A fully Bayesian feature-based phone recognizer

Orhun Uluşahin[1], James M. McQueen[1,2]
[1] Donders Centre for Cognition, [2] Max Planck Institute for Psycholinguistics

Current models of human spoken word recognition (e.g., the Distributed Cohort Model, Gaskell & Marslen-Wilson, 1997, TRACE, McClelland & Elman, 1986, Earshot, Magnuson et al., 2020, and Shortlist B, Norris & McQueen, 2008) rely on a variety of computations to simulate the cognitive mechanisms that underlie human speech processing. However, they share a number of deficiencies across architectures, programming languages, and theoretical frameworks. One such deficiency is ambivalence to speech variation, especially talker-bound variation, which is extensive in real speech. These models all treat speech as if it was spoken by the same talker. A second deficiency is that these models lack the ability to adapt to different forms of speech, including adaptation to talker-specific idiosyncrasies (Eisner & McQueen, 2005).

We present a new model of spoken word recognition, the Adaptive Bayesian Continuous-speech (ABC) Model that does not have these two deficiencies. The model builds on the Bayesian framework of Shortlist B's word recognition process but extends its probabilistic mode of operation to a feature-based Bayesian phone recognizer as well. Thus, instead of relying on hard-coded confusion data from human participants (i.e., as in Shortlist B), ABC dynamically evaluates pseudo-spectral speech input (i.e., numerical representations of acoustic features) of each incoming segment on a feature-by-feature basis to compute relative probabilities for each phone in the listener's inventory. In ABC, phones are a combination of position-specific allophones (where necessary to model positional variability) and position-ambivalent phonemes, in line with allophonic processing at the pre-lexical level (Mitterer et al., 2013, 2018). Knowledge about different talkers' acoustic-phonetic idiosyncrasies is represented in the form of Gaussian or Beta distributions that constitute phone-level priors on a feature-by-feature basis. Thus, ABC enables the simulation of talker-specific perception, accommodating the "talker-contingent" nature of speech recognition (Nygaard et al., 1994). In doing so, our model expands the limited scope of feature-based Bayesian phone perception (e.g., Kleinschmidt & Jaeger, 2015; Xie et al., 2023) in two substantial ways: (1) ABC uses a language-sized inventory of allophones instead of an individual phoneme pair, and (2) compares an expandable array of 12 acoustic features instead of relying on one acoustic-phonetic contrast.

The combination of the Bayesian framework and a modular architecture keep computational demand to a minimum while maximizing user customizability. For instance, users can simulate different "listening profiles" based on between- or within-talker differences. Perceptual expectations can be configured simply by altering the feature-by-feature distributions that constitute phone priors. The lexicon can be expanded, shrunk, or completely replaced as long as it follows ABC's barebones format. Acoustic features can be added or removed so long as they can be represented by a Gaussian or Beta distribution.

Therefore, although the primary use case for the ABC phone recognizer is its integration within the larger model of word recognition, the phone recognizer module alone has a number of other use cases. These include (1) Exploring probabilistic/graded phone perception as opposed to the strict categorical perception of many models, (2) Modelling cue weighting through changes to feature mismatch penalties and feature-by-feature weights, and (3) Simulation of degraded, clinical or otherwise atypical speech or listening conditions.

At Dag van de Fonetiek 2026, we will present a detailed overview of the probabilistic phone recognizer module of ABC, and present simulation results that capture (talker-bound) perceptual learning (Eisner & McQueen, 2005) and its generalization across the lexicon (McQueen et al., 2006) with a phonetic inventory of 49 Dutch phones in a lexicon of 35,000 words. These simulations will demonstrate ABC's adaptability, expandability, and its capacity to handle acoustic variation in speech, with a particular emphasis on between-talker variability, addressing multiple major deficiencies in existing models of spoken word recognition.

References

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. https://doi.org/10.3758/BF03206487

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating Form and Meaning: A Distributed Model of Speech Perception. *Language and Cognitive Processes*, *12*(5–6), 613–656. https://doi.org/10.1080/016909697386646

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

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., & Rueckl, J. G. (2020). EARSHOT: A Minimal Neural Network Model of Incremental Human Speech Recognition. *Cognitive Science*, *44*(4), e12823. https://doi.org/10.1111/cogs.12823

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

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological Abstraction in the Mental Lexicon. *Cognitive Science*, *30*(6), 1113–1126. https://doi.org/10.1207/s15516709cog0000_79

Mitterer, H., Reinisch, E., & McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language*, *98*, 77–92. https://doi.org/10.1016/j.jml.2017.09.005

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361. https://doi.org/10.1016/j.cognition.2013.07.011

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

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech Perception as a Talker-Contingent Process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex*, *166*, 377–424. https://doi.org/10.1016/j.cortex.2023.05.003

Funding

This research is supported by the Dutch Research Council (NWO), Social Sciences and Humanities Open Competition grant 406.23.CTW.005.