A VOICE ACTIVITY DETECTOR EMPLOYING SOFT DECISION BASED NOISE SPECTRUM ADAPTATION

Jongseo Sohn and Wonyong Sung

School of Electrical Engineering,
Seoul National University, Seoul, Korea.
e-mail: sohn, wysung@dsp.snu.ac.kr.

ABSTRACT

In this paper, a voice activity detector (VAD) for variable rate speech coding is decomposed into two parts, a decision rule and a background noise statistic estimator, which are analysed separately by applying a statistical model. A robust decision rule is derived from the generalized likelihood ratio test by assuming that the noise statistics are known a priori. To estimate the time-varying noise statistics, allowing for the occasional presence of the speech signal, a novel noise spectrum adaptation algorithm using the soft decision information of the proposed decision rule is developed. The algorithm is robust, especially for the time-varying noise such as babble noise.

1. INTRODUCTION

Many speech coding applications such as digital voice storage, code division multiple access (CDMA) wireless networks, and packetized communication systems allow variable rate transmission. To reduce the average bit rate, the speech coders in such systems usually employ a voice activity detector (VAD), so that less or no bits can be assigned in the absence of speech.

The most widely used feature for voice activity detection is the difference between speech and background noise in temporal variations of statistics [1]. Typical statistics are the second order moments such as energy, subband energies and power spectrum. The temporal variations of background noise statistics are assumed to be much smaller than those of speech, which makes it possible to estimate the time-varying noise statistics in spite of the occasional presence of a speech signal. The noise statistics are updated during the absence of speech [2], or continuously adapted while imposing some constraints on the adaptation [3][4]. Then, a very sensitive VAD can be designed by formulating a decision rule that compares the estimated noise statistics and the observed signal statistics. Figure 1 shows a paradigm of such VADs. Well known examples of VADs include those employed by 8 kbps and 13 kbps QCELP speech coders in the IS-95 standard, the EVRC in the IS-127 standard, and the VAD adopted for the discontinuous transmission (DTX) mode of the GSM standard.

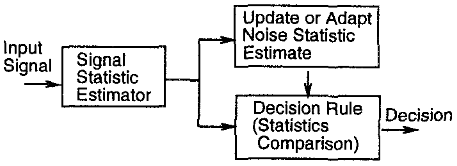

Figure 1: Block diagram of a VAD employing a noise statistic estimator.

Because the performance of a VAD depends on both the decision rule and the noise statistic estimation algorithm, which are coupled together, we optimized each of them separately by applying a statistical model.

2. DECISION RULES OF THE VOICE ACTIVITY DETECTOR

The decision rule of a VAD can be formulated by a *decision statistic*, which is a quantity that measures the difference between noise and observed signal statistics, and a *decision threshold*, which is often determined empirically. In this section, we derive a decision statistic from the generalized likelihood ratio test (LRT) by assuming that the noise statistics are given *a priori* by the noise statistic estimator discussed in the following section. We do not consider any hangover or threshold adaptation scheme during the comparison, as these can be added in a heuristic way after the design of the decision rule.

As most current low bit-rate speech coders operate on a frame basis, voice activity detection is also performed for each frame of L samples over which speech is assumed to be stationary. We use the statistical model in which the speech and noise signals are Gaussian random processes that are independent of each other, then the discrete Fourier transform (DFT) coefficients of each process are asymptotically independent Gaussian random variables [5]. The L-dimensional coefficient vectors of speech, noise, and noisy speech are denoted as $\mathbf{S}$, $\mathbf{N}$, and $\mathbf{X}$, with their kth elements S_k, N_k, and X_k, respectively. In this statistical model, the variances of N_k and S_k are given by [5]:

$$\lambda_N(k) = S_N(2\pi k/L) \qquad (1)$$

$$\lambda_S(k) = S_S(2\pi k/L) \qquad (2)$$

where $S_N(\omega)$ and $S_S(\omega)$ denote the true power spectra of noise and speech, respectively. The variance of X_k is given by:

