Textless NLP: towards language processing from raw audio.

Emmanuel Dupoux
EHESS

The oral (or gestural) modality is the most natural channel for human language interactions. Yet, language technology (Natural Language Processing, NLP) is primarily based on the written modality, and requires massive amounts of textual resources for the training of useful language models. As a result, even fundamentally speech-first applications like speech-to-speech translation or spoken assistants like Alexa, or Siri, are constructed in a Frankenstein way, with text as an intermediate representation between the signal and language models. Besides this being inefficient, this has two unfortunate consequences: first, only a small fraction of the world's languages that have massive textual repositories can be addressed by current technology. Second, even for text-rich languages, the oral form mismatches the written form at a variety of levels, including vocabulary and expressions. The oral medium also contains typically unwritten linguistic features like rhythm and intonation (prosody) and rich paralinguistic information (non-verbal vocalizations like laughter, cries, clicks, etc., nuances carried through changes in voice qualities) which are therefore inaccessible to language models. But is this a necessity? Could we build language applications directly from the audio stream without using any text? In this talk, we review recent breakthroughs in representation learning and self-supervised techniques which have made it possible to learn latent linguistic units directly from audio which unlock the learning of generative language models without the use of any text. We show that these models can capture heretofore unaddressed nuances of the oral language including in a dialogue context, opening up the possibility of speech-to-speech textless NLP applications. We outline existing technical challenges to achieve this goal, including challenges to build expressive oral language datasets at scale.