A Textless Approach for Generative Spoken Language Modeling

Abstract:

An open question for AI research is creating systems that learn from natural interactions as infants learn their first language(s): spontaneously and without access to text or expert labels. Current NLP systems require large amounts of text, which excludes plenty of the world’s languages that have little textual resources or no widely used written form. In addition, textual features do not encode speaker-specific speech properties beyond content (e.g., identity, style, emotion, etc.), as well as structured signals that are part of natural human interaction (intonation, hesitation, laughter, etc.) which are important in the oral form. In this talk, I’ll present our recent studies in developing a Textless Approach for Generative Spoken Language Modeling. The proposed framework is comprised of a pseudo-text Encoder, Sequential modeling, and Speech generation components, all of which were trained in an unsupervised fashion. Lastly, I will present various applications which can benefit from such modeling together with future research directions.