Probing articulatory representation learning for phonological distinctions

Sean Foley
University of Southern California, Los Angeles, California, United States

Abstract

While a growing body of work has aimed to extract spatio-temporal units directly from speech articulatory data, there have been few attempts to probe whether such representations capture phonological contrasts employed in language and to model the mapping between motor plans and phonological representations. This study employs a joint factor analysis and neural convolutive matrix factorization framework to a multi-speaker real-time MRI dataset of vocal tract contours. The framework generates both gestures, the spatio-temporal units that form a given utterance, and gestural scores, which detail the activation of individual gestures in time. Probing of the gestural scores shows some ability to capture phonological distinctions, suggesting that such information is encoded by the model. The gestures, however, show poor discriminability along crucial phonological dimensions, likely limited by cross-speaker spatial variability. The results highlight the difficulties in cross-speaker articulatory modeling, but also show promise in using deep learning to model articulatory representations.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*