Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold

Article *in* IEEE Transactions on Audio Speech and Language Processing · April 2006

DOI: 10.1109/TSA.2005.855842 · Source: IEEE Xplore

Sven Nordholm

Curtin University

344 PUBLICATIONS **2,702** CITATIONS

SEE PROFILE

Roberto Togneri

University of Western Australia

213 PUBLICATIONS **2,734** CITATIONS

SEE PROFILE

Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold

Alan Davis, *Student Member, IEEE*, Sven Nordholm, *Senior Member, IEEE*, and Roberto Togneri, *Senior Member, IEEE*

***Abstract*—Traditionally, voice activity detection algorithms are based on any combination of general speech properties such as temporal energy variations, periodicity, and spectrum. This paper describes a novel statistical method for voice activity detection using a signal-to-noise ratio measure. The method employs a low-variance spectrum estimate and determines an optimal threshold based on the estimated noise statistics. A possible implementation is presented and evaluated over a large test set and compared to current modern standardized algorithms. The evaluations indicate promising results with the proposed scheme being comparable or favorable over the whole test set.**

***Index Terms*—Adaptive voice activity detection, statistical decision, voice activity detection (VAD), voice activity detector.**

I. Introduction

Voice activity detection (VAD) is becoming increasingly important and relevant in modern telecommunication and speech enhancement systems. This increase can be largely attributed to the desire to lower the average bit-rate of speech communication systems, whether this be for mobile telephony or VoIP communications [1].

The VAD mechanism simply decides if speech is present. This information is then used to selectively encode and transmit data in mobile telephony applications or estimate noise statistics in speech enhancement applications. The result of this selectivity is not only the aforementioned data savings, but also power savings in mobile devices [2], co-channel interference reduction in mobile telephony [3] and greater noise suppression in speech enhancement.

Traditionally, VAD algorithms are based on heuristics or fuzzy rules, and general speech properties, for example see [4]–[6]. This design methodology makes it difficult to optimize relevant parameters and obtain consistent results. Recently, attempts have been made to develop a statistical model-based VAD [7], [8]. These schemes adopt the model proposed by Ephraim and Malah. The model assumes Fourier coefficients are statistically independent Gaussian random variables [9] and is motivated by the central limit theorem. Using this model a likelihood ratio is developed and a statistical hypothesis test conducted.

The formulation of the hypothesis test presents some problem. It indicates two key parameters need to be determined, namely the *a priori* and *a posteriori* signal-to-noise ratios [9]. The problem of determining the *a priori* signal-to-noise ratio is addressed by estimating MMSE speech spectral amplitudes. This estimation however is undesirable, introducing complexity and a computational burden. The *a posteriori* signal-to-noise ratio is estimated using a scaled periodogram. Both ratios further depend on an estimate of the variance of the Fourier coefficients during periods of noise. This variance is either determined *a priori* or estimated using an exponential average of a scaled periodogram [10].

Further, the issue of determining the threshold for the hypothesis test is ignored. Bayesian hypothesis testing indicates a threshold should be determined on the basis of a cost or risk function [11]. This however requires *a priori* knowledge of the probabilities of occurrence of each hypothesis, which in this case makes determining a threshold in this manner impractical. Cho *et al.* addressed this and indicated a region for the threshold, but gave no specific analysis [8]. In general the threshold is set by some heuristic rule.

The schemes were reported to produce good results in both babble and vehicle noise. Sohn *et al.* also evaluated the scheme in Gaussian noise; however, results indicated a declining performance below 15 dB [7]. This is due, at least in part, to the method of estimation of the key parameters outlined earlier, namely a scaled periodogram. The periodogram is well known to be an inconsistent spectral estimator [12]. Typically it is shown that the variance is approximately the same size as the square of the power spectrum that is being estimated, and does not decrease with increasing data length. This high variance contributes to the reduced performance in white noise.

Another statistical scheme has been developed by McKinley and Whipple [13]. This scheme, in contrast to other statistical methods compares second-order statistics of the signal to models. Speech models are estimated from a large speech set developed off-line, and noise models are estimated during an initial silence period. The scheme was reported to produce good results in a range of environments; however, only a small test set was used. Further, the scheme is computationally expensive and complex.

Manuscript received October 20, 2003; revised November 22, 2004. This work was supported by the Australian Research Council under Grant A00105530 and the Australian Telecommunications Cooperative Research Centre. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bayya Yegnanarayana.

A. Davis and S. Nordholm are with the Western Australian Telecommunications Research Institute, a Joint Venture Between The University of Western Australia and Curtin University of Technology, Crawley, WA 6009, Australia (e-mail: davisa@watri.org.au; sven@watri.org.au).

R. Togneri is with the School of Electrical, Electronic and Computer Engineering, The University of Western Australia, Crawley, WA 6009, Australia (e-mail: roberto@ee.uwa.edu.au).

Digital Object Identifier 10.1109/TSA.2005.855842

This paper proposes a statistical VAD scheme that makes no assumption about the distribution of the speech in contrast to existing statistical schemes, and instead in some sense attempts to optimally detect the presence of noise. The scheme incorporates a low-variance spectrum estimate and a statistical detection mechanism [14]. The scheme removes the need to estimate the undesirable *a priori* signal-to-noise ratio. Instead the scheme depends on the expected noise power spectral density and the variance of a "signal-to-noise ratio measure" estimated during periods of nonspeech activity. In this way the scheme depends on how the noise varies from frame to frame. These two aforementioned parameters are simple to estimate during an initial silence period [13].

Further, the proposed VAD addresses the issue of threshold determination. An expression is developed for a threshold based on the estimated noise statistics and the desired performance of the VAD. In this manner the proposed VAD adapts to the current noise environment and its performance can be easily altered using only a single meaningful parameter. The combination of these methods results in a statistical test that is computationally efficient and elegant in its implementation.

Finally, this paper presents a possible implementation. The proposed implementation is evaluated and compared to modern standardized algorithms, namely the ETSI AMR VAD options 1 and 2 [4] and the ITU G729 Annex B VAD [5]. The evaluation is an extension of work undertaken by Sarikaya and Hansen [15], whereby the core TIMIT TEST set was used to evaluate VAD performance in a variety of environments. The evaluation indicates that the proposed scheme yields good results through a range of different noise environments and range of signal-to-noise ratios. One of the most interesting aspects of the results is the consistent nature of the proposed algorithm, indicating a good correlation between theory and the observed results.

