A NON-CAUSAL APPROACH TO VOICE ACTIVITY DETECTION IN ADVERSE ENVIRONMENTS USING A NOVEL NOISE ESTIMATOR

A. Esmaeili, S.M. Ahadi & M.A. Fassihi

Electrical Engineering Department, Amirkabir University of Technology
Hafez Avenue, Tehran 15914, Iran

ABSTRACT

Analyzing the characteristics of the LR-based VAD, it was found that the delay associated with the decision directed (DD) *a priori* SNR estimator can lead to detection errors at speech onsets and offsets. In this paper the properties of a non-causal estimator, used before in a speech enhancement system, are investigated. It is shown that the application of the non-causal estimator improves the robustness of the VAD in noisy environments, specifically at low SNRs. In addition, the associated noise estimation procedure has been further improved by the application of a dynamic time varying smoothing factor. Objective tests conducted based on speech/non-speech discrimination show that the proposed VAD outperforms standard VAD algorithms, including ETSI-VADNest, AMR1, AMR2, and also the statistical VADs based on smoothed LR and multiple observation LR, specifically at low SNRs, at the cost of some delay.

Index Terms— Voice Activity Detection (VAD), likelihood ratio, non-causal estimation.

1. INTRODUCTION

Voice activity detection, being a crucial part of many speech processing applications such as speech enhancement [1], variable rate speech coding [2] and speech recognition [1,9], has been the field of study for many researchers. Traditional VADs usually use a combination of different speech features such as short term energy, zero crossing rate, periodicity measures [3], etc. as their decision parameter. While older approaches toward speech detection were based on heuristics, a trend toward robust statistical algorithms has been established in the past few years [4-9]. Recently, successful attempts have been made in developing voice activity detectors based on the likelihood ratio (LR) test, employing different statistical models for the speech and noise [5-9]. In [5] Sohn and Sung developed a LR based VAD assuming speech and noise as Gaussian random processes. An improved version of the VAD was introduced in [6], employing the decision directed (DD) estimate of the *a priori* SNR instead of the ML estimate used in [5]. A novel hangover scheme based on the hidden Markov model (HMM) was used to improve speech detection rate at weak speech tails. However, the VAD's error rate at speech onsets and offsets was rather high. Analyzing the LR as a function of its variables, the cause was found to be the delay term in the DD estimator [7]. To alleviate the problem they suggested the smoothing of the LR. On the other hand, the application of future signal measurements has been found to be beneficial in improving the performance of speech processing systems and to meet the high level of performance required by modern speech processing systems [9-11]. Adopting this approach, a non-causal (NC) a priori SNR estimator was developed for the application of speech enhancement in [11]. In this paper the properties of this NC estimator and its application to voice activity detection is investigated. Also the noise estimator used in [6,7] has been further improved by the application of a dynamic time varying smoothing factor. Finally the performance of the proposed VAD is evaluated and compared to standard VADs (ETSI-VADNest, AMR1&2) and the smoothed likelihood ratio (SLR) and multiple observation likelihood ratio (MO-LR) VADs presented in [6,7] by speech/non-speech discrimination analysis and ROC curves.

2. DESCRIPTION OF THE NON-CAUSAL *A PRIORI* SNR ESTIMATOR

To avoid the delay associated with the DD *a priori* SNR estimator that results in the attenuation of speech onsets by the speech enhancement algorithm, Cohen proposed a non-causal (NC) estimator that takes advantage of future spectral measurements [11]. Adopting the Gaussian model for speech and noise with respective variances of λ_X and λ_N, the proposed NC estimator used the optimal MMSE spectral power (SP) estimate of the signal, derived by Wolfe & Godsill [14], for estimating the *a priori* SNR.