$$\sigma_X^2(k) = \lambda_N(k) + \lambda_S(k) \quad \checkmark \qquad (3)$$

As mentioned before, the noise statistics $\lambda_N(k)$s are assumed to be known *a priori*. Then, the two hypotheses of the voice activity detection problem are as follows:

$$H_0 : \text{speech absent } : \mathbf{X} = \mathbf{N}$$

$$H_1 : \text{speech present} : \mathbf{X} = \mathbf{N} + \mathbf{S}$$

where H_1 is a composite hypothesis with a set of L unknown parameters, $\Theta = \{\lambda_S(k) : k = 0, \ldots L-1\}$. The joint probability density functions conditioned on H_0, and on H_1 and Θ are given by:

$$p(\mathbf{X}|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi\lambda_N(k)} \exp\{-\frac{|X_k|^2}{\lambda_N(k)}\} \qquad (4)$$

$$p(\mathbf{X}|\Theta, H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi[\lambda_N(k) + \lambda_S(k)]} \exp\{-\frac{|X_k|^2}{\lambda_N(k) + \lambda_S(k)}\} \qquad (5)$$

The generalized LRT, which is one of the most powerful methods for composite hypothesis testing [6], replaces Θ with its maximum likelihood estimate, $\hat{\Theta} = \{\hat{\lambda}_S(k) : k = 0, \ldots, L-1\}$, where $\hat{\lambda}_S(k)$s are obtained by the power subtraction method, i.e.,

$$\hat{\lambda}_S(k) = |X_k|^2 - \lambda_N(k) \qquad (6)$$

and the corresponding decision rule using the log likelihood ratio is obtained by substituting Eq. (6) into Eq. (5) as follows:

$$\Lambda_g = \frac{1}{L} \log \frac{p(\mathbf{X}|\hat{\Theta}, H_1)}{p(\mathbf{X}|H_0)}$$

$$= \frac{1}{L} \sum_{k=0}^{L-1} \{\frac{|X_k|^2}{\lambda_N(k)} - \log \frac{|X_k|^2}{\lambda_N(k)} - 1\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \qquad (7)$$

which can be recognized as a discrete approximation of the Itakura-Saito distortion (ISD) measure or as the cross entropy between background noise and observed signal [7].

The VAD used for the DTX mode of GSM estimates the spectral shape of noise in a *P*th order linear predictive coding (LPC) filter form and the corresponding residual energy [2]. The inverse of the LPC filter is used as a noise suppression filter, and the residual energy of the output of the inverse filter is compared with the estimate of noise residual energy to decide whether a signal is speech or noise. If we ignore the memory of the noise suppression filter, the decision rule of the VAD can be formulated as follows:

$$\frac{1}{L} \sum_{k=0}^{L-1} |A_N(k)|^2 |X_k|^2 \underset{H_0}{\overset{H_1}{\gtrless}} (1+\alpha)\sigma_N^2 \qquad (8)$$

where $A_N(k)$ is the kth DFT coefficient of the impulse response of the noise suppression filter, which has length $P+1$, σ_N^2 is the estimated noise residual energy and $L > M + P - 1$ for frame length M. If we recognize that $\sigma_n^2/|A_N(k)|^2$ is a kind of estimator of $\lambda_N(k)$, a generalized form of Eq. (8) is given by:

$$\frac{1}{L} \sum_{k=0}^{L-1} \frac{|X_k|^2}{\lambda_N(k)} \underset{H_0}{\overset{H_1}{\gtrless}} 1 + \alpha \qquad (9)$$

The decision statistic of Eq. (9) can be identified as an average of L subband signal to noise ratios (SNRs). This is almost the same as the statistic proposed by Yang [8]. He used the average subband *a priori* SNR obtained by the power subtraction method.

The VADs of QCELP and EVRC are based on estimates of full band and two subband energies of background noise, respectively [3][4]. Their decision rules can be also interpreted as averaged SNRs over one or two subbands. It can be shown that increasing the number of subbands to L yields the decision rule of Eq. (9), which shows the best performance.

