UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Linguists should learn to love speech-based deep learning models

de Heer Kloots, M.; Boersma, P.; Zuidema, W.

DOI
10.1017/S0140525X26104713
10.48550/arXiv.2512.14506

Publication date
2026

Document Version
Other version

Published in
Behavioral and Brain Sciences

License
CC BY

Link to publication

Citation for published version (APA):
de Heer Kloots, M., Boersma, P., & Zuidema, W. (2026). Linguists should learn to love speech-based deep learning models. *Behavioral and Brain Sciences*, *49*, Article e206. https://doi.org/10.1017/S0140525X26104713, https://doi.org/10.48550/arXiv.2512.14506

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:11 Oct 2026

Linguists should learn to love speech-based deep learning models

Marianne de Heer Kloots[1], Paul Boersma[2], Willem Zuidema[1]

[1] Institute for Logic, Language and Computation, University of Amsterdam
[2] Amsterdam Center for Language and Communication, University of Amsterdam

Correspondence: m.l.s.deheerkloots@uva.nl

Commentary on Futrell, R. & Mahowald, K. (in press). How Linguistics Learned to Stop Worrying and Love the Language Models. *Behavioural and Brain Sciences*.
http://doi.org/10.1017/S0140525X2510112X

Abstract
Futrell and Mahowald present a useful framework bridging technology-oriented deep learning systems and explanation-oriented linguistic theories. Unfortunately, the target article's focus on generative text-based LLMs fundamentally limits fruitful interactions with linguistics, as many interesting questions on human language fall outside what is captured by written text. We argue that audio-based deep learning models can and should play a crucial role.

Main text
To integratively study a human spoken language, linguists investigate how speakers map their communicative intent all the way to their articulatory gestures (language production), as well as how listeners map incoming auditory signals to their interpretation of a speaker's intent (language comprehension). An inclusive linguistics is therefore majorly concerned with structural relations between physical signals and linguistic content — relations typically studied by the fields of *phonetics* (audition, acoustics, articulation) and *phonology* (sound structure). As these fields are largely irrelevant to text-based LLMs, F&M don't address them. This jeopardizes their own endeavour.

The trouble with text. For language comprehension, LLMs operate on text that has already discretized the continuous auditory speech stream into things that look like words, morphemes, syllables, and/or phonological segments (vowels, consonants), depending on the language. Conversely, crucial phonological aspects of spoken language like prosody and intonation aren't usually represented in text. As a result, ambiguities differ between text and speech: English text but not speech distinguishes between *sun* and *son*, while English speech but not text disambiguates the polar vs. alternative readings of *Do you like coffee or tea?* In general, text-based models end up solving fundamentally different problems than human spoken language users, as both cognitive scientists and language technologists have noted before (Dupoux, 2018; Chrupała, 2023).

The structures of natural languages reflect overwhelmingly the properties of speech (or signing), rather than those of text. Merely enriching text-based LLMs with speech capacities through acoustic tokens or separate speech recognition/synthesis components (Arora et al., 2025) cannot resolve the deep trouble caused by textual bottlenecks in linguistic modelling. Researchers interested in modelling linguistic structure should therefore not settle for LLMs, but rather learn to love models designed to capture the more natural form of human spoken language: the speech signal itself.

Linguistic structure in models of speech. Many current speech-technological applications employ self-supervised *speech foundation models*, i.e. deep learning architectures trained to represent audio signals on the basis of unlabelled speech recordings. Parallel to work on the interpretability of text-based LLMs (as covered by F&M), *linguistic interpretability* studies in the speech domain investigate to what extent speech-based models learn to capture any higher-level patterns that make up spoken language. Method-wise, most of these studies use *representational probes* to examine what linguistic

information is represented in speech models' internal states, with insightful results obtained for the encoding of linguistic units like phonemes (Alishahi, Barking, & Chrupała, 2016; Martin et al., 2023) and words (Pasad et al., 2024), as well as morphophonological (Gauthier et al., 2025) and suprasegmental patterns (Shen et al., 2024); other studies use *behavioural (minimal pair) tests* to ask e.g. whether models can distinguish words from pseudowords (Lavechin et al., 2023) or prosodically natural vs. unnatural pauses (de Seyssel et al., 2023). With these types of methods, speech models can be used as "psycholinguistic participants", mimicking human speech perception experiments and their results, e.g. about perceptual similarity (Millet & Dunbar, 2022) and phonetic categorization (de Heer Kloots & Zuidema, 2024).

