Automatic Analysis of Audio Diary Speech Duration and Relative Speech Volume

This paper presents a Python application, DiSpeechEval, for using voice activity detection to summarize the duration and relative loudness of speech in audio files, flagging files with less proportional speech or unclear speech. Unlike other tools that can be used to calculate speech quantity in an audio corpus, DiSpeechEval accepts untranscribed audio. Originally developed for the MI Diaries project (Sneller et al., 2022), other research projects that possess large quantities of varying-quality audio data may benefit from it as well.

The MI Diaries Project is a longitudinal sociolinguistic research project started in 2020 that collects audio diaries from participants living in Michigan. Participants submit diaries remotely through a mobile app and have submitted more than 100,000 minutes of cumulative audio since the project began. One drawback to this method is that audio quality and speech quantity can vary substantially across entries and participants, as some participants speak quietly, move around, record in noisy settings, and sometimes even accidentally mute their microphones.

Manual identification of diaries with poor audio quality or long silences by MI Diaries team members is non-trivial, as these problems may only affect specific portions of relatively lengthy recordings. Therefore, I developed DiSpeechEval to automatically and rapidly evaluate speech quantity and quality in audio files using rVAD voice activity detection (Tan et al., 2020) and the librosa audio analysis package (McFee et al., 2015).

To determine what audio metrics could be used to reliably flag diaries with poor audio quality, I created an exploratory sample of 23 audio diaries. 11 of these diaries were noted by MI Diaries team members to be of low audio quality (containing long silences, quiet speech, or significant background noise). 12 were selected from diaries that team members have featured on the project website, reflecting acceptable audio quality. I also created a 30-diary test set by randomly sampling ten diaries with durations greater than five minutes from each participant age category (adult, teen, kid) of the project, which I coded impressionistically as being of clear or poor audio quality.

The most reliable metrics for distinguishing the featured audio diaries from the poor audio quality diaries were proportion of non-speech to audio duration and difference in median decibel level between speech and non-speech segments as a proportion of total recording decibel range, with thresholds of greater than 40% silence and less than 6.5% median decibel difference, respectively. A flagging function applying these thresholds to the test set correctly flagged all diaries with poor audio quality, while also incorrectly flagging nine clear recordings, scoring a recall of 1.0, a precision of 0.47, and an F1 score of 0.64. All errors resulted from the median decibel level metric, suggesting the metric can be improved. The flagging function took 3.5 minutes to summarize and evaluate 465 minutes of audio.

To allow these techniques to be applied in other contexts, I refined the experimental code into a user-friendly and easily modifiable Python application that is publicly available.