Towards a new model of spoken-word recognition

Orhun Uluşahin[1], James M. McQueen[1,2]
[1] Donders Centre for Cognition, Radboud University, Nijmegen, NL
[2] Max Planck Institute for Psycholinguistics, Nijmegen, NL
orhun.ulusahin@donders.ru.nl

Existing models of spoken-word recognition such as TRACE [1], the Distributed Cohort Model [2], and Shortlist B [3] each offer computationally explicit accounts of how speech is recognized and each can simulate an impressive range of phenomena. But these models fail to account for talker variability: They are effectively models of how words might be recognized if they were always spoken by the same talker. This deficiency is striking, given substantial evidence that word recognition is talker-contingent [4] and that listeners can tune in to the idiosyncrasies of individual talkers' speech [5]. These models have a second striking deficiency. Listeners use prosodic information (e.g., lexical stress cues) in word recognition [6] and yet the models are effectively prosodically deaf (i.e., they treat speech as if it were made up of strings of vowels and consonants with no suprasegmental structure).

We present a new model of spoken-word recognition: the Adaptive Bayesian Continuous-speech (ABC) model. The ABC model addresses the above two deficiencies: It is the first of its kind to handle talker variability and to not be prosodically deaf. The ABC model takes inspiration from three other Bayesian ideal observer models: Shortlist B [3], the Bayesian Prosody Recognizer [6], and the ideal adapter framework [7]. Here, we present ABC's architecture and show how the model handles talker variability in the realization of fricative consonants.

A critical assumption (built into the model's name) is that speech recognition is adaptive: talker variability is dealt with through retuning of perception. ABC therefore has adaptive voice "plug-ins" for multiple talkers. The core idea is that ABC has knowledge of the distributional properties of the acoustic cues to different phonological categories (i.e., pdfs), knowledge about how these likelihood functions vary across talkers, and the ability, through the deployment of plug-ins with this knowledge, to retune the recognition process as the input changes from one talker to another. The first step towards a full ABC model is the implementation of plug-ins for talker-specific cues to fricative consonants using existing data on lexically-guided perceptual learning in Dutch [5,8]. Different likelihood functions for different talkers will be constructed for acoustic cues to the fricative contrast (e.g., spectral centre of gravity for [f] and [s]), modelling the input in the exposure phase of [5]). The model will then be evaluated, as in the test phase of [8], on its ability to recognize words containing the fricatives and, critically, whether this is a talker-contingent process. The model will do talker recognition through bottom-up template matching. Then the talker's plug-in will be deployed: fricative likelihoods will be adjusted in a talker-specific way, leading to changes in the recognition of words containing those fricatives.

This work is ongoing and simulation results are not yet available. We hope to be able to show that the ABC model can simulate generalization of talker-specific learning about fricatives across a large (20,000 word) lexicon. In future work, we will (a) add prosodic structure to the model, (b) address prosodic talker variability, and (c) capture how listeners deal with groups of talkers who share the same regional or foreign accent. The fricative retuning simulations, however, will already offer an existence proof that talker-specific adaptability can be built into a large-lexicon Bayesian model.

References

[1] 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

[2] Gaskell, M.G. & Marslen-Wilson, W.D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, 12, 613–56. https://doi.org/10.1080/016909697386646

[3] Norris, D. & McQueen, J.M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115, 357–95. https://doi.org/10.1037/0033-295X.115.2.357

[4] Nygaard, L.C., Sommers, M. S. & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5, 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

[5] Eisner, F. & McQueen, J.M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, 67, 224–38. https://doi.org/10.3758/BF03206487

[6] McQueen, J.M. & Dilley, L.C. (2020). Prosody and spoken-word recognition. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp. 509–21). Oxford: OUP. https://doi.org/10.1093/oxfordhb/9780198832232.013.33

[7] 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, 148–203. https://doi.org/10.1037/a0038695

[8] McQueen, J.M., Cutler, A. & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, 30, 1113–26. https://doi.org/10.1207/s15516709cog0000_79