Linguists can already draw an important theoretical insight from these findings: speech-only models can indeed learn relevant patterns of linguistic structure, without the pre-existing symbolic categories assumed by earlier connectionist work (McClelland & Elman, 1986; Smolensky, 1990, 1999).

Inductive biases from domain-general perception. Further investigations of what deep learning models learn from audio can potentially inform what mechanisms drive language-relevant perceptual learning in humans. Here, we agree with F&M that deep learning models can helpfully serve as proof-of-concepts for ideas that were previously hard to formalize. At least part of the human perceptual system involved in speech processing is also involved in processing a wide variety of other auditory signals. Deep learning models pre-trained on music and/or environmental sounds show more human-like behaviour than models trained on speech sounds alone, in detecting algebraic patterns in tone or syllable sequences (Orhan, Boubenec, & King, 2025), and in displaying native language effects when processing foreign speech sounds (Poli et al., 2024). Hence, it seems that some inductive biases relevant to the encoding of language-like structures can result from (evolutionarily) optimizing the perceptual system for representing sounds other than speech, perhaps more so than considered before (Aslin & Pisoni, 1980; Soderstrom, Mathis, & Smolensky, 2006).

Inductive biases from bidirectional processing. One property of much cognitively grounded neural modelling, that is not shared with current technological speech or text machines, is that connection weights are symmetric (Kohonen, 1982; Hopfield, 1982; McMurray et al., 2009; Salakhutdinov & Hinton, 2009). In linguistics this corresponds to a specific inductive bias, namely that language employs *bidirectional processing*, i.e. language comprehension and language production utilize the same knowledge.

In phonetics and phonology, neural models with bidirectional connection weights straightforwardly predict the diachronic evolution of auditory dispersion in phoneme inventories (Boersma, Benders & Seinhorst, 2020). Unpublished simulations show that such communicative-success-optimizing effects readily transfer to other linguistic subdomains: with bidirectional connection weights, the maxims of Grice (1975) emerge in pragmatics, as do anti-synonymity effects in semantics, morphology and syntax (Boersma, 2009).

A drawback of small neural models is that they work exclusively on *toy problems*, rather than *whole languages*. Hopefully evidence from toy models can nevertheless inspire speech technologists in architecture design. Most findings cited above were obtained with speech models trained on developmentally realistic amounts of speech (< 1000 hours); beyond improving data efficiency, exploring a variety of inductive biases is primarily crucial in developing more human-like systems with relevance for linguistic theory.

Take-home message. We agree that neural network models developed for technological applications can provide insights into human language and cognition. The early connectionist work that F&M describe as preceding the LLM era included efforts to take the continuous speech signal seriously (Elman & Zipser, 1988; Norris, 1994), and we argue that current linguistic investigations with deep learning models can and should do the same. Once the bottleneck of text can be replaced, we stand a better chance of modelling more human-like linguistic processes, as well as handling the vast majority of local and regional language varieties that are rarely or never written down.

References

Alishahi, A., Barking, M., & Chrupała, G. (2017). Encoding of phonology in a recurrent neural model of grounded speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 368-378). https://doi.org/10.18653/v1/K17-1037

Arora, S., Chang, K. W., Chien, C. M., Peng, Y., Wu, H., Adi, Y., Dupoux, E., Lee, H. Y., Livescu, K., & Watanabe, S. (2025). On the landscape of spoken language models: A comprehensive survey. *arXiv preprint*. https://doi.org/10.48550/arXiv.2504.08528

Aslin, R., & Pisoni, D. (1980). Some developmental processes in speech perception. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), Child phonology: Vol 2, Perception (pp. 67–96). New York: Academic Press.

Boersma, P. (2009). Unidirectional optimization of comprehension can achieve bidirectional optimality. Talk presented at 10th Szklarska Poręba Workshop on the Roots of Pragmasemantics, Szklarska Poręba, March 13, 2009.

