Voice Activity Detection

Jonathan Kola, Carol Espy-Wilson and Tarun Pruthi

Abstract - **Voice activity detectors (VADs) are ubiquitous in speech processing applications such as speech enhancement, signal-to-noise ratio (SNR) estimation, speech recognition, etc. VADs attempt to distinguish between speech and non-speech regions in a signal. Current VADs use measures such as energy differences, periodicity, and spectral differences that exist between different sounds. Some models use heuristic algorithms, while others use statistical models in a supervised or unsupervised learning framework. In this project, different VADs described in the literature were compared and evaluated on a comprehensive set of noises and SNRs. Based on their performance, the algorithm that performed most accurately and consistently was implemented in C programming language.**

***Index Terms*—Voice activity detection.**

I. Introduction

VOICE activity detection is an important step in speech processing applications such as speech enhancement, speech coding and speaker recognition. Voice activity detection approaches consist of feature extraction and discrimination models. Early voice activity detectors (VADs) paid attention to robust features of a signal such as energy, periodicity, dynamics and zero-crossing rates, and based their discrimination techniques on heuristic models. More recent VADs, while utilizing many of the same features, base their discrimination on statistical models. Typical statistical model based classifications use Gaussian distributions to describe various features of noise and speech, develop a likelihood ratio from comparison of measured parameters fitted in different models, and conduct a hypothesis test to make the speech/non-speech decision. Besides good and consistent performance across several different noise types and SNRs, the characteristics of a good VAD include low computational complexity and fast adaptation to changing noise types and SNRs.

The goal of this project was to determine conclusively the best and most consistent VAD algorithm out of the ones proposed in existing literature and standards, and to implement the best one in C programming language. Because proposed VAD algorithms in the literature are not subjected to standardized tests where they are compared against the same set of speech utterances, noise types and SNRs, it is difficult to know which VAD algorithms are the most robust, despite the conclusion of the authors. In this project therefore, an initial literature survey was carried out to determine the VAD algorithms that appeared to have exceptionally good performance. These VADs were determined by taking into account the conclusions of the authors, the novelty of the approach used, and the level of complexity of the algorithm. The complexity of the algorithm was especially crucial if the algorithm was to be practically and successfully implemented in C. After compiling the results from the literature survey based on the aforementioned criteria, a comprehensive test setup was designed, where each of the VADs were ran against the same database of speech utterances in different noise types at different SNRs. Results across different measurements and types of classification of errors were used to evaluate the performance of the different VADs. Finally, based on the results of the test, the best performing VAD was chosen and the algorithm was implemented in C programming language.

The rest of the paper is organized as follows. Section II provides an overview of a generalized VAD algorithm and a summary of the literature survey. Section III describes the test setup and the different measures used to evaluate the VADs. In Section IV the results of the tests are presented. Section V concludes the paper with an evaluation of the results and an evaluation of the properties of the best performing VAD algorithm.

II. VAD Algorithm Overview

VAD algorithms operate by taking in a digitized audio signal, processing this signal, extracting particular features from the processed signal, passing the extracted features of the signal as parameters to a model that describes that feature in noise and in speech, and finally outputting the decision based on thresholds defined in the model. There are many different features that different VAD algorithms model, commonly used among them being Fourier coefficients, periodicity and zero-crossing rates. Similarly there are various models that VAD algorithms use to describe these features, some based on heuristics while others based on statistical models. Popular statistical models include Gaussian distributions and Laplacian

Manuscript received August 1, 2011. This material is based upon work supported by the National Science Foundation under Grant No. 1063035.

Jonathan Kola is a senior at Harvard College (e-mail: kola@fas.harvard.edu)

Carol Espy-Wilson is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: espy@ umd.edu).

Dr. Tarun Pruthi is an affiliate with the Institute of Systems Research, University of Maryland, College Park, MD 20742 USA (e-mail: t.pruthi@ieee.org).

distributions. Based on the decision rule defined in the model, the VAD outputs a flag to indicate the presence or absence of speech.