$$\hat{\xi}_{NC} = G_{SP}^2(\xi'_X, \gamma) \times \gamma, \qquad (1)$$

$$G_{SP} = \sqrt{\frac{\xi'}{(1+\xi')\gamma}\left(1+\frac{\gamma\xi'}{1+\xi'}\right)} \qquad (2)$$

where $\xi = \lambda_X / \lambda_N$ and $\gamma = |Y|^2 / \lambda_N$ are the *a priori* and *a posteriori* SNRs respectively. It was proved that due to the statistical independence of observations assumed in the model, employing future spectral measurements into the system leaves the speech enhancement algorithm intact and only the *a priori* SNR estimator should be modified accordingly. It should be noted that to compute the gain function (G_{SP}), an estimate of the *a priori* SNR (ξ') should be at hand already. To obtain a feasible estimator, this parameter was replaced by a weighted sum of three terms. 1) The enhanced spectral amplitude of the previous frame divided by the noise variance (similar to the DD estimator). 2) The NC estimate obtained for the *a priori* SNR in the previous frame ($\hat{\xi}_{NC}(t-1)$). 3) A smoothed version of the *a posteriori* SNR both in time and frequency (ξ''). So we have:

$$\xi'_k(t) = 0.8\frac{\left|\hat{X}_k(t-1)\right|^2}{\lambda_N} + 0.16\sum_{i=-\omega}^{\omega} b(i)\hat{\xi}_{NC}(k,t-1) + 0.04\xi''_k(t) \quad (3)$$

where k indicates the spectral bin index, t is the time frame index and $b = [0.25 \quad 0.5 \quad 0.25]$ is the normalized window used for local averaging. The third factor, ξ'', is computed as follows:

$$\xi''_k(t) = \frac{\sum_{(i,j)\in\Gamma} b(i)\left|\gamma_{k-i}(t+j)\right|^2}{\sum_{(i,j)\in\Gamma} b(i)} - 1 \quad (4)$$

where $\Gamma = \{(i,j) \mid -\omega \le i \le \omega, 0 \le t \le L, (i,j) \ne 0\}$ designates the time-frequency indices of the measurement. To ensure the positiveness of ξ'' and reduce the musical noise phenomenon, the parameters ξ'' and ξ' were constrained to be larger than zero and $\xi_{\min}$, respectively, as suggested by Cappe [13]. Compared to the DD estimator, the second term used in (3) is new and yields smoother estimates due to its recursive nature. The third term, ξ'', replaces the ML estimate of the *a priori* SNR (i.e. $\xi = \gamma - 1$) used in the DD estimator [12]. Since this factor is computed over the future spectral values we predict a faster response to spectral variations for the NC estimator than the DD estimator.

3. NON-CAUSAL VOICE ACTIVITY DETECTION

The drawbacks considered with traditional, heuristically motivated VADs have caused a trend toward developing robust statistical algorithms by researchers in the past few years. Recently, successful attempts have been made in developing voice activity detectors based on the likelihood ratio test, employing Gaussian statistical models for the speech and noise [5-7]. Forming the likelihood ratio for the kth frequency bin under the assumed model, we have [6]:

$$\Lambda_k = \frac{1}{1+\xi_k}\exp\left\{\frac{\gamma_k\xi_k}{1+\xi_k}\right\} \quad (5)$$

The unknown parameters of the LR are the noise variance, λ_N, and the *a priori* SNR, ξ, where the first is obtained through the noise estimation procedure (described in section 4) and the latter was computed using the DD *a priori* SNR estimator. The final VAD decision parameter will be the geometrical mean of the LRs over all frequency bins. Analyzing the LR as a function of its variables, ξ and γ, it can be seen that the LR reaches its maximum at $(\gamma_{\max}, \xi_{\max})$ while it becomes minimum at $(\gamma_{\min}, \xi_{\max})$, and not $(\gamma_{\min}, \xi_{\min})$. Considering the delay of the DD estimator, we may expect misdetections at speech onsets and offsets. While at speech onsets γ increases, ξ can remain low thus limiting the rise of the LR (this may cause more problems at low SNRs where the increase in γ may not be enough to raise the LR above the decision threshold). On the other hand, at low energy speech tails the instantaneous SNR, γ, declines but ξ may remain high, thus, the LR may fall to a value near to its minimum. The solutions presented so far consist of an HMM-based hangover scheme [6] and smoothing the LR [7]. In fact Cho *et al.* considered designing an adaptive α for the application in the DD estimator. However, their efforts did not yield a generalized adaptive rule. Thus smoothing the LR was considered as an alternative [7] to slow the decay of the VAD's decision parameter and avoid misdetections at