Comparing Eq. (7) and (9) reveals that the average subband SNR-based decision statistic of Eq. (9) is the same as the ISD except for the term $(1/L)\sum \log\{|X_k|^2/\lambda_N(k)\}$. This term becomes dominant only when $|X_k|^2/\lambda_N(k)$ is much less than unity, which is not true in either the H_1 case (because speech is added to the noise), or the H_0 case. Figure 2 compares these two statistics when the VAD operates on noisy speech in Fig. 2(c). $\lambda_N(k)$'s are estimated using a periodogram with analysis windowing. Although the ISD shows more consistent values than the average subband SNR during the absence of speech, these two statistics show similar discrimination performance. Therefore, the decision statistic based on the average subband SNR is a good approximation of the ISD in the voice activity detection problem.

3. ESTIMATION OF NOISE STATISTICS

To estimate or track slow time-varying statistics of nonstationary signals, time averaging with exponential or rectangular weighting is commonly used. In the voice activity detection problem, however, some constraints should be imposed on the tracking to prevent the estimate from being affected by the occasional speech signal. The VADs of

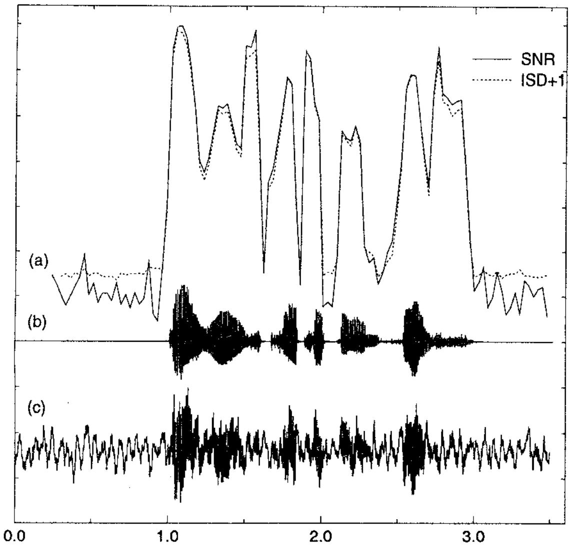

Figure 2: Comparison of the average subband SNR and ISD statistics. The statistics are plotted in log scale. (a) Average subband SNR vs. ISD plus one. (b) Original clean speech, which is shown for easy discrimination by sight. (c) Noisy speech corrupted by vehicular noise at maximum SNR 10 dB. The vehicular noise is obtained from the NOISEX-92 set, NOISE-ROM-0 signal.023.

QCELP and EVRC continuously adapt the noise energy estimate but restrict the rate of increase of the estimate, since a large signal energy indicates the presence of speech. However, as the number of parameters to be estimated is increased by more subband splitting, as in Eq. (9), this heuristic method fails to adapt the spectral shape adequately.

The VAD used in the GSM system updates the noise statistics during the noise-only periods. A secondary VAD is employed to determine the noise-only periods. It is simple but should be designed to be more conservative than the primary decision rule in assuming the presence of speech [2]. It performs well when the noise spectrum changes very slowly such as in vehicular noise. However, as the background noise becomes less stationary, the secondary VAD misses many noise frames and the noise spectrum estimate rapidly becomes outdated. As a result, the decision statistic of the primary decision rule is increased during the absence of speech, and the VAD often false-alarms the speech presence [1].