II. Signal-to-Noise Ratio Measure

Consider the case in which a received speech signal is corrupted by stationary additive noise. The framed received signal may thus be modeled in the following way:

$$x_k(n) = s_k(n) + v_k(n) \tag{1}$$

where $s_k(n)$ and $v_k(n)$ are the clean speech and additive noise of the k^{th} frame respectively. It is assumed that the speech and noise are independent. It is further assumed that the noise environment is long-term stationary and the speech is short-term stationary.

In order to analyze the received signal, spectrum estimation techniques are commonly employed. Typically a periodogram is used; however, it is well known that the periodogram is an inconsistent spectral estimator. Therefore, low-variance spectrum estimation techniques should be used to accurately evaluate the spectral content of the received signal. Further, we wish to minimize the variance inherent in the spectrum estimation technique, because the proposed scheme is dependant on the estimated variance of the background noise. Therefore, it is undesirable to use a "high" variance technique, since it would influence the estimated background noise variance.

Techniques such as the Welch and Bartlett methods were investigated for this purpose. The Welch method of overlapping windows was found to give a good tradeoff between variance reduction and spectral resolution reduction. To generate a reduced variance, reduced resolution power spectral density (PSD) estimate $P_{xx,k}(f_l)$, where $f_l \in \{0, 1, \ldots, L-1\}$ is related to normalized frequency by $\Omega_{f_l} = 2\pi f_l/L$, M overlapping subframes with a length of L each are used. The subframes are overlapped 50% and windowed with a Hanning window. This PSD estimation technique is similar to that used in the ETSI AMR VAD option 2 [4]. The ETSI scheme averages over adjacent FFT bin magnitudes, whereas here we average over adjacent subframes.

The aforementioned low-variance spectrum estimation techniques generally do not produce coefficients that follow a zero mean Gaussian distribution. However, in order to get a tractable detection problem, a zero mean Gaussian distribution is preferable. We define the signal-to-noise ratio (SNR) measure as follows:

$$\psi_k(f_l) = \frac{P_{xx,k}(f_l)}{\hat{P}_{vv}(f_l)} - 1 \tag{2}$$

where $\hat{P}_{vv}(f_l)$ is the expected value of the noise PSD and $P_{xx,k}(f_l)$ is the PSD of the current frame k. The measure represents the ratio of an instantaneous PSD estimate to a long average of the noise PSD. The SNR measure is closely related to the average SNR of the signal, and in fact during speech activity the expected value of the measure is the true average SNR of the signal for a particular spectral bin f_l.

The expected value of the noise PSD is calculated as the sample mean over an initial period of nonspeech activity. This is found as

$$\hat{P}_{vv}(f_l) = \frac{1}{K}\sum_{k=0}^{K-1} P_{xx,k}(f_l) \tag{3}$$

where K is the number of frames during the initial period.

A. Expected Value of SNR Measure

It is important to understand the behavior of the SNR measure. To that end the expected value of the measure during periods of nonspeech activity should be evaluated. During these periods only noise is present, thus the SNR measure becomes

$$\psi_k(f_l) = \frac{P_{vv,k}(f_l)}{\hat{P}_{vv}(f_l)} - 1 \tag{4}$$

where $P_{vv,k}(f_l)$ is the PSD estimate of the noise. Taking the expected value of (4)

$$\begin{aligned} E\left[\psi_k(f_l)\right] &= E\left[\frac{P_{vv,k}(f_l)}{\hat{P}_{vv}(f_l)} - 1\right] \\ &= \frac{1}{\hat{P}_{vv}(f_l)} E\left[P_{vv,k}(f_l)\right] - 1 \\ &= 0. \end{aligned} \tag{5}$$

Therefore, the expected value of the SNR measure during periods of nonspeech activity can be assumed to be close to zero.

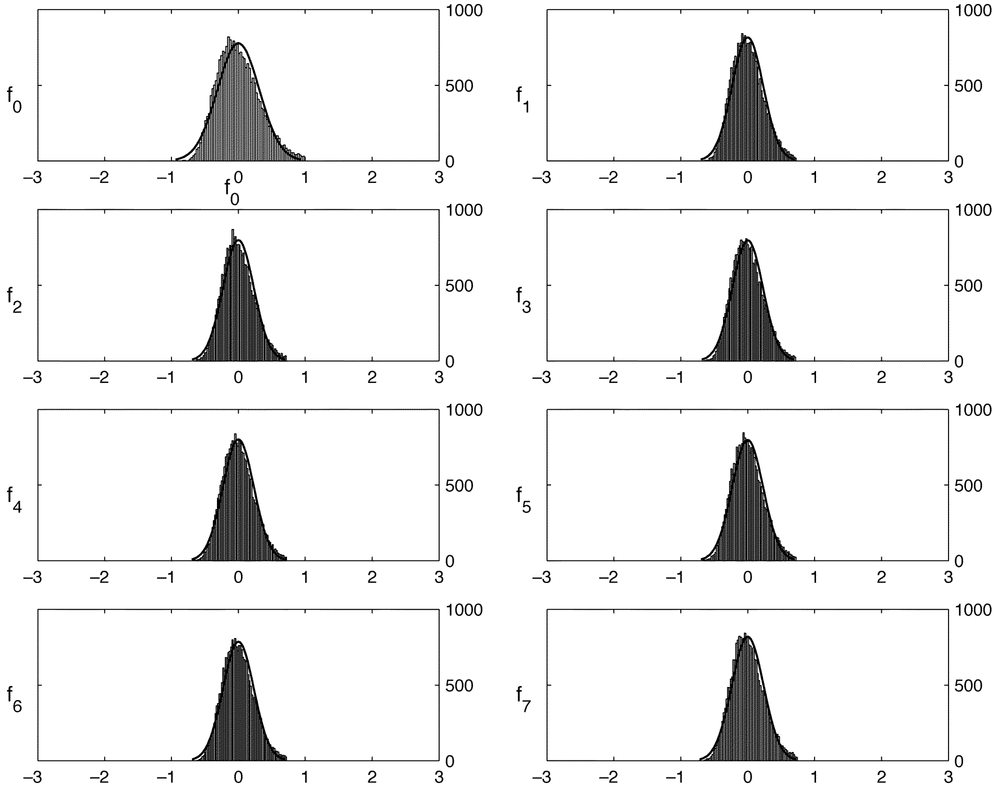


Fig. 1. Histograms of SNR measure in Gaussian noise environment for each frequency bin f_n.