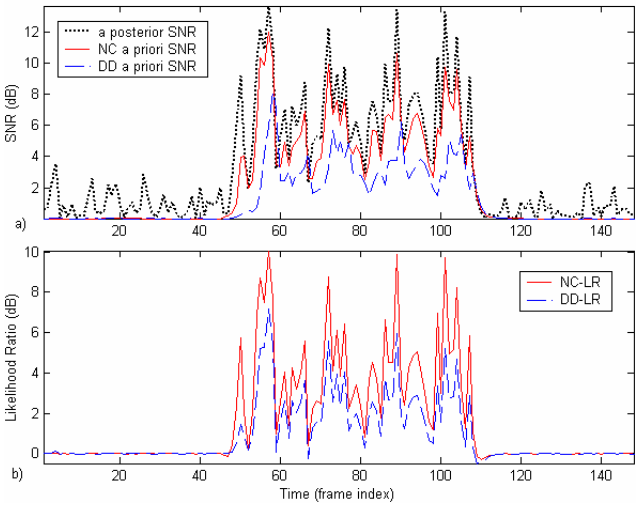

Figure 1: a) SNRs in successive frames for the sample signal corrupted with pink noise at 0dB SNR. b) The LR computed using the NC and DD a priori SNR estimators.

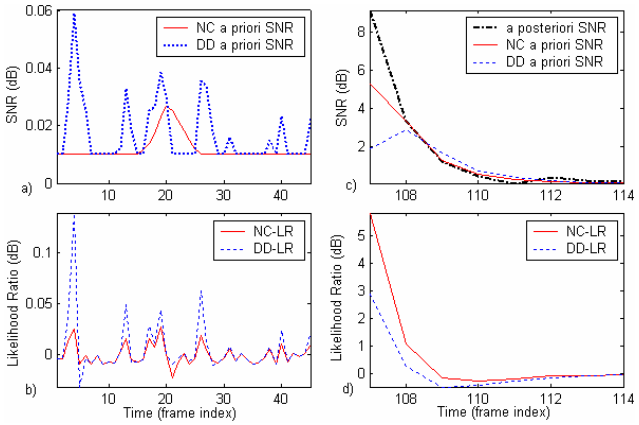

Figure 2: Enlarged views of the *a priori* SNR and LR at the frames containing noise only (a & b) and speech offset (c & d).

speech offsets. To better understand the properties of the DD and NC estimators and their impact on the computed LR a sample signal containing a sinusoidal component was chosen and pink noise was added to it artificially at 0dB SNR. The signal contains only noise in the first 50 and last 40 frames while a sinusoidal component, 60 frames long, comes in the middle. The estimates obtained for the *a priori* SNR using the DD and NC estimators and the corresponding LRs are demonstrated in figures 1(a) and (b), respectively. Comparing the figures, the NC estimator has been helpful in improving the VAD based on the LR in the following ways:

1) Smoother results are obtained by the NC estimator for ξ during noise-only periods causing not only less musical noise in the enhanced signal but also less fluctuations in the calculated LR, thus, causing less false alarms. Figures 2(a) & (b) show an enlarged view of the ξ estimates and their corresponding LRs during the first 45 frames. It is clear that the NC-LR fluctuates less than the DD-LR in noise only frames.

2) At speech onsets where the DD estimator can not respond too fast, the estimate obtained through the NC estimator tracks γ

quickly and consequently the LR increases faster avoiding speech misdetections, Figures 1(a) & (b).

