Analyzing representations of self-supervised speech models

Sharon Goldwater, University of Edinburgh

Recent advances in speech technology make heavy use of pre-trained models that learn from large quantities of raw (untranscribed) speech, using "self-supervised" (ie unsupervised) learning. These models learn to transform the acoustic input into a different representational format that makes supervised learning (for tasks such as transcription or even translation) much easier. However, *what* and *how* speech-relevant information is encoded in these representations is not well understood. I will talk about work in which my group is analyzing the structure of these representations, to gain a more systematic understanding of how word-level, phonetic, and speaker information is encoded.