Speech Synthesis from Found Data

Pallavi Baljekar

AbstrText-to-speech synthesis (TTS) has progressed to such a stage that given a large, clean, phonetically balanced dataset from a single speaker, it can produce intelligible, almost natural sounding speech. However, one is severely limited in building such systems for low-resource languages where there is a lack of such data and there is no access to a native speaker.

Thus, the goal in this thesis is to use the data that is freely available on the web, a.k.a. “Found Data” to build TTS systems. However, since this data is collected from different sources, it is noisy and contains a lot of variations in terms of speaker, language and channel characteristics as well as prosody and speaking style. Conventional TTS systems on the other hand, require a large collection of clean, phonetically balanced, single-speaker data recorded specifically for the purposes of building TTS systems. This presents us with a number of challenges in using found data for building TTS systems within the current pipeline.

In this thesis, we address three of these challenges. First we look at data selection strategies to select good utterances from noisy found data which can produce intelligible speech.

Second, we investigate data augmentation techniques from cleaner external sources of data. Specifically, we study cross-lingual data augmentation techniques from high resource languages. However, often found audio data is un-transcribed. Thus, we also look at methods of using un-transcribed audio along with unrelated text data in the same language to build a decoder for transcription. Furthermore, we address the issue of language, speaker, and channel variations, by training multi-language, multi-speaker models, in a grapheme based neural attention framework.

Lastly, since most of the speech data available on the web is in the form of audiobooks and podcasts, we explore iterative methods of learning a good prosody labeling for long form audio as well as learning prosody embeddings which mimic metrical structure of utterances in an unsupervised fashion to better match the acoustics of the data. In addition, we investigate methods of learning a set of “prominence weights” in attention based neural models, with the goal of improving prosody as well as overall quality of synthesized speech.