3) At low energy speech tails the NC estimates for ξ track γ consistently, avoiding the delay associated with the DD estimator. Thus the fall in LR is slightly less than what can be seen with the DD estimator. This is better seen in the enlarged view of frame numbers 107 to 114, shown in Figures 2(c) & (d).

4) During speech active frames the NC estimates of ξ track γ without the one frame delay associated with the DD estimator, resulting in stronger peaks in the computed LR since, γ and ξ rise almost simultaneously contributing to a strong maximum in the computed LR.

Still, smoothing the computed NC-LR is useful, since it reduces the variance of the VAD's decision parameter, improving its performance in noisy environments. It should be noted that the application of future spectral measurements imposes a delay on the VAD 's decision. To determine the optimal value for the number of subsequent frames used in the NC estimator, the VAD's accuracy was evaluated as a function of the delay term (L) on a subset of the evaluation database. As it can be seen in Figure 3, increasing the delay term to 4 frames significantly improves the VAD's performance. Increasing the delay over this value does not result in further improvements. In [9] a VAD called MO-LR was developed using more observations for forming the LR test. The concept leads to a moving average like smoothing of the LR values over a window of $2w+1$ observations which can be implemented efficiently. This algorithm has also a delay. We will show that the proposed NC-LR based VAD outperforms the SLR, MO-LR and the standard VAD methods.

4. IMPROVED NOISE ESTIMATION

A sensitive part of the VAD is its noise estimation algorithm. To estimate the time varying noise statistics a novel soft-decision technique was developed in [5]. The optimal estimate of the variance of the background noise, λ_N , was obtained in an MMSE sense as

$$\begin{aligned} E\{\lambda_N|Y\} &= E\{\lambda_N|H_0\}P(H_0|Y)+E\{\lambda_N|H_1\}P(H_1|Y) \\ &= Y^2\times P(H_0|Y)+\lambda_N(t-1)P(H_1|Y) \end{aligned} \quad (6)$$

where $P(H_0|Y)$ is the speech absence probability. This can be calculated by the Bayes rule as

$$P(H_0|Y)=\frac{1}{1+\frac{P(H_1)}{P(H_0)}\Lambda} \quad (7)$$

where $P(H_0)$ is the *a priori* speech absence probability, $P(H_1)$ is its complement and Λ is the VAD soft decision parameter. The estimation of $P(H_0)$ has attracted much attention more recently [7,15]. Here it is estimated in an adaptive manner as given in [7]. Our observations of the noise estimator showed that although the noise estimator was fast in following the abrupt changes of the noise signal, it included parts of the speech energy in our estimate of the noise during speech periods. This was more apparent during speech onsets and weak speech tails where, as mentioned earlier, the LR may take small values, thus $P(H_0|Y)$ may increase causing a portion of the speech energy (Y^2) be absorbed in our estimate of noise variance according to (6). This artifact results in misdetection of such frames, but moreover, it causes more attenuation of the signal by the speech enhancement algorithm [12] due to the increase of the noise level (so ξ is further reduced). Thus, the VAD misses to detect the following speech frames and consequently the noise level further increases and causes the VAD to fail to detect even some speech with rather high energy. Although not directly mentioned, to alleviate this problem, Cho *et al.* used the smoothed likelihood ratio (SLR) instead of the LR itself in (7) and further smoothed the value obtained from (6) with a constant smoothing factor [7]. While the proposed solution works well in rather high SNR conditions (+5dB and up, that is the same SNR range used in [7] for their analysis), as the SNR decreases to 0dB and below, the same disturbing artifact is faced. Since γ has an exponential distribution with a mean of one [12], we may consider any deviation of $\bar{\gamma}$ from its expected value as speech or a highly non-stationary noise. In order to overcome the problem, we propose the application of a dynamic time varying smoothing parameter in the recursive averaging as follows