B. Variance of SNR Measure

Similarly, the variance of the measure can be determined, again under the condition that speech is absent. The variance of the SNR measure becomes

$$\begin{aligned}\sigma_v^2(f_l) &= E\left[\left(\psi_k(f_l) - E\left[\psi_k(f_l)\right]\right)^2\right] \\ &= E\left[\psi_k^2(f_l)\right]. \end{aligned} \tag{6}$$

Therefore the variance of the SNR measure during periods of nonspeech activity may be estimated by finding the average square of the SNR, during these periods.

C. Distribution of SNR measure

Finally the statistical nature of the measure should be evaluated. The distribution of the periodogram $P_{xx,k}(f)$ has been investigated on many occasions and has been shown to be Chi-square (χ^2) distributed with two degrees of freedom [16]. Therefore, the low-variance PSD estimate $P_{xx,k}(f_l)$ is made up of a sum of M χ^2 random distributions with two degrees of freedom each where M is the number of subframes, resulting in a χ^2 distribution with $2M$ degrees of freedom. The SNR is made by scaling $P_{xx,k}(f_l)$ by the constant $\hat{P}_{vv}(f_l)$ and shifting the mean by -1. The SNR $\psi_k(f_l)$ may thus be considered to be χ^2 distributed with $2M$ degrees of freedom; however, in practice the distribution may have fewer than $2M$ degrees of freedom due to the overlapping of the subframes.

Using the Welch method with overlapping windows, M becomes large, and thus the number of degrees of freedom becomes large. The distribution of $\psi_k(f_l)$ thus tends toward Gaussian [11]. As such the SNR estimate may be considered to follow a Gaussian distribution.

In order to validate this, the SNR measure was experimentally calculated in a range of noise environments. The measure was found to follow closely a Gaussian distribution in stationary noise environments such as Gaussian noise, pink noise, and HF channel noise as taken from the NOISEX-92 database. In highly variable environments such as babble and vehicle noise, the assumption is violated. Normalized histograms along with a Gaussian fit with zero mean can be seen in Figs. 1–3. The figures represent the Gaussian, vehicle, and babble noise environments as taken from the NOISEX-92 database. The effect of the Gaussian assumption failing is reduced performance in the affected environment. This is generally manifested as false alarms, due to the long tails on the probability density function (pdf). This phenomenon can be seen in the evaluation where the babble noise environment is considered.

III. Statistical Detection Using the SNR Measure

In general, signal detection may be viewed as the problem of deciding between two possibilities. When applied here, this problem becomes one of deciding if speech is present or not. In order to make this decision, two hypotheses are considered, the null and alternative. The null represents the case where only noise is present and the alternative represents the case where

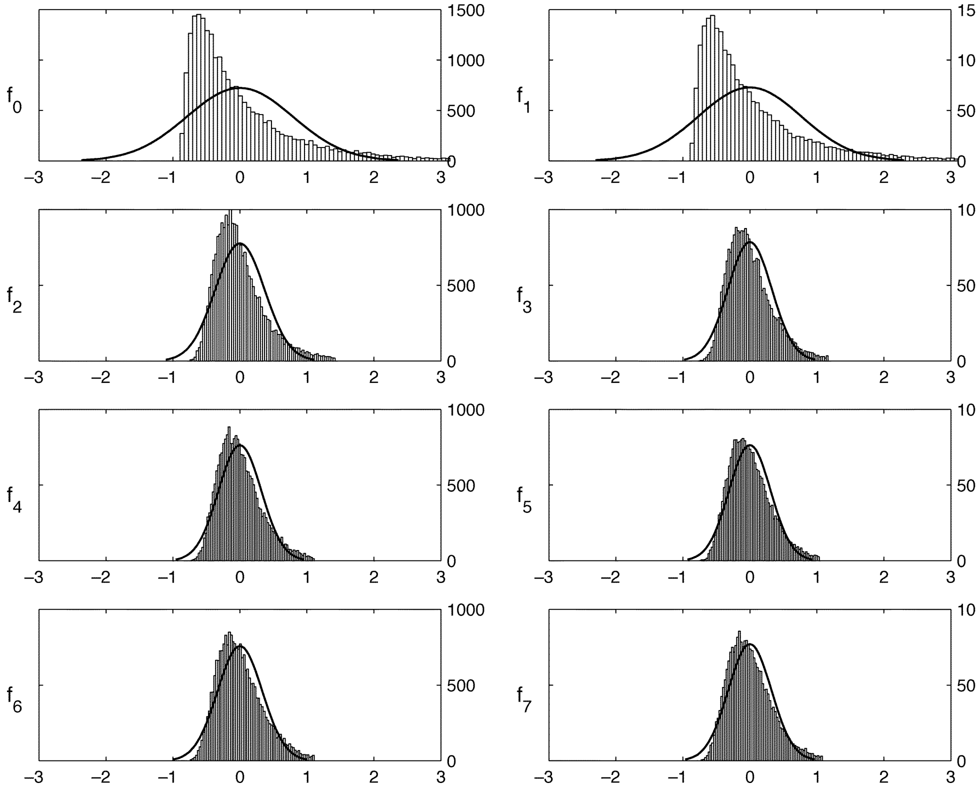


Fig. 2. Histograms of SNR measure in Vehicle noise environment for each frequency bin f_n.

both speech and noise are present. These can be represented in the following manner:

$$H_0 : \psi_k(f_l) = \frac{P_{vv,k}(f_l)}{\hat{P}_{vv}(f_l)} - 1$$
$$H_1 : \psi_k(f_l) = \frac{P_{vv,k}(f_l) + P_{ss,k}(f_l)}{\hat{P}_{vv}(f_l)} - 1$$

where H_0 represents the null hypothesis, H_1 represents the alternative and $P_{ss,k}(f_l)$ is a PSD estimate of the speech in the f_l^{th} spectral bin. This representation comes about as per the independence assumption in (1).

In order to decide between the two possibilities, traditionally a likelihood ratio is found. This is realized by finding the ratio of the assumed pdfs, under the null and alternative. The likelihood ratio is then compared to a threshold. This threshold ideally should be found by considering a cost function and *a priori* probabilities of occurrence of each hypothesis. This threshold however is often heuristically found due to problems inherent in estimating the cost function and *a priori* probabilities. A decision is finally made by determining if the likelihood ratio is larger than the threshold or not, and thus if the alternative or null hypothesis is present.

