IMPROVED VOICE ACTIVITY DETECTION BASED ON A SMOOTHED STATISTICAL LIKELIHOOD RATIO

Yong Duk Cho, Khaldoon Al-Naimi, and Ahmet Kondoz

Centre for Communication Systems Research
University of Surrey
Guildford, Surrey GU2 7XH, UK
{Y.Cho, K.Al-Naimi, A.Kondoz}@eim.surrey.ac.uk

ABSTRACT

This paper presents the behavioural mechanism of a statistical model-based voice activity detector (VAD), featuring a likelihood ratio test for the activity decision. From investigation of the VAD, it is found that detection errors could occur frequently at speech offset regions because of the delay term in the decision-directed parameter estimator, employed for the estimation of an unknown parameter of the likelihood ratio. Hence, this paper proposes a smoothed likelihood ratio so as to alleviate the detection errors at the offset region. Objective test results show that the proposed scheme is useful for achieving a considerable performance improvement for the VAD. Additionally, the proposed VAD gives detection performances superior to G.729B VAD and comparable with AMR VAD option 2.

1. INTRODUCTION

Silence compression by means of voice activity detection can provide a significant improvement in the channel capacity for bandwidth-limited communication systems, such as mobile and packet-based networks. Many algorithms for the speech detection were proposed with the growth of mobile communications, and some of them were adopted as standard methods [1][2]. Traditionally, the detection is performed on the basis of heuristics, thus it is not easy to analyse the characteristics of the detector and to improve the performance.

Recently, Sohn et al. proposed a novel voice activity detector (VAD) based on a statistical model, and reported that it can produce a high detection accuracy [3]. The reason for the high performance is attributed to the adoption of Ephraim and Malah's noise suppression rules [4] for the voice activity decision rules, conducted by a likelihood ratio test using a decision-directed parameter estimator for an unknown parameter.

From investigation of the VAD, however, it is observed that the VAD may generate relatively high numbers of detection errors at the offset region of speech signals. Sohn et al. have circumvented this problem by a hangover scheme, but the reason for the undesirable phenomenon is not mentioned. In this paper, we analyse the behavioural characteristics of Sohn et al.'s VAD, identify the rationale of the unwanted phenomenon, and then propose a solution enabling a significantly improved VAD.

2. DESCRIPTION OF DECISION RULES BASED ON LIKELIHOOD RATIO TEST

Voice activity decision can be considered as a test of two hypotheses: H_0 and H_1, which indicate speech absence and presence, respectively. Assuming that each spectral component of speech and noise has complex Gaussian distribution [4], in which the noise is additive and uncorrelated with the speech, the conditional probability density functions (PDF) of a noisy spectral component Y_k, given $H_{0,k}$ and $H_{1,k}$, are

$$p(Y_k|H_{0,k}) = \frac{1}{\pi\lambda_{N,k}} \exp\left\{-\frac{|Y_k|^2}{\lambda_{N,k}}\right\} \quad (1)$$

$$p(Y_k|H_{1,k}) = \frac{1}{\pi(\lambda_{N,k}+\lambda_{X,k})} \exp\left\{-\frac{|Y_k|^2}{\lambda_{N,k}+\lambda_{X,k}}\right\} \quad (2)$$

where k indicates the spectral bin index, and $\lambda_{N,k}$ and $\lambda_{X,k}$ denote the variances of the noise and speech spectra, respectively.

The likelihood ratio (LR) of the kth spectral bin, Λ_k, is defined from the above two PDFs as [3]

$$\Lambda_k = \frac{p(Y_k|H_{1,k})}{p(Y_k|H_{0,k})} = \frac{1}{1+\xi_k} \exp\left\{\frac{(1+\gamma_k)\xi_k}{1+\xi_k}\right\} \quad (3)$$

where γ_k and ξ_k are the *a posteriori* and *a priori* signal-to-noise ratios (SNR) defined as, $\gamma_k = |Y_k|^2/\lambda_{N,k} - 1$ [1] and $\xi_k = \lambda_{X,k}/\lambda_{N,k}$. The noise variance is assumed to be known through noise adaptation (see Section 4). However, the variance of the speech is unknown, thus the *a priori* SNR of the nth frame, $\xi_k^{(n)}$, is estimated using the decision directed (DD) method [4] as

$$\hat{\xi}_k^{(n)} = \alpha \frac{\left|\hat{X}_k^{(n-1)}\right|^2}{\lambda_{N,k}^{(n-1)}} + (1-\alpha)MAX\{\gamma_k^{(n)}, 0\} \quad (4)$$

where α is a weighting term, e.g. 0.98, and the enhanced spectral amplitude, $|\hat{X}_k|$, is estimated using the minimum mean square error of the short-time spectral amplitude estimator [4].