3.1. Noise Spectrum Adaptation Using Soft Decision Information

The optimal estimate of the variance of the background noise Fourier expansion coefficients $\lambda_N(k)$ in terms of the minimum mean square error is given by:

$$\begin{aligned}\hat{\lambda}_N(k) &= E(\lambda_N(k)|X_k)\\ &= E(\lambda_N(k)|H_0)P(H_0|X_k) + E(\lambda_N(k)|H_1)P(H_1|X_k)\end{aligned} \tag{10}$$

Using Bayes rule:

$$\begin{aligned}P(H_0|X_k) &= \frac{p(X_k|H_0)P(H_0)}{p(X_k|H_0)P(H_0)+p(X_k|H_1)P(H_1)}\\ &= \frac{1}{1+\varepsilon\Lambda(k)}\end{aligned} \tag{11}$$

where $\boxed{\varepsilon = \frac{P(H_1)}{P(H_0)}}$ and $\boxed{\Lambda(k) = p(X_k|H_1)/p(X_k|H_0)}$. Similarly, the following holds:

$$P(H_1|X_k) = \frac{\varepsilon\Lambda(k)}{1+\varepsilon\Lambda(k)} \tag{12}$$

Substituting Eq. (11) and Eq. (12) into Eq. (10) yields:

$$\begin{aligned}E(\lambda_N(k)|X_k) &= \frac{1}{1+\varepsilon\Lambda(k)}\boxed{E(\lambda_N(k)|H_0)}\\ &+ \frac{\varepsilon\Lambda(k)}{1+\varepsilon\Lambda(k)}\boxed{E(\lambda_N(k)|H_1)}\end{aligned} \tag{13}$$

Since the estimation is performed for each frame, we add the superscript (m) so that $\lambda_N^{(m)}(k)$ denotes $\lambda_N(k)$ at the mth frame. To obtain a feasible estimator of $\lambda_N(k)$ rather than Eq. (13), we use the current frame measurement $|X_k^{(m)}|$ instead of $E(\lambda_N^{(m)}(k)|H_0)$ when speech is absent. When speech is present, for the observed information $|X_k^{(m)}|$ not to be reflected on $\hat{\lambda}_N^{(m)}(k)$, we replace $E(\lambda_N^{(m)}(k)|H_1)$ by the estimate of the previous frame, $\hat{\lambda}_N^{(m-1)}(k)$, and a recursive formula for $\hat{\lambda}_N(k)$ is obtained as follows:

✓

$$\boxed{\begin{aligned}\hat{\lambda}_N^{(m)}(k) &= \frac{1}{1+\varepsilon\Lambda^{(m)}(k)}|X_k^{(m)}|^2\\ &+ \frac{\varepsilon\Lambda^{(m)}(k)}{1+\varepsilon\Lambda^{(m)}(k)}\hat{\lambda}_N^{(m-1)}(k)\end{aligned}} \tag{14}$$

As we have no estimate of the speech parameter set Θ, it seems reasonable to use the generalized likelihood ratio instead of $\Lambda^{(m)}(k)$ in Eq. (14), which is defined as:

$$\Lambda_g^{(m)}(k) = \frac{p(X_k^{(m)}|\hat{\lambda}_S^{(m)}(k), H_1)}{p(X_k^{(m)}|H_0)} \tag{15}$$

Since the decision is not made for each frequency band k, but made once by observing all the frequency bands, we replace the $\Lambda_g(k)$s with their geometric mean Λ_g in Eq. (7), as follows:

$$\hat{\lambda}_N^{(m)}(k) = \frac{1}{1+\varepsilon\Lambda_g^{(m)}}|X_k^{(m)}|^2 + \frac{\varepsilon\Lambda_g^{(m)}}{1+\varepsilon\Lambda_g^{(m)}}\hat{\lambda}_N^{(m-1)}(k) \tag{16}$$

If $\Lambda^{(m)}$ were fixed for frame index m, Eq. (16) would be an estimator of the time-varying spectrum with an exponentially weighted averaging. However in this case it can be

