

MIT Press • Open Encyclopedia of Cognitive Science

Speech Recognition

James M. McQueen¹

¹**Donders Institute for Brain, Cognition and Behaviour, Radboud University**

MIT Press

Published on: May 27, 2025

DOI: <https://doi.org/10.21428/e2759450.626f4797>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

For listeners to understand speech, they must recognize the constituent words of spoken utterances. The cognitive science of speech recognition has therefore focused on developing theories of how words are recognized and on constructing computational models instantiating those theories. Researchers have sought to discover solutions to core computational problems, including how listeners deal with the enormous physical variability of speech, how they cope with the fact that words usually sound like many other words, and how they segment continuous speech into discrete words. Researchers have also asked about the nature of lexical (word) representations, about their constituent parts (representations of vowels and consonants and of prosodic structures such as syllables and lexical stress patterns), and about the cognitive processes all these representations engage in. Controversies and current issues include the following: (1) Is there lexical–prelexical feedback? (2) Are there individual differences in speech recognition? (3) How are words recognized in multimodal settings? (4) What is the neurobiology of speech recognition?

History

Research on speech recognition contributed to the formation of the field of cognitive science. For example, core theoretical questions (e.g., the role of contextual information, the nature of perceptual representations) were already being asked in the 1950s ([Ladefoged & Broadbent, 1957](#); [Liberman et al., 1957](#); [Miller et al., 1951](#)). Work since then can broadly be characterized as reflecting increasing awareness about the richness of acoustic phonetic information (e.g., spoken words are not strings of phonemes like written words are strings of letters) and about the need for cross-linguistic comparisons testing the language universality of theoretical claims.

The field's history can also be traced in the development of computational models [see [Computational Models of Language Learning](#)]. Later models included new insights (e.g., the cohort model was more attuned to the temporally continuous nature of spoken-word recognition than the older Logogen model) and greater plausibility (e.g., Shortlist had a more realistically sized lexicon—20,000 words—than the older TRACE model—200 words). For references to and summaries of these and other models, see [Weber and Scharenborg \(2012\)](#).

Core concepts

Variability problem

Words sound different when spoken by talkers varying, for example, in age, gender, and (regional or foreign) accent. Even within one talker, the realization of the same word will vary as a function of, amongst other things, speaking rate, accentuation, and speaking style. The solution to this variability problem appears to be partially representational (e.g., listeners store variant forms of phonetically reduced words; [Ernestus et al., 2002](#)), partially procedural (e.g., mechanisms adapt to the acoustic consequences of differences in the shape of talkers' vocal tracts; [Ladefoged & Broadbent, 1957](#)), and partially a mixture (e.g., listeners store knowledge about talkers' ways of speaking and use it to adapt to those talkers; [Zhang & Holt, 2018](#)).

Lexical embedding problem

Because lexica with tens of thousands of words are built from phoneme inventories with rarely more than 40 entries, words necessarily sound like other words. Word recognition thus entails the parallel evaluation of multiple lexical hypotheses (similar-sounding words that [partially] match the input; e.g., hypotheses “blue,” “glue,” “bloom,” and “moon” given the input “blue moon”). Most researchers agree there is a form of competition among those hypotheses ([McClelland & Elman, 1986](#)).

Segmentation problem

Unlike the words in this text, the boundaries between words are not reliably marked; continuous speech must thus be segmented into discrete words. Listeners solve this segmentation problem in multiple ways ([Mattys et al., 2005](#)): they use lexical competition (e.g., if “blue” wins in competition with “bloom,” it follows that there is a word boundary before the [m]) and a variety of cues to likely word boundary locations (e.g., metrical cues; [Cutler & Norris, 1988](#)).

Cognitive architecture: Representations

Researchers have asked about the nature of lexical representations. One extreme view was that they are phonologically abstract (e.g., phoneme strings with no acoustic detail); another was that they are episodic memory traces (e.g., memories of exactly what was said in full acoustic detail). Neither extreme appears tenable; models are probably abstractionist–episodic hybrids that store some but not all acoustic details ([McQueen et al., 2006](#)). A separate question concerns how morphologically complex words are represented ([Marslen-Wilson et al., 1994](#)) [see [Morphology](#)]. Relatedly, researchers have asked about words’ phonological constituents: Are vowels and consonants represented as, for example, phonemes, allophones, or (bundles of) phonetic features? This question remains unanswered ([Samuel, 2020](#)), although it is clear that, in addition to these segmental constituents, words also have suprasegmental (prosodic) constituents (e.g., syllables, and varying cross-linguistically, stress, and tonal patterns). Segmental processing is informed by suprasegmental processing, and vice versa ([Mitterer et al., 2016](#)).

