Multimodal Input Aids a Bayesian Model of Phonetic Learning

Sophia Zhi
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Roger Levy
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Stephan Meylan
MIT, Cambridge, Massachusetts, United States

Abstract

One of the many tasks facing the typically-developing child language learner is learning to discriminate between distinctive sounds that make up words in their native language. We investigate whether multimodal information—specifically adult speech coupled with video frames of speakers' faces—benefits a computational model of phonetic learning. We introduce a method for creating high-quality synthetic videos of speakers' faces for an existing audio corpus. Our learning model, when trained and tested on audiovisual inputs, achieves 8.1% relative improvement on a phoneme discrimination battery compared to a model trained and tested on audio-only input. It outperforms the audio model by 3.9% when tested on audio-only data, suggesting that visual information facilitates the acquisition of acoustic distinctions. In noisy audio environments, our audiovisual model recovers 67% of the loss in performance of the audio model relative to non-noisy environments. These results demonstrate that visual information benefits an ideal learner and illustrate multiple ways that children might leverage visual cues when learning to discriminate speech sounds.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*