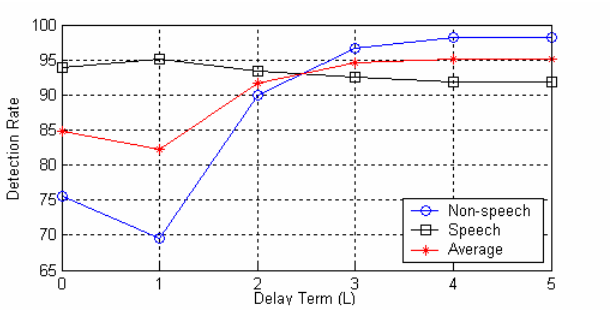

Figure 3. VAD's performance as a function of the delay term (L).

$$\hat{\lambda}_N(t)=\alpha_d\hat{\lambda}_N(t-1)+(1-\alpha_d)E\{\lambda_N|Y\} \quad (8)$$

$$\alpha_d = 0.92+0.05|\bar{\gamma}-1| \quad (9)$$

Where $\bar{\gamma}$ is the smoothed version of γ, with a constant factor of 0.95 and α_d is limited to 0.98. Through speech active frames $|\bar{\gamma}-1|$ may rise, thus increasing the dynamic smoothing parameter avoiding the absorption of speech energy into our estimate of noise. In addition, the proposed NC-LR takes larger values in speech onsets and offsets than the LR, alleviating the aforementioned problem in such regions according to (6,7). Our investigation of the VAD performance showed that the application of the smoothed NC-LR and the dynamic smoothing parameter, compared to a constant smoothing parameter [7] leads to significant improvements in terms of speech detection rate, specifically at low SNRs.

5. PERFORMANCE EVALUATION

The proposed VAD was extensively evaluated by means of objective tests based on speech detection rate (SDR) and non-speech detection rate (NDR) and the results were compared to standard VADs including AMR1, AMR2 [2] and ETSI-VADNest [1] as well as the smoothed and multiple observation likelihood ratio-based VADs presented in [7] and [9], respectively. It should be mentioned that the proposed modifications made to the VAD were separately evaluated. Speech data of about 440 sec duration from two female and two male speakers have been taken from the persian AKHBAR database. The database is collected at the Speech lab, Electrical Engineering Department, Amirkabir

University of Technology from the local radio news broadcasts. A set of four noises were chosen from the NATO RSG-10 database [17] (white, pink, car and f16) and artificially added to the speech at SNRs of 15, 10, 5, 0 and -5 dB. The clean data was manually labeled as speech or silence and used as a reference for discrimination analysis. Finally the above mentioned VADs were applied to the noisy data. SDR/NDR is computed by the division of the true number of speech/noise detected frames to the total number of frames labeled as speech/noise manually. The results for the working point of the VADs are presented in Table1. As it can be seen the multiple observation likelihood ratio (MO-LR) based VAD (where the window length was set to 17 frames, thus, imposing an eight frame delay on the VAD's decision) has improved the detection rates over the smoothed likelihood ratio (SLR) based VAD. Yet the non-causal (NC) approach outperforms both of them. This is specifically noticeable in environments with non-stationary noise. Another test that is commonly conducted to reveal the tradeoff between speech detection (P_d=SDR/100) and false-alarm (P_f=1-NDR/100) probabilities is the receiver operating characteristic (ROC) curves. The ROC curves, for white, pink and f16 noise at 0dB, are shown in figure 4 where each of the curves are found by inspecting the VAD's performance (P_d & P_f) as its threshold changes. The multiple observations LR based VAD which uses more observations for making the decision as suggested in [9] shows improved performance over the SLR based VAD. Further improvements can be achieved by applying the proposed changes to the noise estimator used in [7]. Meanwhile, the NC-LR, having a delay of 4 frames, is still superior to all, including the standard VADs. Similar results hold over the rest of the SNR range and noise types.