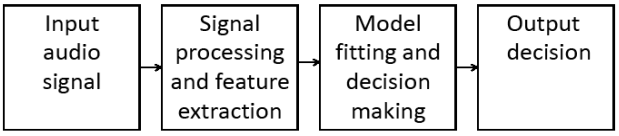

Figure 1. Block diagram for VAD algorithms

A. *Ying et al. 2011(in press)*

D. Ying proposed a VAD algorithm based on an unsupervised learning framework [1]. This particular algorithm was based on a sequential Gaussian Mixture Model (SGMM), and it utilized as its feature parameter the energy distribution in Mel-spaced frequency bands of the signal. The Gaussian Mixture Model used comprised of two Gaussian distributions, each trying to model either noise or speech. The models were trained using an unsupervised learning process, whereby the initial frames from a signal were clustered into the two Gaussians, with the distribution with the lowest mean representing noisy regions and the distribution with the higher mean representing speech regions. The estimated distributions were also used to determine a decision threshold to discriminate speech from non-speech. This algorithm performed the detection process in each sub-band, independently of all other sub-bands, and the results from each sub-band were used to determine the final output through a voting procedure decided by some threshold determined experimentally. A hangover scheme which simply delayed the transition from a speech declaration to a non-speech declaration was also implemented to account for the low energy regions of the tail end of utterances

B. *J. Sohn et al. (1999)*

The much cited Sohn VAD based on a statistical model [2] has a very similar approach to the Ying VAD [1], in terms of its use of the energy in a signal as its primary parameter for model comparison, and its use of Gaussian distributions to model the distribution of the speech and non-speech energies. The Sohn VAD however does not separate the signal into different frequency bands, but takes into account the distribution of the entire spectrum. The algorithm employs the Decision Directed method to estimate the *a priori* SNR in the signal. A likelihood ratio is then computed using the SNR in the current frame and the estimated *a priori* SNR which is then compared to some threshold determined by the distribution model to make the speech/non-speech decision. This algorithm also implements a hangover scheme to prevent the clipping of weak speech tails, however rather than implementing a simple delay in transition from a speech to a non-speech indicator, the hangover scheme is based on a Hidden Markov Model whereby the speech decision of a current frame only depends on the current frame and the previous frame, making the correlation between consecutive speech frames explicit. A major distinguishing factor of the Sohn VAD is the semi-supervised training of its Gaussian model. The noise statistics are estimated by assuming an initial non-speech region in a signal to train a noisy model, which amounts to supervised learning, before subsequent frames are then used to update the model in an unsupervised manner.

C. *ITU G.729B*

The ITU G.729B VAD is an algorithm used widely in comparisons of different VAD algorithms [3]. It uses four features as its parameters, the full and low-band frame energies, the set of line spectral frequencies (which are Linear Predictive Coding (LPC) coefficients), and zero crossing rates. Running averages are calculated of the long-term signal parameters and the characteristic energies of the background noise. Difference measures are then computed that compare the feature parameters in a particular frame to the running averages, and a decision is made based on the union of the speech/non-speech results given by the four different measures. Based on models of each of these features, the final decision is determined by the combination of the decisions made by each model. This VAD also implements a hangover scheme to smooth the voice activity decision.

D. *ETSI AMR VAD*

There are two implementations of the ETSI AMR VAD, option 1 (AMR 1) and option 2 (AMR 2) [4].

The AMR 1 algorithm operates by first separating the audio signal into different frequency bands and then detecting pitch and tone (variations in pitch) presence in the sub-bands as indications of speech. A hangover scheme is added to account for low power endings of speech bursts.

The AMR 2 algorithm uses as its feature parameter the energy in a particular sub-band and its power spectral density estimate. The sub-band energy in a current frame is compared to long-term energy estimates and a decision is made based on the SNR difference measure. Running estimates of the background noise are computed based on the deviation of the spectral density in order to provide an adaptive measure of the SNR. This algorithm also provides a hangover scheme to smooth the decision making process.

The Ying, Sohn, ITU G.729B, AMR 1 and AMR 2 VADs were subsequently tested against the same database and for the same measures to provide a comprehensive comparison. All these VADs represented both the heuristic approach and the statistical approach to decision making, and they used a variety of features and combinations of features as the parameters in their models The Sohn VAD is based on robust statistical modeling and hypothesis testing which is a novel approach that many subsequent algorithms have been based on. The Ying VAD algorithm refined the Sohn algorithm by introducing a completely unsupervised learning training method for the statistical models which avoided the problem of non-speech assumptions at the beginning of audio signals. The ITU G.729B and the AMR VADs use different features in combination and base their decision on heuristics. These VADs therefore to a large extent represent a comprehensive set of the most commonly used VAD approaches, and the results of the tests give a good indication of the most robust approaches to voice activity detection for future application.

