-Maestri, P. (**2025**). "Attractive synthetic voices," Comput. Hum. Behav. Artif. Hum. **6**, 100211.

Candini, M., Zamagni, E., Nuzzo, A., Ruotolo, F., Iachini, T., and Frassinetti, F. (**2014**). "Who is speaking? Implicit and explicit self and other voice recognition," Brain Cogn. **92**, 112–117.

Conde, T., Gonçalves, Ó. F., and Pinheiro, A. P. (**2016**). "The effects of stimulus complexity on the preattentive processing of self-generated and nonself voices: An ERP study," Cogn. Affect. Behav. Neurosci. **16**, 106–123.

Conde, T., Gonçalves, Ó. F., and Pinheiro, A. P. (**2018**). "Stimulus complexity matters when you hear your own voice: Attention effects on self-generated voice processing," Int. J. Psychophysiol. **133**, 66–78.

Crow, K. M., van Mersbergen, M., and Payne, A. E. (**2021**). "Vocal congruence: The voice and the self measured by interoceptive awareness," J. Voice **35**, 324.e15–324.e28.

Franken, M. K., Acheson, D. J., McQueen, J. M., Hagoort, P., and Eisner, F. (**2018**). "Opposing and following responses in sensorimotor speech control: Why responses go both ways," Psychon. Bull. Rev. **25**, 1458–1467.

Gessinger, I., Cohn, M., Cowan, B. R., Zellou, G., and Möbius, B. (**2023**). "Cross-linguistic emotion perception in human and TTS voices," in *Proceedings Interspeech 2023*, pp. 5222–5226.

Guerouaou, N., Vaiva, G., and Aucouturier, J.-J. (**2024**). "Social affective inferences in the era of AI-filters: Towards the Bayesian reshaping of human sociality?," HAL-04801092 (HAL, Lyon, France).

Herrmann, B. (**2023**). "The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults," Int. J. Speech Technol. **26**, 395–415.

JASA
https://doi.org/10.1121/10.0039927

Iannotti, G. R., Orepic, P., Brunet, D., Koenig, T., Alcoba-Banqueri, S., Garin, D. F. A., Schaller, K., Blanke, O., and Michel, C. M. (**2022**). "EEG spatiotemporal patterns underlying self-other voice discrimination," Cereb. Cortex **32**, 1978–1992.

Khalilian-Gourtani, A., Wang, R., Chen, X., Yu, L., Dugan, P., Friedman, D., Doyle, W., Devinsky, O., Wang, Y., and Flinker, A. (**2024**). "A corollary discharge circuit in human speech," Proc. Natl. Acad. Sci. U.S.A. **121**, e2404121121.

Kreiman, J. (**2024**). "Information conveyed by voice quality," J. Acoust. Soc. Am. **155**, 1264–1271.

Le Maguer, S., King, S., and Harte, N. (**2024**). "The limits of the Mean Opinion Score for speech synthesis evaluation," Comput. Speech Lang. **84**, 101577.

McGettigan, C., Bloch, S., Rosi, V., Dinkar, T., Lavan, N., Bowles, C., and Reus, J. (**2024**). "Voice cloning: Psychological and ethical implications of intentionally synthesising familiar voice identities".

Nussbaum, C., Frühholz, S., and Schweinberger, S. R. (**2025**). "Understanding voice naturalness," Trends Cogn. Sci. **29**, 467–480.

Ohata, R., Asai, T., Imaizumi, S., and Imamizu, H. (**2022**). "I hear my voice; therefore I spoke: The sense of agency over speech is enhanced by hearing one's own voice," Psychol. Sci. **33**, 1226–1239.

Orepic, P., Kannape, O. A., Faivre, N., and Blanke, O. (**2023**). "Bone conduction facilitates self-other voice discrimination," R Soc. Open Sci. **10**, 221561.

Pandey, A., Le Maguer, S., and Harte, N. (**2025**). "What is naturalness?," in *Proceedings of the 13th edition of the Speech Synthesis Workshop*, pp. 215–221.

Perrotin, O., Stephenson, B., Gerber, S., Bailly, G., and King, S. (**2025**). "Refining the evaluation of speech synthesis: A summary of the Blizzard Challenge 2023," Comput. Speech Lang. **90**, 101747.

Pinheiro, A. P., Rezaii, N., Nestor, P. G., Rauber, A., Spencer, K. M., and Niznikiewicz, M. (**2016**). "Did you or I say pretty, rude or brief? An ERP study of the effects of speaker's identity on emotional word processing," Brain Lang. **153-154**, 38–49.

Pinheiro, A. P., Sarzedas, J., Roberto, M. S., and Kotz, S. A. (**2023**). "Attention and emotion shape self-voice prioritization in speech processing," Cortex **158**, 83–95.

Pinheiro, A. P., Schwartze, M., Amorim, M., Coentre, R., Levy, P., and Kotz, S. A. (**2020**). "Changes in motor preparation affect the sensory consequences of voice production in voice hearers," Neuropsychologia **146**, 107531.

Rosi, V., Payne, B., and McGettigan, C. (**2025a**). "Effects of self-similarity and self-generation on the perceptual prioritization of voices," J. Exp. Psychol. Hum. Percept. Perform. **51**, 996–1007.

Rosi, V., Soopramanien, E., and McGettigan, C. (**2025b**). "Perception and social evaluation of cloned and recorded voices: Effects of familiarity and self-relevance," Comput. Hum. Behav. Artif. Hum. **4**, 100143.

Stern, J., Schild, C., Jones, B. C., DeBruine, L. M., Hahn, A., Puts, D. A., Zettler, I., Kordsmeyer, T., Feinberg, D., Zamfir, D., Penke, L., and Arslan, R. C. (**2021**). "Do voices carry valid information about a speaker's personality?," J. Res. Personal. **92**, 104092.

Tajadura-Jiménez, A., Banakou, D., Bianchi-Berthouze, N., and Slater, M. (**2017**). "Embodiment in a child-like talking virtual body influences object size perception, self-identification, and subsequent real speaking," Sci. Rep. **7**, 9637.

AQ2