Automated Forced Temporal Alignment of MULTISIMO Transcripts with Speech Signals

Kavya Bhadre Gowda, Master of Science in Computer Science

University of Dublin, Trinity College, 2020

Supervisor: Dr. Carl Vogel

This research work explored the various tools and techniques available to perform the time-based alignment of lengthy MULTISIMO audio files, which contain lengthy MULTISIMO audio "Continuous Speech Signals" of approximately 10 minutes to its respective transcripts automatically and accurately. A user-friendly web application was designed and developed to solve this time-based alignment problem, that provides accurate results even with noisy audio inputs and audio containing long pauses. Temporal alignment of audio and transcripts helps linguists to analyze the language at Phonetic and Word level. Analyzing audio signals using the "Forced Aligner Application" that has a higher signal-to-noise ratio, is easy and accurate.

The focus of Forced Aligner Application is to identify the Pronunciations and Syllables in the Speech Signals, which are utilized in training Speech Recognition Systems. This web application is developed by analyzing the existing state-of-the-art tools and techniques in the Forced Alignment area. This application was built using the HTK toolkit and has achieved an accuracy of 77.11% compared to the manual alignment of start and end time of utterances. We have achieved this accuracy by cleaning the transcripts, re-sampling the audio signals, and changing the configurations of the HTK toolkit in the ProsodyLab tool. Our analysis also showed that there is a significant correlation between the end time of the utterances and the mismatch in the alignment.

With Spearman's correlation, it is also proved that the alignment of the start-time of utterances is easier than compared to end-time.