A Robust Voice Activity Detection for Real-Time Automatic Speech Recognition

Omid Ghahabi, Wei Zhou, Volker Fischer
EML European Media Laboratory GmbH, Berliner Straße 45, 69120 Heidelberg, Germany
omid.ghahabi@eml.org

Abstract: Voice Activity Detection (VAD), locating speech segments within an audio recording, is a main part of most speech technology applications. Non-speech segments, e.g., silence, noise, and music, usually do not carry any interesting information in speech recognition applications and they even degrade the performance of the recognition system in terms of both the accuracy and computational cost. Various VAD techniques have been developed, but not all of them are appropriate for a real-time application where the robustness, accuracy, and the processing time are the main keys. In this paper, we propose a fast and robust VAD for a real-time Automatic Speech Recognition (ASR) task. The main goal is to effectively filter out the non-speech segments before processing the speech segments of the audio signal by the decoder. The proposed technique is a hybrid supervised/unsupervised model based on zero-order Baum Welch statistics obtained from a Universal Background Model (UBM). We will show that not only the processing time for the whole speech recognition task is decreased to a great extent, but also the recognition accuracy is increased by mainly reducing the insertion of undesired words.