The *a posteriori* SNR γ_k fluctuates highly from frame to frame because of the high fluctuation of the short-time spectral amplitude $|Y_k|$. On the other hand, the *a priori* SNR $\hat{\xi}_k$ changes slowly due to the smoothing effect. As the value of α increases, so that of

[1] The definition of the *a posterioi* SNR is slightly different from the original one, $\gamma_k = |Y_k|^2/\lambda_{N,k}$ [4].

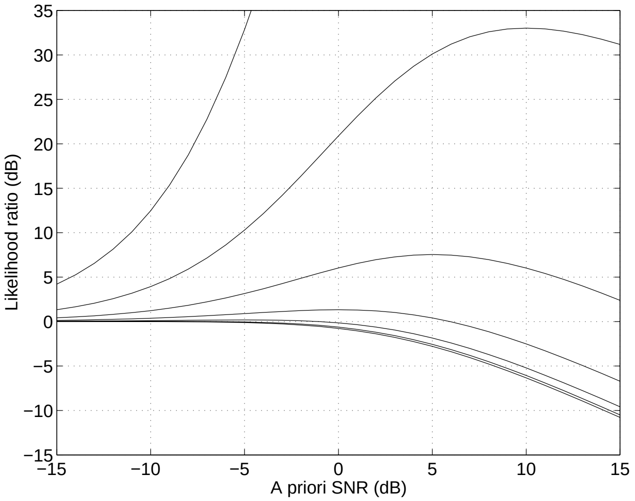

Fig. 1. Likelihood ratio versus *a priori* SNR versus *a posteriori* SNR. The solid lines from top-most to bottom are *a posteriori* SNRs of 15, 10, 5, 0, -5, -10, -15 dB, respectively.

$\hat{\xi}_k$ becomes more smoother. The features of the SNRs, γ_k and $\hat{\xi}_k$, can be compensated each other in the calculation of Λ_k, and consequently enable to enhance the performance of the VAD. The DD estimator for the *a priori* SNR is useful not only for avoiding the musical noise phenomenon in speech enhancement [5], but also for reducing the error rate in voice activity detection.

3. ANALYSIS AND IMPROVEMENT OF THE LIKELIHOOD RATIO

The behavioural characteristics of the LR in (3) are observed with respect to the *a priori* and *a posteriori* SNRs, as shown in Fig. 1. The maximal peaks correspond to the result of maximum likelihood (ML) estimation of the *a priori* SNR. The ML estimator [3] results in lower performance in comparison with the DD estimator because of the inherent high-fluctuation of the *a posteriori* SNR. The LR employing the DD estimator has the following properties:

1. If the *a posteriori* SNR is very high, i.e. $\gamma_k \gg 1$, and the range of the *a priori* SNR is limited properly, the LR becomes very high, i.e. $\Lambda_k \gg 1$.
2. If the *a posteriori* SNR is low, i.e. $\gamma_k < 1$, the *a priori* SNR becomes a key parameter in the calculation of the LR.

In practice, the threshold of the LR is set between 0.2 dB and 0.8 dB, and both the *a posteriori* and *a priori* SNRs are bounded between -15 dB and 15 dB.

The delay of the noise variance $\lambda_{N,k}^{(n-1)}$ in (4) does not seriously affect the *a priori* SNR $\hat{\xi}_k^{(n)}$, assuming that the noise statistics change slowly. However, the spectral amplitude of the speech signal may change abruptly, particularly in onset and offset regions, in which the power of the spectral bins could increase and decrease rapidly, respectively. At the offset region, γ_k can be low but $\hat{\xi}_k$ can be much higher than γ_k due to the delay term $|\hat{X}_k^{(n-1)}|^2$ in (4). Thus Λ_k becomes too low according to the LR property 2, and consequently the geometric mean of Λ_k over all spectral bins may become lower than the threshold of the VAD. On the other hand, the delay rarely causes a problem at the onset regions, according to the LR property 1, as $\gamma_k^{(n)}$ in (3) is large enough.