III. Test Setup

The five VADs were run across a database of utterances in 22 different noises and at 7 different SNRs (-12dB, -3dB, 0dB, 3dB, 6dB, 12dB, 18dB). For each noise type and SNR there were 6 phonetically diverse utterances from the TIMIT database, 1 for each of 6 speakers comprising of 3 males and 3 females. The ground truth to which their output was compared to was based on the phonetic transcriptions of the utterances provided with the TIMIT database. Only the begin/end markers (h#) and the pause markers (pau) were considered to be non-speech, with every other transcription indicating speech. VAD performance is typically measured by its accuracy in detecting speech as speech, and in detecting noise as noise. These measures are defined as the speech hit rate and non-speech hit rate respectively. However, in the event that a VAD makes an error, different errors can be considered more acceptable than others depending on the purpose. For example, in VAD performance an error that calls a noisy region a speech region may be more desirable than one which calls a speech region a noisy region. Measures to characterize different errors as described below, outlined by A. Davis et al. 2006 [6], are therefore used as well to fully describe and evaluate performance of different VADs.

- *Front End Clipping (FEC):* Errors that misclassify the beginning of an utterance or a word as non-speech.
- *Mid Sentence Clipping (MSC):* Errors that classify regions in the middle of sentences as non-speech.
- *Overhang error (OVER)*: Errors that misclassify non-speech regions at the end of utterances as speech due to the implementation of the hangover scheme.
- *Noise detected as speech (NDS)*: Errors whereby noisy regions are misclassified as speech.
- *Speech hit rate:* Percentage of correct declarations of speech.
- *Non-speech hit rate:* Percentage of correct declarations of non-speech.
- *Average hit rate:* Average of speech hit rate and non-speech hit rate.

The most meaningful measure in this set up was the average speech/non-speech hit rate. This measure was a simple average of the speech and non-speech hit rates in a particular noise type at a particular SNR on a scale from 0 to 1, with scores closer to 1 indicating better accuracy. This measure however did not explicitly take into account any preferences to speech bias, as the weights of the speech and non-speech hit rates in computing the average were equal. It was however observed from experimental data that all the VADs were biased towards correctly classifying speech. Therefore in their calculations the average hit rate was already reflecting the implicit bias.
Another measure considered was the consistency in performance of the VADs in different noise types at the same SNR. VADs that have comparable performances in different noise types at the same SNR may be better suited to certain applications than VADs whose performance varies greatly with noise type at the same SNR.

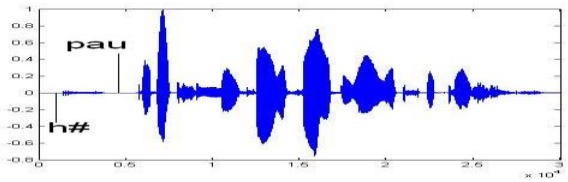

Figure 2.
Pauses (pau) and begin/end markers (h#) are declared non-speech. Every other transcription is declared speech.

IV. Results

Based on the test set up, the Ying VAD emerged as being the most accurate and most consistent algorithm against the utterances in the different noise types and different SNRs.
It was consistently among the best performers at all SNRs. The second best performer was the AMR 2 VAD, followed by the Sohn VAD, ITU VAD and the AMR1 VAD.

Table 1.
VAD average speech/non-speech hit rates

	Ying	AMR 2	Sohn	ITU	AMR 1
-12 dB	0.4768	0.2687	0.4139	0.3133	0.5174
-3 dB	0.6351	0.5986	0.5484	0.3974	0.5260
0 dB	0.6795	0.694	0.6006	0.5133	0.5141
3 dB	0.7148	0.7411	0.6496	0.5570	0.5039
6 dB	0.7524	0.7637	0.6956	0.5962	0.5009
12 dB	0.8057	0.7790	0.7606	0.6575	0.5031
18 dB	0.8454	0.7890	0.8029	0.7055	0.5164
Average Hit Rate	0.7014	0.6620	0.6388	0.5343	0.5117

Table 2.
Speech and non-speech hit rates in clean speech

	Ying	AMR2	Sohn	ITU	AMR1
Speech hit rate	0.9864	0.9665	0.9879	0.9960	0.9970
Non-speech hit rate	0.7177	0.6482	0.6628	0.4663	0.5045
Average speech/non-speech hit rate	0.8521	0.8074	0.8254	0.7312	0.7508

Most of the errors in the Ying VAD were due to front end clipping, and the least errors were due to the overhang scheme as shown in Table 3.

