IMPROVED VOICE ACTIVITY DETECTION IN THE PRESENCE OF PASSING VEHICLE NOISE

Stephen W. Laverty

Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609, USA
Email: steve@alum.wpi.edu

Donald R. Brown

Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609, USA
Email: drb@wpi.edu

Introduction

Voice activity detection (VAD) is an important enabling technology for a variety of speech-based applications including speech recognition, speech encoding, and hands-free telephony. The primary function of a voice activity detector is to provide an indication of speech presence in order to facilitate speech processing as well as possibly provide delimiters for the beginning and end of a speech segment. While VAD is often quite effective in benign acoustical environments, e.g. a conference room, it tends be less accurate in vehicular environments due to the strong noise present in the automobile cabin. Historically, vehicular voice activity detectors have relied on the fact that the noise in the automobile cabin tends to be stationary over long periods of time and, as such, can be suppressed to a large extent by an adaptive filter with coefficients obtained during non-speech periods [1]. While adaptive filtering does tend to improve the accuracy of VAD in the automotive environment, it is not capable of suppressing short-term nonstationary noise signals, e.g. noise from passing vehicles. In driving scenarios with frequent passing vehicle events, traditional vehicular voice activity detectors may suffer from an unacceptable number of false detections of speech and, as a consequence, the overall performance of the speech application may be significantly degraded.

This paper describes a new approach to improve the accuracy of VAD in automotive scenarios with frequent passing vehicle events. We focus on the multichannel far-field microphone case relevant to hands-free speech acquisition in automotive scenarios. In our system model, a total of four states are possible: $\{X, S, P, SP\} = \{$[no speech + no pass], [speech + no pass], [no speech + pass], [speech + pass]$\}$. Traditional VAD tends to be fairly accurate at distinguishing state X from states $\{S, P, SP\}$ but is less effective at discriminating between states S, P, and SP. Our focus in this contribution is on discrimination between states P and SP or, in other words, detecting the presence or absence of speech during passing vehicle events. Our proposed solution uses both power and pitch information from the noisy speech signal and leverages standard techniques from classification theory to optimally discriminate between the P and SP states. The proposed solution was tested on actual multichannel in-vehicle recordings and our results suggest that the proposed voice activity detector can significantly improve VAD accuracy in driving scenarios with frequent passing vehicle events.

Improved Voice Activity Detection

The objective of the passing-vehicle-noise tolerant voice activity detector (PVNT-VAD) is to determine, given observations from microphones in the cabin, whether a pass without speech (P) or a pass with speech (SP) is more likely to have generated those observations. Given this hypothesis testing structure we need to select a feature vector $\vec{x}$ that produces a conditional distribution $f_x(\vec{x}|P)$ that differs substantially from $f_x(\vec{x}|SP)$. The most common choice of feature is signal power. In the vehicular environment, however, the overall power tends to be dominated by the background noise and the noise of the passing vehicle. Although the passing vehicle noise is fairly broadband, both it and the background noise are heavily weighted toward lower frequencies. Accordingly, using a high pass filter before computing the power helps mitigate the influence of both sources of noise. Unvoiced speech, due to its broadband nature, immediately benefits from this filtering since its ratio of power in the passband to power in the stopband is much greater than that of the noise. Voiced speech, on the other hand, tends to be weighted toward lower frequencies. While high pass filtering applied to the voiced speech improves the separation of the distributions, the improvement is insufficient. Voiced speech can be accommodated by augmenting the feature vector with measurements from a pitch detector (which exploits the structure of voiced speech). The joint distribution of this two element feature vector can then be used to classify which state, P or SP, is more likely to be present.

The power and pitch measurements made on sample data can be analyzed by one of the many techniques made available by classification theory (e.g. linear or quadratic discriminant analysis [2] or kernel discriminant analysis [3]) to produce optimal decision regions. These decision regions can then be used as a classifier producing either P or SP as its decision and completing the final stage of the PVNT-VAD system as shown in Figure 1.