A few techniques are investigated to overcome the problem in the LR-based VAD. Firstly, it is possible to consider an adaptive weighting factor in the estimation of the *a priori* SNR in (4). In other words, a lower α can be assigned for the active region, and a higher α for the inactive region. When a low α is assigned at the offset region, it reduces the effect of the delay in (4), produces a lower $\hat{\xi}_k$, and therefore may prevent the abrupt decay of Λ_k. However, in our experiment, it was not easy to design a generalised adaptive rule that results in higher performance over various kinds of speech and noise signals. Thus, more investigation is required for attaining a consistent rule concerning the adaptive α. Secondly, a smoothed likelihood ratio (SLR) $\Psi_k^{(n)}$ is considered and defined as

$$\Psi_k^{(n)} = \exp\left\{\kappa \log \Psi_k^{(n-1)} + (1-\kappa) \log \Lambda_k^{(n)}\right\} \qquad (5)$$

where κ is a smoothing factor, and $\Lambda_k^{(n)}$ is defined in (3) for the nth frame. The decision of the voice activity is finally carried out by

$$\Psi^{(n)} = \left\{\prod_{k=1}^{K} \Psi_k^{(n)}\right\}^{\frac{1}{K}} \qquad (6)$$

where K denotes the number of spectral bins. An nth input frame is classified as voice-active if $\Psi^{(n)}$ is greater than a threshold, and voice-inactive otherwise.

An example of the LR and the SLR over a segment of speech signals is shown in Fig. 2(a), (b), and (c). The SLR seems to overcome the problem outlined for the LR. As shown in Fig. 2(b), the SLR is relatively higher than the LR at the offset regions. The comparison over inactive frames is also shown in Fig. 2(c), which indicates that the SLR fluctuates less than the LR.

4. NOISE ESTIMATION BASED ON THE SLR

The short-time spectral amplitudes of the noise signal could fluctuate strongly from frame to frame, depending on the characteristics of the noise source. In order to mitigate this phenomenon, parameter smoothing techniques are considered in the estimation of the variance of noise spectra. Moreover, in order to cope with time-varying noise signals, the variance of the noise spectrum is adapted to the current input signal by a soft decision-based method.

The speech absence probability (SAP) of the kth spectral bin, $p(H_{0,k}|Y_k)$, can be calculated by Bayes' rule as

$$\begin{aligned} p(H_{0,k}|Y_k) &= \frac{p(H_{0,k})p(Y_k|H_{0,k})}{p(H_{0,k})p(Y_k|H_{0,k}) + p(H_{1,k})p(Y_k|H_{1,k})} \\ &= \frac{1}{1 + \frac{p(H_{1,k})}{p(H_{0,k})}\Psi_k} \qquad (7) \end{aligned}$$

where $p(H_{1,k}) = 1 - p(H_{0,k})$, and the unknown *a priori* speech absence probability (PSAP), $p(H_{0,k})$, is estimated in an adaptive manner given as

$$\begin{aligned} \hat{p}(H_{0,k}^{(n)}) &= MIN\{MAX\{\beta\hat{p}(H_{0,k}^{(n-1)}) + (1-\beta) \\ &\quad p(H_{0,k}^{(n)}|Y_k^{(n)}), H_0^{(L)}\}, H_0^{(U)}\} \qquad (8) \end{aligned}$$

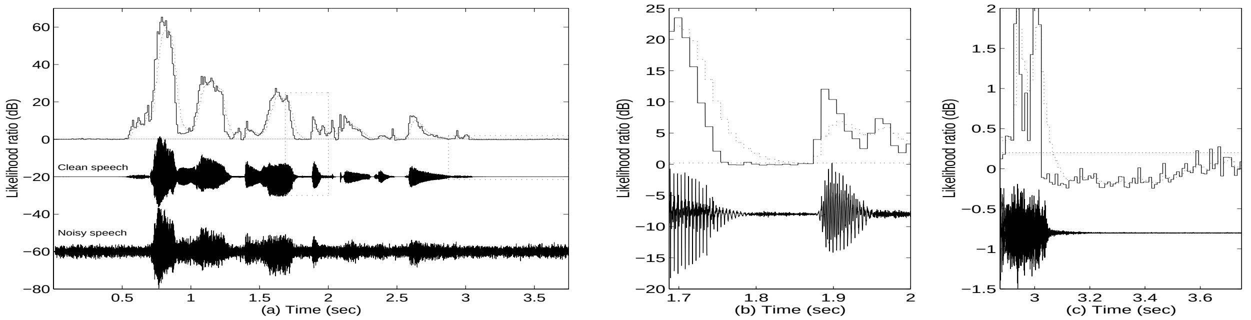

Fig. 2. An example of the computed LR (solid line) and SLR (dotted line) of a segment of vehicle noisy signals of 5 dB SNR. The dotted horizontal-line indicates the VAD threshold. The boxed regions in Fig. (a) are enlarged in Fig. (b) and (c).

where β is a smoothing factor, e.g. 0.65. The lower and upper limits, $H_0^{(L)}$ and $H_0^{(U)}$, of the PSAP are determined through experiments, e.g. 0.2 and 0.8, respectively. Note that the SLR, Ψ_k, instead of the LR, Λ_k, is applied to the calculation of the SAP.