Table 3.
Ying VAD Error Summary

	FEC	MSC	OVER	NDS
-12dB	0.2277	0.0251	0.0018	0.0701
-3dB	0.1352	0.0184	0.0015	0.0589
0dB	0.1129	0.1629	0.0015	0.0552
3dB	0.0956	0.0213	0.0016	0.0519
6dB	0.0796	0.0139	0.0015	0.0472

12dB	0.0622	0.0117	0.0015	0.0369
18dB	0.0455	0.0089	0.0017	0.0322
Average	0.1084	0.0166	0.0016	0.0504

The accuracy of the VADs improved as the SNR progressed from the very low level of -12dB to 18dB and clean signals as shown in Figure 3.

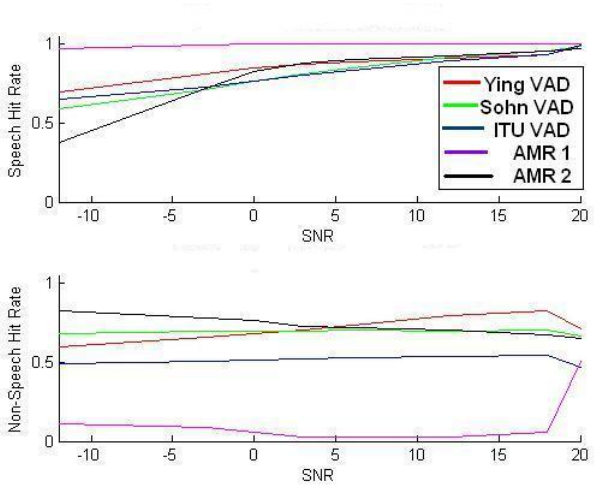

Figure 3. Speech/non-speech hit rates across SNRs

All the VADs performed poorly in music2 type noise (periodic noise) and all VADs except the Ying VAD performed worst in pass type noise. In the Ying VAD, the best results were recorded in pass noise. Despite poor performance in music2 noise, the VADs performed well in other periodic noises such as babble noise and music1 noise (music1 noise is instrumental (heavy metal), music2 noise is lyrical (reggae)), therefore the performance of the VADs was not generally worse in periodic noises, though the worst performance was recorded in a periodic noise. The best performances were in a variety of noises, however the VADs achieved the overall best results in the fire60nosiren noise type. (See Figure 6 – 10 (Note different scales for presentation clarity)).