Here we wish to make a statistical detection without assuming any prior knowledge of the distribution of the speech signal, *a priori* probabilities of occurrence of each hypothesis or the cost function. We will only assume that the SNR measure, during periods of nonspeech activity, is zero mean and distributed in a Gaussian manner, and that under the alternative, the introduction of speech introduces some significant shift in mean. For high-SNR conditions, this assumption will be valid; however, as the SNR decreases the assumption will begin to become violated during low-energy portions of speech. We will thus model the pdf of the SNR measure during periods of nonspeech activity as

$$p\left(\psi_k(f_l)|H_0\right) = \frac{1}{\sqrt{2\pi\sigma_{v,k}^2(f_l)}}\exp\left(\frac{-\psi_k^2(f_l)}{2\sigma_{v,k}^2(f_l)}\right) \qquad (7)$$

where $\sigma_{v,k}^2(f_{l)}$ is the variance of the SNR measure during periods of nonspeech activity in the f_l^{th} spectral bin.

Rather than comparing a test statistic like the likelihood ratio to a threshold, we will instead directly compare the SNR measure to a threshold. We will represent this comparison as

$$\psi_k(f_l) \gtrless_{H_0}^{H_1} \eta_k'(f_l) \qquad (8)$$

where $\eta_k'(f_l)$ is the threshold in the f_l^{th} spectral bin.

The problem at hand now becomes one of determining this threshold in some best manner. Considering the Gaussian assumption, we investigate the false-alarm probability, $Pr(\eta_k'(f_l) < \psi_k(f_l)|H_0)$. A false alarm is realized when the SNR measure is larger than the threshold, given the null hypothesis is present. We may analytically determine the

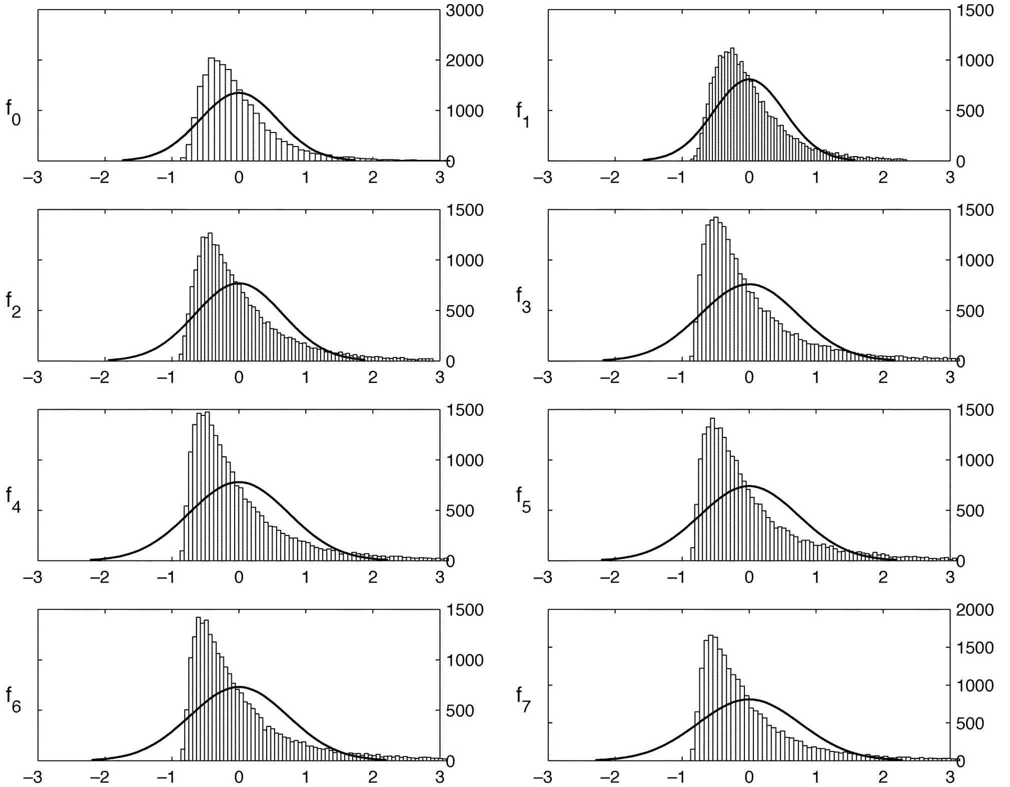


Fig. 3. Histograms of SNR measure in Babble noise environment for each frequency bin f_n.

threshold by considering the false-alarm probability and the assumed pdf under the null hypothesis

$$Pr\left(\eta'_k(f_l) < \psi_k(f_l)|H_0\right) = \int_{\eta'_k(f_l)}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2_{v,k}(f_l)}}\exp\left(\frac{-\psi^2_k(f_l)}{2\sigma^2_{v,k}(f_l)}\right) d\psi_k(f_l). \tag{9}$$

It is clear that this threshold $\eta'_k(f_l)$ can be determined from the stated false-alarm probability $Pr(\eta'_k(f_l) < \psi_k(f_l)|H_0)$. Manipulating (9) we find

$$\begin{aligned} Pr\left(\eta'_k(f_l) < \psi_k(f_l)|H_0\right) &= \int_{\eta'_k(f_l)}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2_{v,k}(f_l)}}\exp\left(\frac{-\psi^2_k(f_l)}{2\sigma^2_{v,k}(f_l)}\right) d\psi_k(f_l) \\ &= \int_{\frac{\eta'_k(f_l)}{\sigma_{v,k}(f_k)}}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-y^2}{2}\right) dy \\ &= \frac{1}{2}erfc\left(\frac{\eta'_k(f_l)}{\sqrt{2\sigma^2_{v,k}(f_l)}}\right) \end{aligned} \tag{10}$$

where $erfc(u)$ is the complementary error function [17].

Solving for $\eta'_k(f_l)$ we find the threshold

$$\eta'_k(f_l) = \sqrt{2\sigma^2_{v,k}(f_l)} \cdot erfc^{-1}(2P_{FA}) \tag{11}$$

where P_{FA} is the probability of false alarm. In this manner we may set the threshold as determined by the noise statistics, and desired system performance, namely the false-alarm probability.

The final decision is made by a comparison of means

$$\frac{1}{L}\sum_{f_l=0}^{L-1}\psi_k(f_l) \gtrless^{H_1}_{H_0} \frac{1}{L}\sum_{f_l=0}^{L-1}\eta'_k(f_l) \tag{12}$$

where H_1 is decided if the average SNR is larger than or equal to the average threshold, otherwise H_0 is decided. This final decision rule was determined empirically; however, it is interesting to note that a decision may also be made in each spectral bin independently by comparing $\psi_k(f_l)$ to $\eta'_k(f_l)$ for each f_l. Such a comparison would be equivalent to an optimal test in each spectral bin based on the noise statistics and further, would provide an independent VAD decision in each bin.