Cognitive architecture: Processes

Researchers have also asked which computations are performed, in what order, during word recognition and how information flows between representations. Computations include different forms of abstraction, normalization, compensation, integration, adaptation, and prediction. These processes often interrelate; for example, abstract perceptual categories adapt to changing input, supporting cross-talker generalization ([Kleinschmidt & Jaeger, 2015](#)). It is agreed that word recognition is modulated by the information in the sentential (syntactic, prosodic, semantic, pragmatic) context but not which computations (e.g., prediction vs. integration) are responsible. There is broad consensus that there is cascaded (continuous) bottom-up flow of information from the prelexical to the lexical level, such that, for example, lexical processing is not delayed until prelexical (e.g., phonemic) decisions have been made ([Andruski et al., 1994](#)).

Questions, controversies, and new developments

(1) It remains controversial whether there is top-down feedback, whereby lexical processing (which words have been accessed) influences prelexical processing (i.e., the processing that led to lexical access). Evidence suggests there is off-line lexical–prelexical feedback for perceptual learning (reviewed in [Kleinschmidt & Jaeger, 2015](#)), but there is continuing debate on whether there is feedback during on-line processing ([Luthra et al., 2024](#)).

(2) Are there individual differences in speech recognition [see [Linguistic Variation](#)]? There has been increasing interest in whether listeners differ in how they recognize speech and, if so, whether those differences are stable listener traits ([Giovannone & Theodore, 2023](#)). Much remains to be discovered here because most earlier research on speech recognition focused on highly educated young (predominantly female) adults, simply because they were the easiest to recruit [see [WEIRD](#)].

(3) Much current research focuses on multimodal language. How do listeners recognize words in natural conversational settings where there is not only auditory speech information but also visual information available (e.g., lip movements and hand gestures; [Holler & Levinson, 2019](#)) [see [Gesture](#)]?

(4) Neuroscientific methods have been used to address the issues summarized above and to localize speech recognition to particular brain regions and networks ([Hickok & Poeppel, 2007](#)) [see [Neuroscience of Language](#)]. An important open question is how the biophysical properties of neurons support the computations necessary to recognize words.

Broader connections

Speech recognition has similarities and dissimilarities with reading and speech production [see [Language Production](#)]. A core question is whether the same representations and processes support language use across modalities (speech vs. print, recognition vs. production). Another connection is that to language acquisition: How did the adult recognition system develop? A final connection is to automatic speech recognition: Which aspects of current recognition technology have psychological or biological plausibility?

Further reading

- Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148–203. <https://doi.org/10.1037/a0038695>
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86. [https://doi.org/10.1016/0010-0285\(86\)90015-0](https://doi.org/10.1016/0010-0285(86)90015-0)
- Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357–395. <https://doi.org/10.1037/0033-295X.115.2.357>

- Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *WIREs Cognitive Science*, 3(3), 387–401. <https://doi.org/10.1002/wcs.1178>

References

- Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, 52, 163–187. [https://doi.org/10.1016/0010-0277\(94\)90042-6](https://doi.org/10.1016/0010-0277(94)90042-6)

↵

- Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113–121. <https://doi.org/10.1037/0096-1523.14.1.113>

↵

- Ernestus, M., Baayen, H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, 81, 162–173. <https://doi.org/10.1006/brln.2001.2514>

↵

- Giovannone, N., & Theodore, R. M. (2023). Do individual differences in lexical reliance reflect states or traits? *Cognition*, 232, 105320. <https://doi.org/10.1016/j.cognition.2022.105320>

↵

- Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8, 393–402. <https://doi.org/10.1038/nrn2113>

↵

- Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8), 639–652. <https://doi.org/10.1016/j.tics.2019.05.006>

↵

- Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148–203. <https://doi.org/10.1037/a0038695>

↵

- Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29(1), 98–104. <https://doi.org/10.1121/1.1908694>

↵

- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358–368.

↑

- Luthra, S., Crinnion, A. M., Saltzman, D. S., & Magnuson, J. S. (2024). Do they know it's Christmash? Lexical knowledge directly impacts speech perception. *Cognitive Science*, 48, e13449. <https://doi.org/10.1111/cogs.13449>

↑

- Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, 101(1), 3–33. <https://doi.org/10.1037/0033-295X.101.1.3>

↑

- Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, 134, 477–500. <https://doi.org/10.1037/0096-3445.134.4.477>

↑

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

↑

- McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, 30, 1113–1126. https://doi.org/10.1207/s15516709cog0000_79

↑

- Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 41, 329–335. <https://doi.org/10.1037/h0062491>

↑

- Mitterer, H., Cho, T., & Kim, S. (2016). How does prosody influence speech categorization? *Journal of Phonetics*, 54, 68–79. <https://doi.org/10.1016/j.wocn.2015.09.002>

↑

- Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, 111, 104070. <https://doi.org/10.1016/j.jml.2019.104070>

↑

- Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *WIREs Cognitive Science*, 3(3), 387–401. <https://doi.org/10.1002/wcs.1178>

↑

- Zhang, X., & Holt, L. L. (2018). Simultaneous tracking of coevolving distributional regularities in speech. *Journal of Experimental Psychology: Human Perception and Performance*, 44, 1760–1779.

