A diagnostic approach to evaluating TTS prosody in conveying information structure

Na Hu[a,e*], Jiseung Kim[b,e], Michele Gubian[a], Riccardo Orrico[c,e], Stella Gryllia[d,e], James Kirby[a], Amalia Arvaniti[e]

[a] IPS, LMU München, Munich, Germany

[b] Chosun University, Gwangju, South Korea

[c] École Normale Supérieure, Paris, France

[d] Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

[e] Radboud University, Nijmegen, the Netherlands

Prosody plays a crucial role in conveying linguistic and communicative information in human speech. Although state-of-the-art TTS systems generate highly natural-sounding speech, the extent to which they resemble human speech in conveying meaning through prosody remains unclear, due to a lack of appropriate evaluation approaches.

This paper addresses the gap by proposing an evaluation approach that identifies *how* TTS prosody deviates from human prosody in conveying meaning. Combining Functional Principal Component Analysis with Bayesian multivariate linear regression, this approach quantifies differences in continuous prosodic properties along interpretable dimensions of variation. We demonstrate its diagnostic value by examining the intonational realization of information structure using OpenAI's TTS system as a case study, focusing on accent placement and the F0 contour shapes of accented words, with accentuation and information status manually annotated. Results reveal that, despite high agreement in accentuation between human and TTS speech, differences remain in the F0 contour shapes of accented words. While human speech distinguishes information status categories through F0 scaling, slope, and curvature, TTS speech relies only on F0 scaling. The proposed approach thus provides direct, interpretable evidence of how TTS prosody differs from human prosody and highlights the importance of using context-rich materials in TTS evaluation.

The proposed evaluation approach can be extended to other linguistic phenomena and continuous properties of prosody, given relevant human speech data and annotations. It complements existing global naturalness-focused evaluation metrics, providing a more comprehensive understanding of how TTS systems convey linguistic meaning through prosody.