A. Discussion

The proposed VAD scheme only considers the statistics of the noise, and disregards statistical information about the speech signal. Further, the proposed scheme relies solely on the assumption that there will be a significant shift in mean during periods of speech activity. If the proposed scheme were implemented with no constraints on the estimated noise statistics, the scheme would fail in highly variable noise environments. This is because the fundamental assumption that the speech introduces a significant shift in mean will be violated, due to the high variance of the noise. The resulting VAD would then exhibit undesirable behavior, whereby significant false rejections (or miss-detections) would occur in highly variable noise environments such as the babble noise environment. To guard against this, appropriate constraints need to be placed on the estimated variance of the noise. This is implemented in the proposed scheme by limiting the threshold. These constraints effectively tradeoff false rejections for false alarms.

Further, an interesting point is that the developed threshold is in fact SNR independent. The threshold depends only on the background noise statistics. The lower the variance in a particular spectral bin, the lower the threshold. Hence, the less the noise background environment changes with time (i.e., it has a low variance), the better the scheme will perform. However, as the SNR becomes lower, the fundamental assumption that there will be a significant shift in mean during periods of speech becomes weaker. Low energy portions of speech are first to be falsely rejected. In order to counter this a hangover scheme is incorporated.

It is interesting to examine the detection problem if it were derived by including the speech statistics. Usually a likelihood ratio is formed then a Bayes test or similar performed, for example see [7] and [8]. Here, however, we will compare the SNR measure directly to a threshold determined by considering the probability of detection, rather than the probability of false alarm. By doing this we may directly compare the different optimization criteria. To examine this, we assume a Gaussian pdf under the alternative, with a nonzero mean

$$p\left(\psi_k(f_l)|H_1\right) = \frac{1}{\sqrt{2\pi\sigma^2_{s+v,k}(f_l)}} \cdot \exp\left(\frac{-\left[\psi_k(f_l)-\mu_s(f_l)\right]^2}{2\sigma^2_{s+v,k}(f_l)}\right) \qquad (13)$$

where $\mu_s(f_l)$ is the mean of the SNR measure in the f_l^{th} spectral bin during speech periods and $\sigma^2_{s+v,k}(f_l)$ is the variance of the SNR measure during speech periods in the f_l^{th} spectral bin. We may follow a similar procedure as for the false-alarm rate and determine the best threshold based on the correct detection rate (deciding the alternative, given the alternative is present)

$$\begin{aligned} &Pr\left(\eta'_k(f_l) < \psi_k(f_l)|H_1\right) \\ &= \int_{\eta'_k(f_l)}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2_{s+v,k}(f_l)}} \cdot \exp\left(\frac{-\left[\psi_k(f_l)-\mu_s(f_l)\right]^2}{2\sigma^2_{s+v,k}(f_l)}\right) d\psi_k(f_l) \\ &= \int_{\frac{\eta'_k(f_l)-\mu_s(f_l)}{\sqrt{\sigma^2_{s+v,k}(f_k)}}}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-y^2}{2}\right) dy \\ &= \frac{1}{2} erfc\left(\frac{\eta'_k(f_l)-\mu_s(f_1)}{\sqrt{\sigma^2_{s+v,k}(f_k)}}\right). \end{aligned} \qquad (14)$$

Rearranging as before, we find

$$\eta'_k(f_l) = \sqrt{2\sigma^2_{s+v,k}(f_l)} \cdot erfc^{-1}(2P_C) + \mu_s(f_l) \qquad (15)$$

where P_c is the probability of correct detection. Here we see that a threshold determined in such a manner is dependant on one extra parameter, the mean of the SNR measure during speech periods and also on the variance of the SNR measure during speech periods. Further, it is common to assume that background noise is long-term stationary, whereas speech is not long-term stationary. Thus, the problem becomes one of estimating these parameters. Herein lies the difficulty, since speech is only quasi-present, therefore during periods of nonspeech activity it would not be possible to determine the threshold, such as before the initial occurrence of speech. This nonstationary manner makes it counter-intuitive to estimate the threshold based on the speech statistics.

However, if a threshold were to be determined from speech statistics two possible scenarios should be considered. Firstly generate short-term estimates of the speech statistics from frame to frame, resulting in a threshold that will vary from frame to frame. Or, secondly use long-term estimates to determine the speech statistics, this will result in poor performance because of the nonstationary nature of the speech. It therefore makes intuitive sense to optimize the threshold based on the assumed stationary noise statistics.

IV. Hangover Scheme

In a practical implementation, a hangover scheme is required to lower the probability of false rejections [7]. The hangover scheme does this by reducing the risk of a low-energy portion of speech at the end of an utterance being falsely rejected, by arbitrarily declaring a period of speech activity after a period of speech activity has already being detected. This is based on the idea that speech occurrences are highly correlated with time. A hangover scheme can be implemented as a state machine and is also well visualized in this manner.

Fig. 4 shows the state machine hangover scheme as implemented in the VAD. The parameter D in the figure indicates the decision as made by testing the SNR with the threshold. D is assigned 1 if the SNR is larger than or equal to the threshold or 0 if the SNR is less than the threshold. This slightly biases the VAD toward an active decision. This value is then used to determine which state the machine should be in. The parameter VAD then specifies the final VAD decision.

Initially the hangover scheme is in the noise state, indicating no speech is present. The final VAD decision at this point is inactive $(VAD = 0)$. The parameter D is now used to determine how the hangover scheme should proceed. When $D = 1$, the state machine begins to progress through the transition states toward the speech state. At this point it is not clear if speech is present or not, because the parameter D may have been the result of a false alarm.

After four consecutive indications that speech is present $(D = 1)$, the hangover scheme will enter the speech state. Four states are chosen because of the low probability of four consecutive false alarms. The hangover scheme will remain in this state until the parameter D indicates speech is no longer present $(D = 0)$.