6. CONCLUSION

Voice activity detection has become a crucial part of many speech processing applications. Analyzing the characteristics of the LR-based VAD, it was found that the delay associated with the DD *a priori* SNR estimator can lead to detection errors at speech onsets and offsets. Cho *et al.* smoothed the LR to alleviate the detection errors [7]. While the proposed solution works well in rather high SNR conditions (+5dB and up, that is the same SNR range used in [7] for their analysis), as the SNR decreases to 0dB and below the SLR-based VAD makes numerous detection errors. In this paper the properties of a NC *a priori* SNR estimator, used before in a speech enhancement system, is studied. It is shown that the application of the NC estimator improves the robustness of the VAD in noisy environments, specifically at low SNRs. Since the proposed solution takes advantage of subsequent spectral measurements, the algorithm has a delay of a few frames. Moreover, the soft noise estimation technique was investigated and further improved by the application of a dynamic time varying smoothing parameter. This prevents our noise estimate to capture speech energy during weak parts of speech that occurs frequently at low SNRs. It has been found that the proposed VAD outperforms standard VAD algorithms, including ETSI-VADNest, AMR1, AMR2, and also the statistical VADs based on SLR and MO-LR [9].

Due to the simultaneous high speech/non-speech hit rates of the proposed VAD, we expect it to be a very good choice for frame dropping in a typical speech recognition system which is not sensitive to the VAD's delay. However this needs to be further verified.

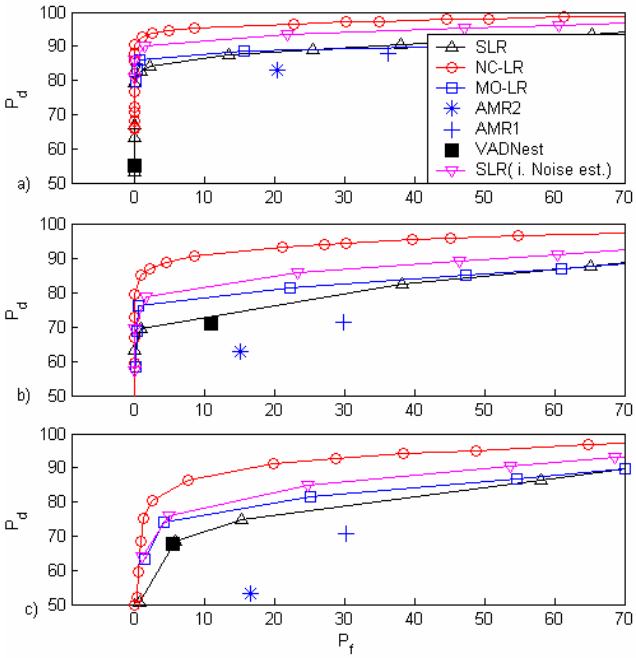

Figure 4. Roc curves for VADs at 0dB SNR: (a) white noise, (b) pink noise, (c) F16 noise.

7. REFERENCES

[1] ETSI, "Speech Processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm," *ETSI ES 202 212 v1.1.1*, Nov. 2003.

[2] ETSI, "Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, 1999.

[3] S.G. Tanyer and H. Özer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 478–482, July 2000.

[4] B. Ahmed and W. Holmes, "A Voice Activity Detector Using The Chi-Square Test" in *Proc. ICASSP*, vol. 2, pp. 737–740, 2001.

[5] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. ICASSP,* 1998, pp. 365–368.

[6] 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.

[7] Y.D. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Proc. ICASSP*, vol. 2, pp. 737–740, 2001.

[8] S. Gazor and W. Zhang, "A Soft Voice Activity Detector Based on a Laplacian–Gaussian Model," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, Sep. 2003.

[9] J. Ramírez, J.C. Segura, C. Benítez, A. Torre, and A.J. Rubio, "Statistical Voice Activity Detection using a Multiple Observation Likelihood Ratio Test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, 2005.

[10] ____, "Efficient voice activity detection algorithms using long-term speech information", *Speech Communication*, vol. 42, pp. 271-287, 2004.