The variance of the noise spectrum of the kth spectral component in the nth frame, $\lambda_{N,k}^{(n)}$, is updated in a recursive way as

$$\lambda_{N,k}^{(n)} = \eta\lambda_{N,k}^{(n-1)} + (1-\eta)E(\left|N_k^{(n)}\right|^2|Y_k^{(n)}) \qquad (9)$$

where η is a smoothing factor, e.g. 0.95. The expected noise power-spectrum $E(|N_k^{(n)}|^2|Y_k^{(n)})$ is estimated by means of a soft-decision technique [6] as

$$\begin{aligned} E(\left|N_k^{(n)}\right|^2|Y_k^{(n)}) = \left|Y_k^{(n)}\right|^2 p(H_{0,k}|Y_k^{(n)}) \\ +\lambda_{N,k}^{(n-1)} p(H_{1,k}|Y_k^{(n)}) \end{aligned} \qquad (10)$$

where $p(H_{1,k}|Y_k^{(n)}) = 1 - p(H_{0,k}|Y_k^{(n)})$. Through the experiments, it is observed that the SLR-based adaptation is useful for the estimation of the noise spectra with high fluctuation, such as babble noise source.

5. PERFORMANCE EVALUATION

An objective test is conducted to evaluate the performance of the proposed VAD scheme. Speech materials of duration 96-sec are collected, filtered by the modified IRS, and then mixed with vehicle and babble noises of 5, 10, 15, and 25 dB SNR. The active and inactive regions of the speech material are marked manually. Furthermore each active region is sub-classified into speech onset, speech offset, and strongly active speech (SAS) in order to obtain more detailed information depending on the characteristics of speech signals. The proportions of the inactive, speech onset, speech offset, and strongly active speech (SAS) regions of the speech material are 0.43, 0.13, 0.17, and 0.27, respectively. The VAD test is carried out every 10-ms frame of the processed noisy signal.

The effect of the smoothing factor κ in (5) is investigated, as shown in Fig. 3. Note that the case of $\kappa = 0$ reduces (5) to the LR-based method. It is obvious from the results that the detection accuracy, with an increase of κ, can be improved considerably at the offset region without serious degradation in the performance at the onset region for both the vehicle and the babble noisy signals. Concerning the detection of the inactive frames, interesting experimental results are observed. In the case of vehicle noisy signals, as κ increases, the false alarm rate in the inactive frames increases: gradually, for $\kappa < 0.9$, and then substantially, for $\kappa > 0.9$. However, in the case of babble noisy signals, it is observed that the error rate decreases gradually as κ increases, for $\kappa < 0.9$, and then increases like the case of the vehicle noisy signal, for $\kappa > 0.9$. Therefore, if κ is selected properly, the SLR-based method gives significantly improved performance over the LR-based method. The reason for this result is explained in Section 3.

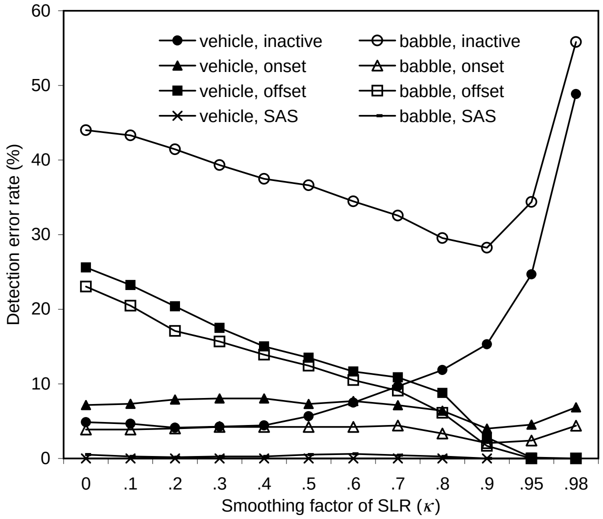

Fig. 3. Analysis of the smoothing factor κ of the SLR in terms of the noise source and the inactive and the active, i.e. onset, offset, and SAS, regions. The noise level is 10 dB SNR.

