Word stress in self-supervised speech models: A cross-linguistic comparison

Martijn Bentum[1], Louis ten Bosch[1], Tomas O. Lentz[2]
[1]Centre for Language Studies, Radboud University
[2]Department of Communication and Cognition, Tilburg University

Self-supervised speech models (S3Ms) learn general-purpose representations of spoken language that can be fine-tuned for tasks such as speech recognition, speaker identification, and emotion detection. Yet the end-to-end nature of S3Ms makes them difficult to interpret. A common approach is diagnostic classification, where simple classifiers probe model layers for linguistic information. Prior work has shown that S3Ms encode phonetic, semantic, syntactic, and prosodic cues (e.g., Pasad, Chou & Livescu, 2021; Bentum, ten Bosch & Lentz, 2024). This study extends that approach to investigate how word stress is represented in S3Ms across languages. Specifically, we investigate the S3M representations of word stress for five different languages: Three languages with variable or lexical stress (Dutch, English and German) and two languages with fixed or demarcative stress (Hungarian and Polish).

Word stress refers to the relative prominence of syllables within words (Gussenhoven, 2004), realized acoustically through correlates such as duration, intensity, pitch, spectral tilt, and formant peripherality (e.g., van Heuven, 2018). These cues vary in reliability across languages: in fixed-stress languages (e.g., Hungarian, Polish) stress occurs in predictable positions, while in variable-stress languages (e.g., Dutch, English, German) stress is lexically distinctive and less predictable. Human listeners show corresponding sensitivity, with “stress deafness” observed in fixed-stress language speakers (PeperKamp, Vendelin & Dupoux, 2010).

We used the multilingual Wav2vec 2.0 XLS-R model (Babu et al., 2021), trained on 128 languages, and examined bisyllabic words in read-aloud sentences from Common Voice. Stress labels were assigned using CELEX for variable-stress languages and rule-based methods for fixed-stress languages. Classifiers were trained on both acoustic features and model embeddings extracted from different model layers.

Results show that stress can be reliably decoded from S3M embeddings in all five languages, with peak performance around transformer layer 17. Unlike acoustic correlates, model representations consistently revealed language-specific clustering, separating fixed- from variable-stress languages. These findings suggest that S3Ms encode abstract, language-specific stress representations beyond acoustic correlates, offering new insights into how prosody is captured in multilingual models

References

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. *XLS-R: Self-supervised cross-lingual speech representation learning at scale*. arXiv preprint arXiv:2111.09296, 2021.

Bentum, M., ten Bosch, L., & Lentz, T. (2024). The processing of stress in end-to-end automatic speech recognition models. In *Proceedings of Interspeech 2024* (pp. 2350-2354).

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge University Press.

Pasad, A. Chou, J. C.& Livescu, K. (2021). Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. (pp. 914–921).

Peperkamp, S., Vendelin, I., & Dupoux, E. (2010). Perception of predictable stress: A cross-linguistic investigation. *Journal of Phonetics*, 38(3), 422-430.

van Heuven, V. J. (2018). Acoustic correlates and perceptual cues of word and sentence stress: towards a cross-linguistic perspective. In R. Goedemans, J. Heinz and HulstH. van der (Eds.), *The Study of word Stress and accent: Theories, Methods and data.* (15–59). England: Cambridge University Press.