[11] I. Cohen, "Speech enhancement using a noncausal *A Priori* SNR estimator," *IEEE Signal Process. Letters*, vol. 11, no. 9, pp. 725–728, Sep. 2004.
[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-32, pp. 1109–1121, 1984.
[13] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2, pp. 345–349, Apr. 1994.
[14] P.J. Wolfe and S.J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. 11th IEEE Workshop Statist. Signal Processing*, Singapore, pp. 496–499, Aug. 6–8, 2001.
[15] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.
[16] ETSI "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunication Systems (UMTS); AMR speech codec; C-source code," *ETSI TS 126 073 V6.0.0*, 2004.
[17] Available from http://spib.rice.edu/spib/select_noise.html.

Environment		VADNest		AMR1		AMR2		SLR		MO-LR		**NC-LR**	
Noise	SNR	SDR	NDR	SDR	NDR	SDR	NDR	SDR	NDR	SDR	NDR	**SDR**	**NDR**
White	15	91.88	99.82	91.95	68.47	97.70	77.55	98.50	99.14	92.93	99.53	**99.02**	**97.95**
	10	90.57	99.91	92.21	68.70	97.55	76.69	96.94	99.62	92.46	99.56	**98.27**	**98.41**
	5	86.73	99.94	90.75	67.28	95.85	77.60	92.79	99.83	91.56	99.89	**96.59**	**98.77**
	0	54.95	100.00	87.88	63.80	83.05	79.55	79.23	100.00	88.54	99.95	**92.78**	**98.98**
	-5	11.06	100.00	81.26	54.98	33.92	91.25	50.19	100.00	72.05	100.00	**79.26**	**99.06**
Pink	15	93.41	88.91	92.59	70.75	97.65	77.15	97.30	99.54	92.83	99.56	**98.80**	**98.19**
	10	92.53	88.94	90.99	70.07	97.44	77.49	93.52	99.84	91.76	99.59	**97.24**	**98.66**
	5	89.92	89.01	87.08	69.37	93.88	78.65	83.43	99.97	89.72	99.96	**94.17**	**98.98**
	0	71.03	89.11	71.30	70.22	62.83	84.97	62.99	100.00	78.90	100.00	**84.88**	**99.13**
	-5	26.94	89.06	57.33	67.24	18.03	96.66	33.98	100.00	45.90	100.00	**56.96**	**99.08**
Vehicle	15	99.95	6.36	99.53	76.91	99.98	75.92	99.9870	84.5510	100.00	92.06	**99.99**	**84.48**
	10	99.94	6.28	99.76	76.10	99.93	75.29	99.95	85.33	100.00	92.15	**99.97**	**84.68**
	5	99.94	6.26	99.61	73.45	99.87	75.43	99.74	86.52	99.83	92.38	**99.91**	**85.44**
	0	99.95	6.26	98.53	71.92	99.34	76.25	99.14	87.63	99.16	92.88	**99.81**	**86.64**
	-5	99.95	6.26	96.75	71.54	98.53	77.70	97.66	88.52	97.02	93.63	**99.37**	**87.74**
F16	15	93.14	94.26	94.06	70.30	97.89	76.42	97.73	94.24	97.08	86.97	**99.43**	**73.46**
	10	91.90	94.43	91.87	70.06	97.01	76.97	94.16	94.42	94.88	90.40	**98.61**	**77.15**
	5	88.76	94.50	86.81	69.63	90.54	79.58	85.88	94.33	92.28	92.46	**96.64**	**80.17**
	0	67.76	94.58	70.63	69.79	53.16	83.39	68.33	94.18	84.97	92.53	**91.11**	**80.19**
	-5	24.97	94.54	65.28	55.06	17.16	94.66	46.33	93.99	62.30	92.16	**75.40**	**79.54**

Table 1. Detection rates for the proposed and benchmark VADs for various environmental conditions