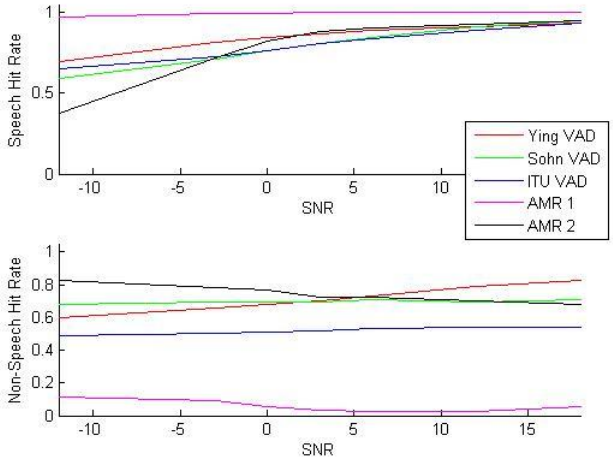

Figure 4. Speech/non-speech hit rate in aperiodic noise

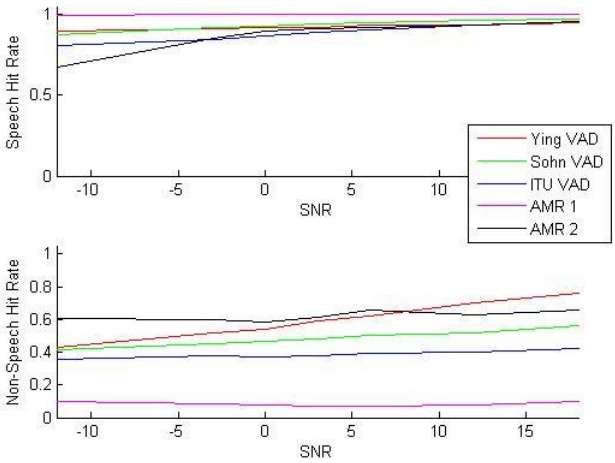

Figure 5. Speech/non-speech hit rate in periodic noise

The performance over different noise types also varied with VADs, with some experiencing more variation over different noise types at a particular SNR than others. Table 3 shows the variance of the speech/non-speech hit rates caused by different noise types at a particular SNR.

Table 4.
Speech/non-speech hit rate variance (x 10^{-3})

	Ying	Sohn	AMR 2	ITU	AMR 1
-12 dB	14.1	36.9	51.1	7.8	0.5
-3 dB	5.2	23.0	13.8	3.5	0.9
0 dB	4.0	16.4	5.4	1.7	0.3
3 dB	3.8	9.9	2.4	0.85	0.08
6 dB	2.6	5.9	0.9	0.82	0.0083
12 dB	1.7	3.4	0.1	1.9	0.099
18 dB	0.6	2.5	0.04	3.0	1.6
Average (from 0 – 18db)	2.54	7.62	1.77	1.65	0.42

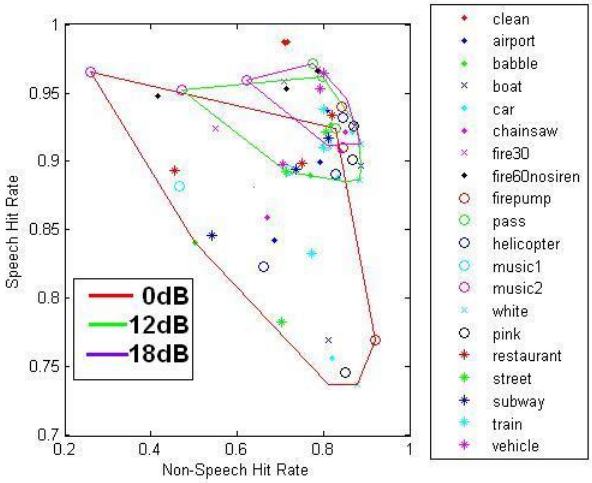

Figure 6. Ying VAD hit rates in different noise types at different SNRs

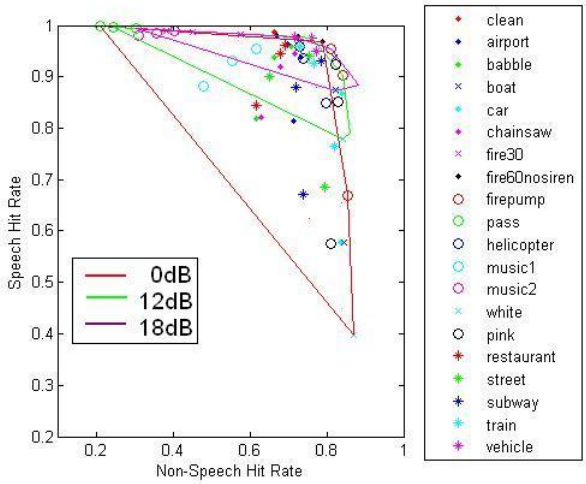

Figure 7. Sohn VAD hit rates in different noise types at different SNRs.

