Prosody in the Age of AI: Insights from Large Speech Models

Samuel Sohn
Rutgers University, New Brunswick, New Jersey, United States

Sten Knutsen
Rutgers University, New Brunswick, New Jersey, United States

Karin Stromswold
Rutgers University, New Brunswick, New Jersey, United States

Abstract

Prosody affects how people produce and understand language, yet studies of how it does so have been hindered by the lack of efficient tools for analyzing prosodic stress. We fine-tune OpenAI Whisper large-v2, a state-of-the-art speech recognition model, to recognize phrasal, lexical, and contrastive stress using a small, carefully annotated dataset. Our results show that Whisper can learn distinct, gender-specific stress patterns to achieve near-human and super-human accuracy in stress classification and transfer its learning from one type of stress to another, surpassing traditional machine learning models. Furthermore, we explore how acoustic context influences its performance and propose a novel black-box evaluation method for characterizing the decision boundaries used by Whisper for prosodic stress interpretation. These findings open new avenues for large-scale, automated prosody research with implications for linguistic theory and speech processing.

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