Under various noise levels and sources, the performance of the SLR-based VAD is compared with those of standard VADs, such as ITU-T G.729 annex B VAD (G.729B) [1] and ETSI AMR VAD option 2 (AMR2) [2], and the LR-based VAD with and without the hangover scheme [3], as shown in Table 5. Originally AMR2 produces the detection result every 20-ms by the *logical OR* operation of two 10-ms detection results, thus the 10-ms result can be obtained easily by slight modification to the original code. Taking

Table 1. The comparison of detection error rates among the SLR-based, the LR-based, AMR2, and G.729B VADs. The LR+HO means the LR-based VAD with the hangover scheme.

		Detection error rate (%)							
SNR		Vehicle noise				Babble noise			
(dB)	VAD	Inactive	Onset	Offset	SAS	Inactive	Onset	Offset	SAS
5	SLR	13.87	6.42	7.51	0.00	29.40	2.43	4.93	0.52
	LR	4.49	12.88	30.92	0.00	46.25	6.90	27.77	2.49
	LR+HO	5.33	12.05	12.86	0.00	46.50	4.52	11.28	1.48
	AMR2	18.64	9.13	0.00	0.00	41.66	4.75	0.26	0.00
	G.729B	8.58	70.23	60.21	5.14	48.17	56.79	45.88	5.12
15	SLR	17.12	3.48	0.73	0.00	29.20	2.05	0.00	0.00
	LR	5.07	5.34	19.85	0.00	41.76	3.70	16.83	0.08
	LR+HO	7.52	4.75	6.80	0.00	42.67	3.32	4.18	0.00
	AMR2	20.15	3.78	0.00	0.00	51.53	2.19	0.26	0.00
	G.729B	8.57	31.19	39.41	0.00	49.79	25.90	32.73	0.00
25	SLR	23.01	2.82	0.00	0.00	30.77	1.54	0.00	0.00
	LR	6.64	3.29	11.79	0.00	34.38	1.54	8.75	0.00
	LR+HO	10.94	1.56	2.75	0.00	36.45	0.89	1.59	0.00
	AMR2	20.28	2.68	0.00	0.00	20.61	2.31	0.12	0.00
	G.729B	8.85	12.75	19.06	0.00	44.30	11.34	15.49	0.00

into account the results in Fig. 3, $\kappa = 0.9$ is selected for the SLR-based VAD. G.729B generates considerably high error rates at the active regions in comparison with other methods. It is important to note that frequent detection errors of speech frames lead to serious degradation in speech quality, thus the error rate of speech frame detection should be low enough. The LR-based VAD gives consistently superior performance to G.729B, but the VAD without the hangover scheme produces relatively high detection error rates in the active regions. The hangover scheme can considerably alleviate this problem, but the speech detection error rate is still somewhat high in comparison with the results of the SLR-based VAD and AMR2. The performances of the SLR-based VAD and AMR2 seem to be comparable.

6. CONCLUSION

In this paper, we have analysed the behaviour of a VAD based on the statistical likelihood ratio (LR), and found that the delay term in the decision-directed parameter estimator employed for the estimation of the LR can cause frequent detection errors in the offset regions of speech frames. In order to circumvent this problem, we have proposed the smoothed likelihood ratio (SLR), which provides a graceful decrease of the likelihood ratio at the offset region. Moreover, the SLR-based parameter smoothing technique is applied for adaptation of the noise variance so as to cope with high fluctuation of the noise spectra. Through the experiments, it has been shown that the proposed SLR scheme is highly desirable for the improvement of the LR-based VAD. Additionally, the SLR-based VAD gives detection performances superior to G.729B and comparable with AMR VAD option 2.

7. ACKNOWLEDGEMENT

The authors would like to thank Dr. D. Burgess for his precious proof reading and S. Villette for his invaluable suggestions for performance evaluation

8. REFERENCES

[1] ITU-T, "A silence compression scheme for G.729 optimised for terminals conforming to ITU-T V.70," *ITU-T Rec. G.729 Annex B*, Nov. 1996.

[2] ETSI, "Voice activity detector (VAD) for adaptive multi-rate (AMR) speech teaffic channels," *ETSI EN 301 708 v7.1.1*, Dec. 1999.

[3] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Sig. Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–20, Dec. 1984.

[5] O. Cappé, "Elimination of musical noise phonomenon with the Ephraim and Malah noise suppression," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 345–9, Apr. 1994.

[6] J. Sohn and W. Sung, "A voice activity detection employing soft decision based noise spectrum adaptation," in *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, May 1998, pp. 365–8.