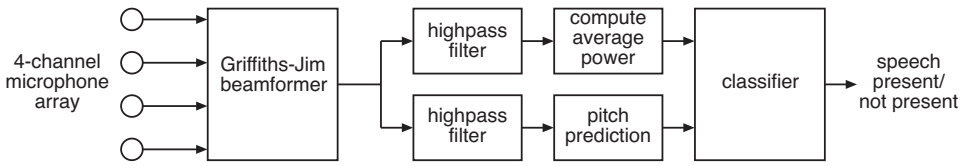
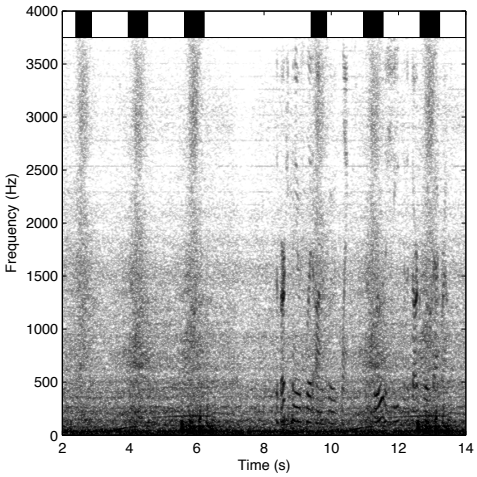


Figure 1: Block diagram of proposed passing-vehicle-noise tolerant voice activity detector (PVNT-VAD).

Results

The PVNT-VAD was tested on experimental data acquired from an array of four microphones attached to the driver's visor in a 1993 Dodge Intrepid at a sample size of sixteen bits and a sampling rate of 48kHz, later downsampled to 8kHz. Speech was recorded in a quiet, open area with the windows down and engine off. Passing vehicle noise was collected driving on two lane roads at speeds between 25 and 55 miles per hour also with the windows down. A test sample was generated by simply adding a portion of the speech recording to six passes from the noise recording. Each ten millisecond block of this signal was then classified manually. A spectrogram of the test signal is shown in Figure 2.

Quadratic discriminant analysis was chosen for classification. This data was then processed to produce a receiver operating characteristic (ROC) curve. For comparison, two other ROC curves were also generated. One curve demonstrates a simple power threshold after an optimally chosen bandpass filter. The other curve, suggested by recommendations in [1], results from using linear prediction in an attempt to eliminate pseudostationary noise before applying the power threshold. These three curves are shown in Figure 3 along with points generated by the G.729 Annex B VAD [4] before and after the Griffiths-Jim beamformer.



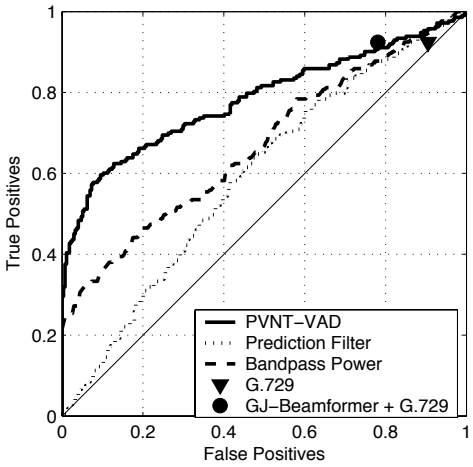


Figure 2: Spectrogram of sample used. Passes are highlighted. Only the highlighted areas are used for evaluation. The three passes on the left are without speech, the three passes on the right contain speech as well. The speech consists of a portion of a Harvard sentence list, specifically: "The birch canoe slid on the smooth plank. Glue the sheet to the dark blue background. Its easy to tell".

Figure 3: Performance of PVNT-VAD versus energy detector, prediction filter energy detector [1], and G.729 before and after the Griffiths-Jim beamformer for passing vehicle (P) versus passing vehicle and speech (SP).

Conclusions

This paper presented a new technique for the detection of speech in the presence of passing vehicle noise. The results presented in this paper suggest that the PVNT-VAD system provide a substantial gain in detection accuracy when compared to baseline methods.

References

[1] European Telecommunication Standard Institute (ETSI) EN 580-6 (GSM 06.32), "Voice Activity Detector (VAD) for full rate speech traffic channels,", 1996.

[2] M. James, *Classification Algorithms*, New York: John Wiley & Sons, 1985.

[3] D. J. Hand, *Kernel Discriminant Analysis*, Chichester: Research Studies Press, 1982.

[4] International Telecommunication Union Telecom Standardization Sector (ITU-T), "A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70 (G.729 Annex B),", 1996.