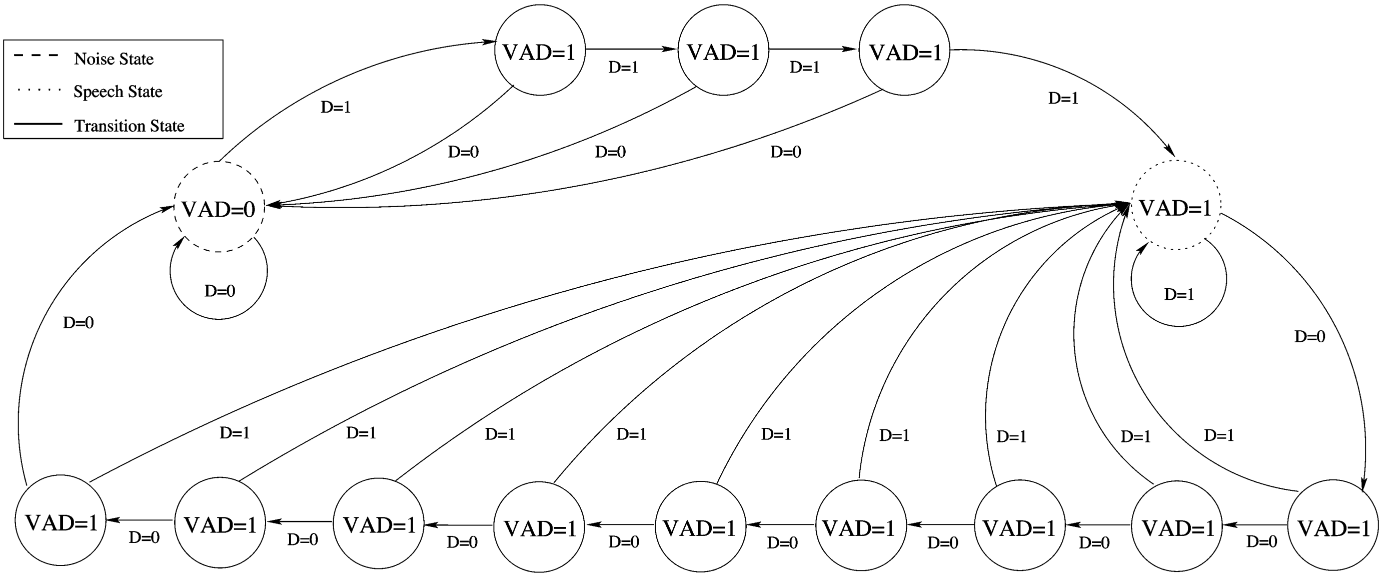


Fig. 4. Hangover scheme state machine.

When this event occurs, the hangover scheme will begin to progress through transition states toward the noise state. This is required because there is some ambiguity whether speech is still present, due to the fact that the result $D = 0$ may be attributed to a false rejection. After ten consecutive noise indications $(D = 0)$, the hangover scheme will again enter the noise state and wait for a speech indication $(D = 1)$. Ten consecutive states were chosen because this represents 100 ms, at 8000-Hz sampling frequency with 50% overlap and 20-ms length frames, which we found was generally enough to bridge a falsely rejected phone.

The hangover scheme influences the behavior of the VAD in two distinct ways.

- First, the scheme provides a quick transition from inactive to active. This is because speech activity is unconditionally declared if the statistical test indicates speech is present, even if the result may be attributed to a false alarm.
- Second, the scheme delays the transition from active to inactive. Meaning even if the statistical test indicates no speech is present, the VAD will not necessarily decide no speech activity, but will begin to progress through the transition states to the noise state. This effectively delays the transition from active to inactive and results in a reduction in false rejections. This behavior however may result in an increase of false alarms, especially in high-SNR environments.

The hangover scheme in [7] is based on a complex Markov modeling architecture requiring iterative evaluation of joint observation and state probabilities. The empirical scheme we propose involves a simple implementation in either hardware or software, very low memory and computational requirements and high effectiveness. The scheme is however based solely on the idea that speech occurrences are highly correlated with time and further that some hangover will always be required. This assumption may possibly result in an increase of false alarms through the addition of erroneous hangover, especially in high-SNR situations.

TABLE I
PARAMETERS FOR VAD IMPLEMENTATION

Measure	Value
$\eta'_k(f_l)_{MAX}$	1.5
$\eta'_k(f_l)_{MIN}$	0.45
$\hat{P}_{vv,MIN}$	0.001
$\alpha_{C,\psi_k(f_l)}$	0.75
$\alpha_{P_{vv}}$	0.999
$\alpha_{\sigma_v^2}$	0.35
$\alpha_{\eta'}$	0.75
L	16
M	19
P_{FA}	5%

V. IMPLEMENTATION

Several factors need to be considered when implementing the proposed VAD for a practical application. In order to ensure proper operation in a variety of environments, factors such as the threshold, noise power and spectral resolution need to be considered. The presented parameter values are suitable for general applications.

One key factor is the adaptive threshold. The threshold should be constrained between an upper and lower limit. This effectively limits the estimated variance of the background noise. This constraint implies that there is always some assumed noise variability; however, the variability is less than some upper boundary. The use of this upper boundary effectively trades false rejections for false alarms in highly variable environments such as the babble noise environment. This limiting is applied by clamping $\eta'_k(f_l)$ between an upper limit $\eta'_k(f_l)_{\mathrm{MAX}}$ and a lower limit $\eta'_k(f_l)_{\mathrm{MIN}}$. Through observation appropriate values for the constraints were found and are presented in Table I.

Another important aspect of the VAD implementation is the expected noise power estimate $\hat{P}_{vv}(f_l)$. If this value becomes small, it is possible for the SNR measure to tend toward infinity. Therefore, $\hat{P}_{vv}(f_l)$ should be constrained to a lower level.

This constraint however may introduce some bias to the SNR measure during periods of high SNR. The constraint is applied by limiting $\hat{P}_{vv}(f_l)$ to be larger or equal to a minimum value $\hat{P}_{vv,MIN}$. Through observation a value of 0.001 was found to give good performance for normalized input data; see Table I.

A further point to consider is smoothing of the SNR measure. The authors of [8] noted that smoothing of the likelihood ratio across time resulted in increased performance, this can be contributed to the variance reduction the smoothing generates. The SNR measure is however a low-variance estimate and no further variance reduction is required, the SNR measure however should be smoothed using a conditional exponential average. This average should be applied during decaying periods [18]. The averaging is applied in the following manner:

$$\hat{\psi}_k(f_l) = \left(1 - \alpha_{\psi_k(f_l)}\right) \cdot \psi_k(f_l) + \alpha_{\psi_k(f_l)} \cdot \hat{\psi}_{k-1}(f_l). \quad (16)$$

The coefficient $\alpha_{\psi_k(f_l)}$ is determined as follows:

$$\alpha_{\psi_k(f_l)} = \begin{cases} \alpha_{C,\psi_k(f_l)}, & \psi_k(f_l) \leq \psi_{k-1}(f_l) \\ 0, & \psi_k(f_l) > \psi_{k-1}(f_l) \end{cases} \quad (17)$$

where $\alpha_{\psi_k(f_l)}$ is the averaging coefficient and $\alpha_{C,\psi_k(f_l)}$ is a constant value. The conditional smoothing is not used as a variance reduction mechanism, but as a method to delay the transition from a high-energy portion of speech to a low-energy portion. This is done in an attempt to bridge low-energy portions of speech in the middle of an utterance, that directly follow a high-energy portion, in effect an energy dependant short hangover mechanism. Through practical observations an appropriate value for $\alpha_{C,\psi_k(f_l)}$ was found and is presented in Table I.