Boersma, P., Benders, T. & Seinhorst, K. (2020). Neural networks for phonology and phonetics. *Journal of Language Modelling* 8(1): 103–177. https://doi.org/10.15398/jlm.v8i1.224

Chrupała, G. (2023). Putting Natural in Natural Language Processing. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 7820-7827). https://doi.org/10.18653/v1/2023.findings-acl.495

de Heer Kloots, M., Zuidema, W. (2024) Human-like Linguistic Biases in Neural Speech Models: Phonetic Categorization and Phonotactic Constraints in Wav2Vec2.0. *Proc. Interspeech 2024*, 4593-4597. https://doi.org/10.21437/Interspeech.2024-2490

de Seyssel, M., Lavechin, M., Titeux, H., Thomas, A., Virlet, G., Revilla, A.S., Wisniewski, G., Ludusan, B., Dupoux, E. (2023) ProsAudit, a prosodic benchmark for self-supervised speech models. *Proc. Interspeech 2023*, 2963-2967. https://doi.org/10.21437/Interspeech.2023-438

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43-59. https://doi.org/10.1016/j.cognition.2017.11.008

Elman, J. L., & Zipser, D. (1988). Learning the hidden structure of speech. *The Journal of the Acoustical Society of America*, *83*(4), 1615-1626. https://doi.org/10.1121/1.395916

Gauthier, J., Breiss, C., Leonard, M. K., & Chang, E. F. (2025). Emergent morpho-phonological representations in self-supervised speech models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 28055-28074). https://doi.org/10.18653/v1/2025.emnlp-main.1425

Grice, H. P. (1975). Logic and conversation. In P. Cole & J.J. Morgan (eds.), *Syntax and semantics 3: Speech acts*, pp. 41–58. Academic Press, New York NY.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* 79:2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43: 59–69. https://doi.org/10.1007/BF00337288

Lavechin, M., Sy, Y., Titeux, H., Blandón, M.A.C., Räsänen, O., Bredin, H., Dupoux, E., Cristia, A. (2023) BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models. *Proc. Interspeech 2023*, 4588-4592. http://doi.org/10.21437/Interspeech.2023-978

Martin, K., Gauthier, J., Breiss, C., Levy, R. (2023) Probing Self-supervised Speech Models for Phonetic and Phonemic Information: A Case Study in Aspiration. *Proc. Interspeech 2023*, 251-255. http://doi.org/10.21437/Interspeech.2023-2359

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive psychology*, *18*(1), 1-86. https://doi.org/10.1016/0010-0285(86)90015-0

Millet, J., & Dunbar, E. (2022). Do self-supervised speech models develop human-like perception biases?. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7591-7605). https://doi.org/10.18653/v1/2022.acl-long.523

McMurray, B., Horst, J. S., Toscano, J. C., & Samuelson, L. K. (2009). Integrating Connectionist Learning and Dynamical Systems Processing: Case Studies in Speech and Lexical

Development. In John P. Spencer, Michael S.C. Thomas, and James L. McClelland (eds.), *Toward a unified theory of development: connectionism and dynamic systems theory re-considered*, pp. 218–249. Oxford University Press, New York NY. https://doi.org/10.1093/acprof:oso/9780195300598.003.0011

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189-234. https://doi.org/10.1016/0010-0277(94)90043-4

Orhan, P., Boubenec, Y., & King, J. R. (2025). The detection of algebraic auditory structures emerges with self-supervised learning. *PLOS Computational Biology*, *21*(9), e1013271. https://doi.org/10.1371/journal.pcbi.1013271

Pasad, A., Chien, C. M., Settle, S., & Livescu, K. (2024). What do self-supervised speech models know about words?. *Transactions of the Association for Computational Linguistics*, *12*, 372-391. https://doi.org/10.1162/tacl_a_00656

Poli, M., Schatz, T., Dupoux, E., & Lavechin, M. (2024). Modeling the initial state of early phonetic learning in infants. *Language Development Research*, *5*(1). https://doi.org/10.34842/y89t-6q31

Salakhutdinov, R. R., and Hinton, G. E. (2009). Deep Boltzmann machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*. Clearwater, Florida. https://proceedings.mlr.press/v5/salakhutdinov09a.html

Shen, G., Watkins, M., Alishahi, A., Bisazza, A., & Chrupała, G. (2024). Encoding of lexical tone in self-supervised models of spoken language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 4250-4261). https://doi.org/10.18653/v1/2024.naacl-long.239

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, *46*(1-2), 159-216. https://doi.org/10.1016/0004-3702(90)90007-M

Smolensky, P. (1999). Grammar-based connectionist approaches to language. *Cognitive Science*, *23*(4), 589-613. https://doi.org/10.1016/S0364-0213(99)00017-8

Soderstrom, M., Mathis, D., & Smolensky, P. (2006). Abstract genomic encoding of universal grammar in Optimality Theory. In P. Smolensky & G. Legendre (eds.), *The Harmonic Mind, From Neural Computation to Optimality-Theoretic Grammar (Volume II: Linguistic and Philosophical Implications)*, pp. 403-471. MIT Press.