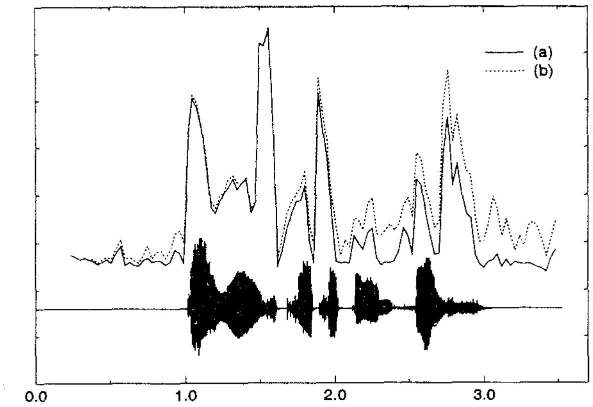

Figure 3: Performance comparison of the two noise spectrum estimation methods. ISDs are plotted in log scale. Only the original speech signal is shown. (a) Soft decision-based adaptation method. (b) Secondary VAD-based update method.

identified as a first order time-varying infinite impulse response (IIR) filtering. Equation (16) implies that the closer the observed signal spectrum is to the current estimate, the faster is the convergence speed of the adaptation system. Also, ε can be interpreted as a system parameter that determines the convergence speed rather than the *a priori* probability ratio. Using the soft decision information Λ_g, the noise spectrum is continuously adapted whether speech is present or not present.

3.2. Experimental Results

Babble noise, which is commonly encountered in the mobile communication environment, is known as a time-varying signal [1] and is less stationary than vehicular noise. Babble noise is added to a clean speech signal at maximum SNR 10 dB. ISDs are measured for the noisy speech to compare the two noise spectrum estimation methods, the secondary VAD-based update and the soft decision-based adaptation method. Results are shown in Fig. 3. Because the secondary VAD requires long periods of noise to determine the noise-only periods, it cannot update the noise spectrum during short silence intervals such as between speech utterances or before plosives. Therefore, the ISDs are found to be large after a burst of speech utterances, while the soft decision-based adaptation method tracks the noise spectrum properly.

4. DISCUSSION AND CONCLUDING REMARKS

To analyse a VAD, we decomposed it into two parts: the decision rule and the noise statistic estimation algorithm. These are optimized separately by applying a statistical model. We derived a robust decision rule from the generalized LRT by assuming that the noise statistics are known *a priori*. For the noise statistic estimation part, a robust noise spectrum adaptation method is developed by using the soft decision information of the proposed decision rule. This shows better tracking performance than existing methods, especially for time-varying noise such as babble noise.

When our VAD is used in an LPC-based speech coder, the noise spectrum can be estimated using an LPC model to eliminate the DFT operations that require a large amount of computation. The ISD between the noise LPC model and the observed signal spectrum is approximated by the modified ISD. Procedures for the efficient computation of the modified ISD are discussed in [9]. With this approximation, the adaptation of the noise spectrum can be performed in the autocorrelation domain, since the relationship between the current and previous noise estimates in Eq. (16) is linear.

5. REFERENCES

[1] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Proc. IEEE Speech Coding Workshop*, Oct. 1993, pp. 85–86.

[2] D.K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd, "The voice activity detector for the pan-European digital cellular mobile telephone service," in *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, 1989, pp. 369–372.

[3] W. Gardner, P. Jacobs, and C. Lee, "QCELP: A variable rate speech coder for CDMA digital cellular," in *Speech and Audio Coding for Wireless Networks*, pp. 85–92, MA: Kluwer, 1993.

[4] TIA/EIA/IS-127, *Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*, 1996.

[5] W.A. Pearlman and R.M. Gray, "Source coding of the discrete Fourier transform," *IEEE Trans. Inform., Theory*, vol. IT-23, pp. 683–692, Nov. 1978.

[6] H.L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*, MIT Press, Cambridge, MA, 1968.

[7] J.E. Shore, "Minimum cross-entropy spectral analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, no. 2, pp. 230–237, April 1981.

[8] J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems," in *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, 1993, pp. 363–366.

[9] R.M. Gray, A.H. Gray, Jr., G. Rebolledo, and J.E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Trans. Inform., Theory*, vol. IT-27, no. 6, pp. 708–721, Nov. 1981.