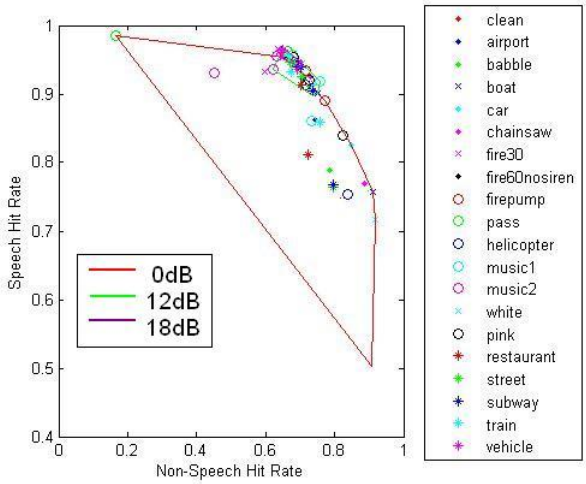

Figure 8. AMR 2 hit rates in different noise types at different SNRs

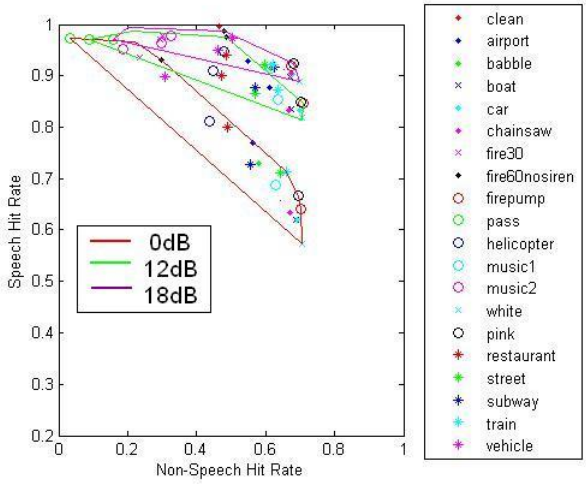

Figure 9. ITU G.729B hit rates in different noise types at different SNRs

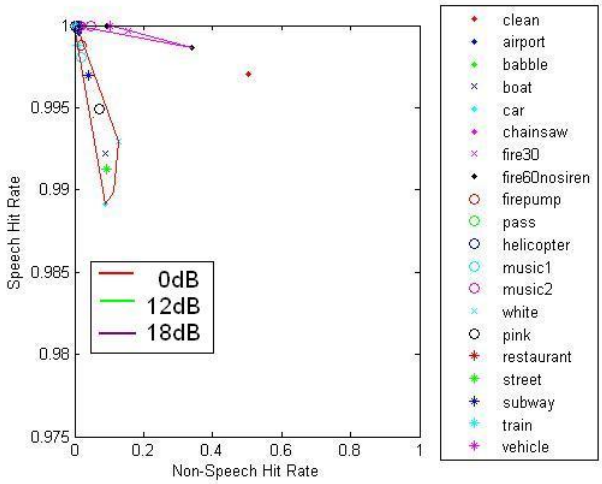

Figure 10. AMR 1 hit rates in different noise types at different SNRs.

V. Conclusion

The approach outlined by D. Ying et al. achieved the best and most consistent performance at all SNRs and different noise types with a comparatively low variance in performance in different noise types.

The unsupervised learning of the models proved to be a robust approach to modeling parameter distributions, while the sub-band level decision-making process lead to greater accuracy. The overhang scheme implemented was also very robust as there were very few errors associated with it.

There are however some concerns about the Ying algorithm. The first concerns the startup time of the algorithm, which is greater than 0.5s due primarily to the unsupervised model training which, in the default setting, requires 60 10ms frames in order to converge to an accurate model. This time lag may be inappropriate for some applications of the VAD. Another concern is the computational load of the algorithm. Although the decision making process in each sub-band of the signal is very efficient, as a whole, carrying out the same process for several bands becomes a high computational cost, which similarly may be undesirable in certain VAD applications. Despite these shortcomings, the Ying VAD was shown to perform exceptionally well against other VADs in a standard testing framework, and the potential of the unsupervised learning framework in voice activity detection has been demonstrated as being a high performing and robust approach.

Acknowledgment

J. Kola thanks D. Ying for providing a MATLAB implementation of the VAD outlined in [1].

References

[1] D. Ying, Y. Yan, J. Dang and F. Soong, "Voice Activity Detection Based On An Unsupervised Learning Framework (Accepted for publication)," *IEEE Transactions on Audio, Speech and Language Processing*, to be published.

[2] J. Sohn, N.S Kim and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1,pp. 1-3, Jan. 1999.

[3] ITU, "Coding of speech at 8 kbit/s using conjugate structure algebraic code-excited linear –prediction (CS-ACELP). Annex B: A silence

compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," International Telecommunication Union, 1996.

[4] ETSI, "Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," ETSI EN 301 708 Recommendation, 1999.

[5] LDC - Linguistic Data Consortium. "TIMIT Acoustic-Phonetic Continuous Speech Corpus." <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1>.

[6] A. Davis, S. Nordholm and R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation And An Adaptive Threshold," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 412-423, March 2006.