Further averaging is required to smooth the threshold $\eta'_k(f_l)$ and variance $\sigma^2_{v,k}(f_l)$ estimates. This averaging is applied in an exponential manner similar to (16). The smoothing coefficients $\alpha_{\eta'}$ and $\alpha_{\sigma^2_v}$, that have been found to produce good results, respectively, are presented in Table I.

It is also possible to update the noise PSD estimate $\hat{P}_{vv}(f_l)$. This may be done recursively during periods of nonspeech activity. The update should be slow and reflect the assumed stationarity of the background noise. With this in mind the noise PSD estimate is updated using an exponential average with coefficient $\alpha_{P_{vv}}$. An appropriate value for this was found through observation and is presented in Table I.

Spectral resolution also needs to be carefully considered. The resolution needs to be such that it is wide enough to suitably reduce the variance while simultaneously not encompassing too much bandwidth. For a sampling frequency of 8000 Hz, it was found that eight bands with a bandwidth of 500 Hz each were appropriate. Therefore, the appropriate subframe length L, was found to be 16. With a frame length of 160 samples, this results in $M = 19$ overlapping subframes at 50% overlap.

Finally, the probability of false alarm should also be considered. This value changes the sensitivity of the VAD. A low value will result in a sensitive VAD producing increased false alarms, whereas a high value will result in a less sensitive VAD that exhibits increased missrejections. This one parameter can be used to easily control the behavior of the VAD. Through testing an appropriate value for this was found to be 5%.

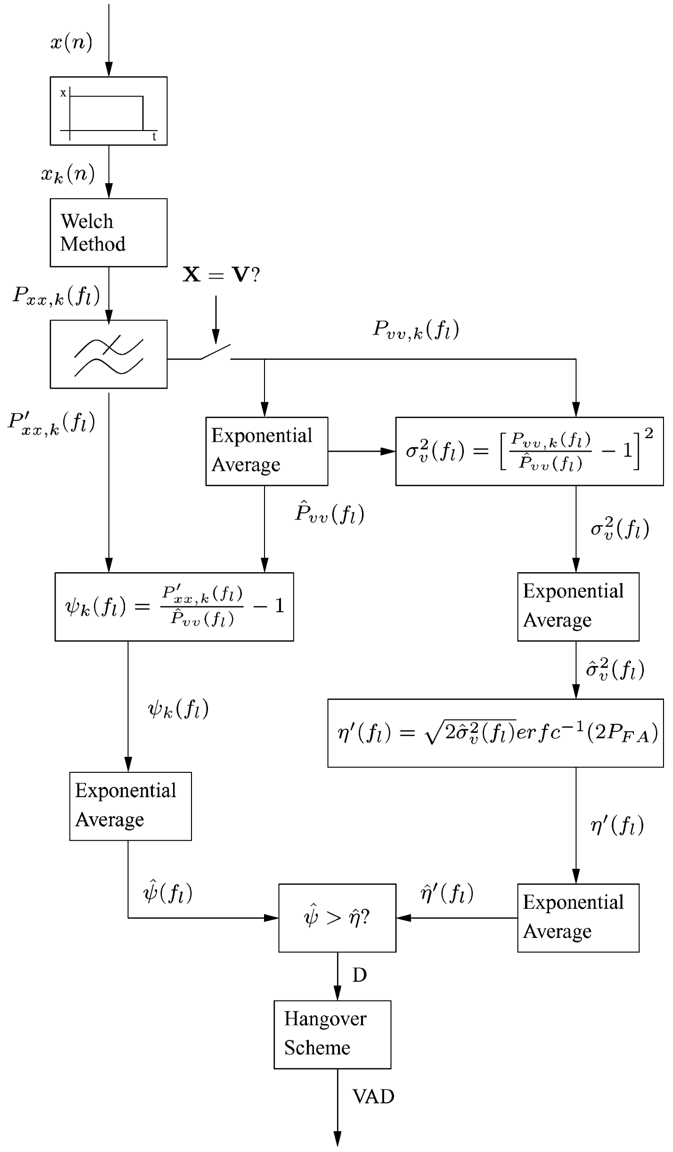


Fig. 5. Implemented system block diagram.

A. Final System

The final implemented system is shown in Fig. 5. Initially the raw data $x(n)$ is framed into 20-ms frames with 50% overlap between frames. The Welch method of overlapping subframes is used to estimate the reduced-variance, reduced-resolution PSD, $P_{xx,k}(f_l)$. A high-pass filter is also applied at this point to remove undesirable low-frequency components. Following that, the SNR $\psi_k(f_l)$ is calculated using the noise power $\hat{P}_{vv}(f_l)$ and the current PSD estimate. The current SNR $\psi_k(f_l)$ is then applied to a short exponential average over time. The arithmetic mean over frequency is then found and compared to the threshold η'.

The threshold η' is found by first calculating the variance of the SNR measure during periods of noise $\sigma^2_{v,k}(f_l)$. This variance is then exponentially averaged over time. The threshold $\eta'_k(f_l)$ is then calculated according to (11). This threshold is also exponentially averaged over time. The arithmetic mean over frequency of the threshold is then found and compared to the SNR, resulting in the preliminary decision D. After the comparison,

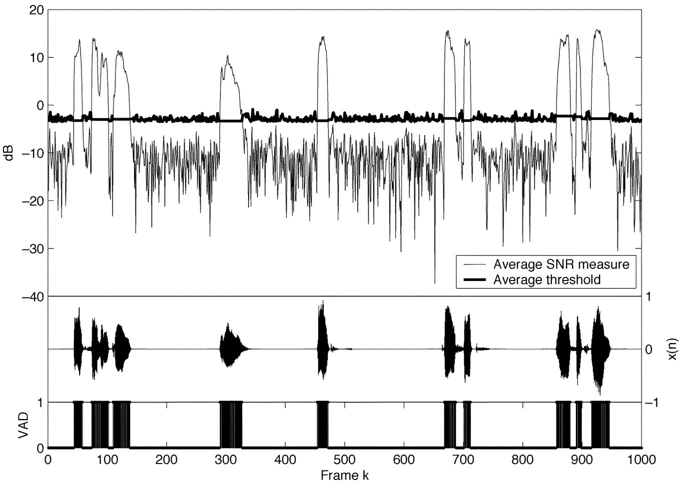


Fig. 6. Average SNR measure over frequency with average threshold over frequency for a 10-s utterance in Gaussian noise, where the hangover scheme and averaging have been disabled. The final VAD decision is shown in the lower section.

