Corpus generation for research in pause-internal phonetic particles

Oral Submission

Within our research on pause-internal phonetic particles (PINTS) [http://pauseparticles.org/], there is a sub-project to collect spoken speech data with appropriate annotations. A primary usage for this corpus is to develop a pipeline for modeling different PINTS in speech synthesis. A secondary goal is to develop a corpus that is consistent, re-usable, and can add value to projects outside of our immediate needs, with a focus on unscripted spontaneous speech. Initially, the corpus will be created in English with the possibility of implementation into additional languages, such as, German and French.

The corpus will contain audio recordings collected from YouTube, scraped from a variety of speakers and topics, in spontaneous and semi-spontaneous situations. This material can be automatically trained with our PINTS models or used in unsupervised training. Important categories for the annotation of PINTS includes: acoustically silent pauses, filler particles ('filled pauses'), clicks, and a variety of breathing phenomena. During this testing phase only breath noises were investigated.

The current pipeline is a multi-step process. First, both the audio and subtitles are downloaded from YouTube. The subtitles are generated automatically by YouTube's speech recognition algorithms, or from user-generated subtitles. Next, the individual words from the subtitles are converted into their corresponding parts of speech (POS). Both the words and the POS are converted into time-aligned annotations for further analysis. The final step is modeling the predictions for the pause particles (in this case breath noises). This initial testing phase used a neural network, trained on audio previously annotated with breath noises, to model the breath noise predictions.

In this test phase, the proposed pipeline effectively gathered the relevant annotations. Moving forward, this pipeline can be used to create a curated corpus for our project. Additionally, other researchers could implement this pipeline for conducting their own basic and applied research in the speech sciences.