Voice Activity Detection

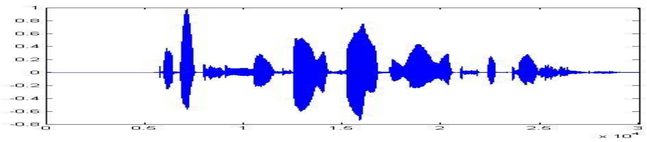
Jonathan Kola, Dr. Carol Espy-Wilson, Dr. Tarun Pruthi

MERIT FAIR
BIEN 2011

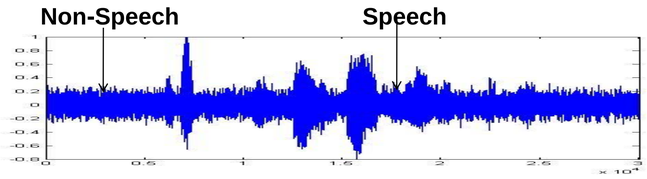
Introduction

• Voice activity detection is the process by which algorithms called Voice Activity Detectors (VADs) are able to distinguish regions that contain speech from regions that do not contain speech in an audio signal
• Several features distinguish speech from non-speech, however, where the speech signal is corrupted by background noise it becomes more and more difficult to characterize these features and make a decision

Speech (blue regions), non-speech (quiet regions)



Similarity between speech and non-speech in noise

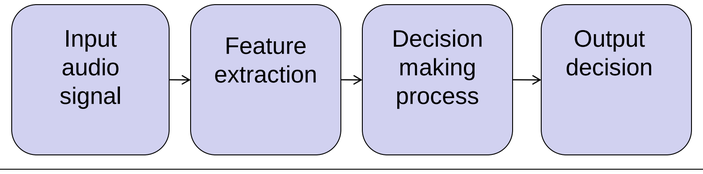


Goals

• Identify the best performing VAD algorithms in the literature
• Develop a comprehensive testing setup in which to compare these different VADs
• Determine the best performing algorithm and fully implement it in C programming language

Detection Process

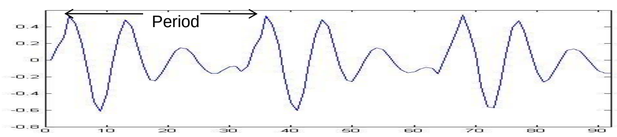
• Audio input signal processing
• Extraction of particular feature or set of features from processed audio signal
• Comparison of extracted features to heuristic or statistical model in order to determine likelihood of speech or non-speech events
• Decision making based on a likelihood threshold determined heuristically or derived from a statistical model



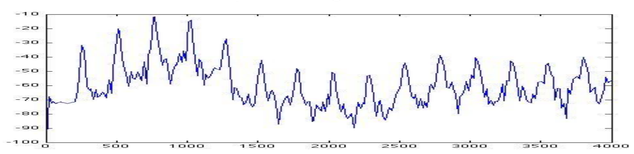
Feature Extraction

Commonly used features used to distinguish speech from non-speech are:
• Periodicity
• Fourier Coefficients
• Zero Crossing Rates

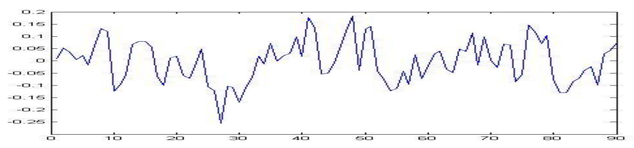
Speech signal showing periodicities in time domain



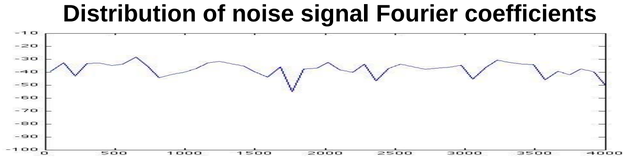
Distribution of speech signal Fourier coefficients



Aperiodic noise in time domain



Distribution of noise signal Fourier coefficients



Decision Making Process

• After feature extraction, these parameters are fitted into models in order to generate an output decision

A common statistical model setup

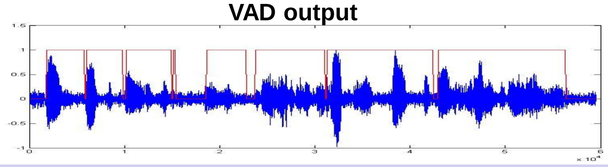
H_0 : Speech absent: X = N
H_1 : Speech present: X = N + S
$P(X \mid H_0) = N(\mu_N, \lambda_N)$
$P(X \mid H_1) = N(\mu_{N+S}, \lambda_{N+S})$

Where X is the parameter of the feature extracted, N and S are the parameter values in noise and speech respectively, μ_N and λ_N are the mean and variance of the parameter distribution in noise and μ_{N+S} and λ_{N+S} are the mean and variance of the parameter distribution in speech + noise.

Output Decision

• A VAD outputs a “1” for every signal frame it decides contains speech, and a “0” for every frame it decides does not contain speech

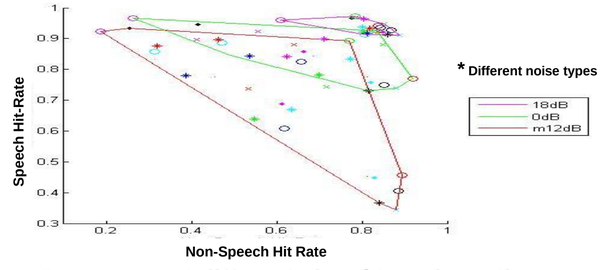
VAD output



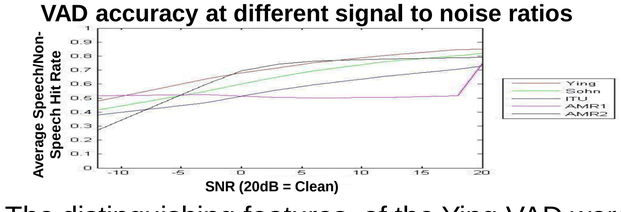
Experimental Results and Evaluation

The criteria used to evaluate VADs were based on:
• Accuracy in detecting speech (speech hit rate) and accuracy in detecting non-speech (non-speech hit rate) in different noise types at different signal-to-noise ratios (SNRs), on a scale from 0 to 1
• Consistency of performance across different SNRs and noise types
• From this analysis, an algorithm outlined in “VAD based on an Unsupervised Learning Framework”, proposed by D.Ying et al (IEEE Transactions on Audio, Speech and Language Processing (accepted)) proved to be the best performer

Ying VAD score in different noises at -12dB, 0dB and 18dB SNR



VAD accuracy at different signal to noise ratios



The distinguishing features of the Ying VAD were:
• Unsupervised noise and speech parameter learning techniques
• Adaptable decision threshold based on a Gaussian Mixture Model
• Scaling of the decision making procedure down to the level of individual frequency bands, which then contribute to determine aggregate results