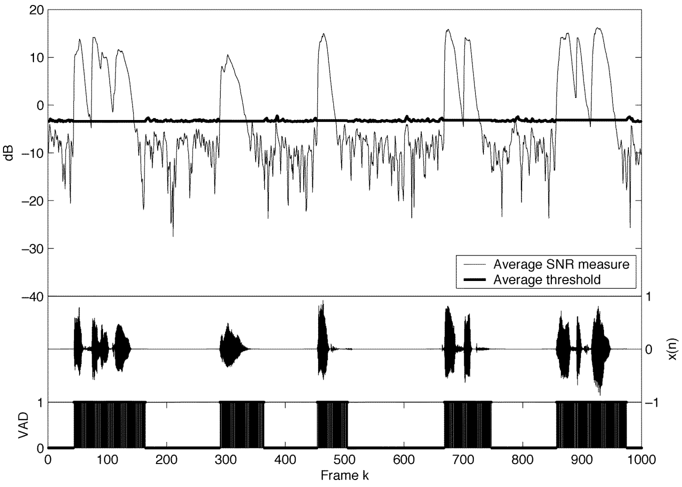


Fig. 7. Average SNR measure over frequency with average threshold over frequency for a 10-s utterance in Gaussian noise, where the hangover scheme and all averaging are enabled. The final VAD decision is shown in the lower section.

the preliminary decision D is subjected to the hangover scheme as described in Section IV. Finally, a decision is made.

The averaging and hangover scheme play an integral role in the VAD decision. Fig. 6 shows an example of the SNR measure and threshold as estimated with no exponential averaging or hangover, for a 10-s speech utterance in Gaussian noise with an average SNR of 5 dB. Note, updating of the noise PSD is still active. The utterance chosen was a sequence of spoken digits, both continuous and discrete, “eight seven three, silence, nine, silence, eight, silence, four six, silence, one six one.” The final VAD decision is also shown along with the clean utterance waveform. The final VAD decision shows poor detection of low-energy phonemes such as the fricatives /s/ and /th/ in seven and three respectively. Conditional exponential averaging and a hangover scheme may be employed to reduce this behavior.

Fig. 7 indicates the SNR measure and threshold as estimated with exponential averaging and the hangover scheme active. The effect of the averaging and hangover are clear. The false rejections have been reduced; however, there is an accompanying increase in false alarms. This increase is, however, preferable to a large amount of false rejections.

VI. Voice Activity Detector Evaluation

Using the implemented system outlined in Section V-A, the effectiveness of the proposed algorithm was evaluated. Surveying literature indicates two distinct schools pertaining to VAD evaluation, namely subjective and objective evaluation. In general, subjective evaluation methods attempt to determine the effect of erroneous VAD decisions on human perception [19]. Tests such as the ABC [19] however do not take into consideration the effect of false alarms and as such are inappropriate for a thorough evaluation of VAD performance. Therefore, in order to evaluate the performance of the proposed scheme objective evaluation was used. A testing strategy has previously been presented by Freeman *et al.* [20] and further by Beritelli *et al.* [6]. We apply that strategy whereby the output of the VAD is compared to a set of ideal decisions. This comparison yields five parameters indicating VAD performance,[1] after Beritelli *et al.* [6].

- *FEC (front end clipping):* Clipping due to speech being misclassified as noise in passing from noise to speech activity.
- *MSC (mid speech clipping):* Clipping due to speech misclassified as noise during an utterance.
- *OVER (over hang):* Noise interpreted as speech due to the VAD flag remaining active in passing from speech activity to noise.
- *NDS (noise detected as speech):* Noise interpreted as speech within a silence period.
- *Correct (correct VAD decision):* Correct decisions made by the VAD.

The clipping parameters (FEC and MSC) collectively maybe interpreted as indicators of false rejections and therefore should be minimized. The OVER and NDS parameters give an indication of false alarms and should also be minimized to yield best system performance. The correct parameter should be maximized and indicates the amount of correct decisions made by the VAD under test.

In order to gain a comparative analysis of the proposed VADs performance, several modern standardized VAD schemes were also evaluated. The schemes used were the two ETSI AMR VADs options 1 and 2 [4] and the ITU G.729 Annex B VAD [5]. The implementations were taken from the authors C implementations respectively [21], [22]. All baseline schemes include a hangover scheme within their algorithm description and are standardized, thus suitable for a general comparison. Data was resampled to 8000 Hz and requantized to 13-bit precision if required.

One important aspect of the comparison is the differing frame lengths used. The proposed scheme produces a decision every 10 ms, as does the G.729B VAD. The AMR VADs produce decisions every 20 ms. In order to compare the two, the decisions produced by the AMR VADs were compared to a set of ideal

[1] This varies from the traditional four through the introduction of the correct parameter.

TABLE II
VAD PERFORMANCE FOR VARIOUS SNRS AND NOISE ENVIRONMENTS

ENVIRONMENT		AMR VAD OPTION 1					AMR VAD OPTION 2				
Noise	*SNR*	*Correct*	*FEC*	*MSC*	*NDS*	*OVER*	*Correct*	*FEC*	*MSC*	*NDS*	*OVER*
Gaussian	0.00dB	75.32	1.70	17.27	5.62	0.09	88.39	0.93	9.72	0.15	0.81
Gaussian	5.00dB	85.73	0.85	8.40	4.68	0.35	94.60	0.50	2.92	0.21	1.78
Gaussian	10.00dB	91.67	0.51	3.64	3.43	0.76	96.21	0.28	0.75	0.28	2.48
Gaussian	15.00dB	95.32	0.28	1.59	1.69	1.13	96.47	0.15	0.20	0.38	2.79
Gaussian	20.00dB	97.09	0.17	1.10	0.47	1.17	96.35	0.07	0.08	0.49	3.01
Gaussian	25.00dB	97.34	0.10	1.24	0.30	1.01	96.30	0.05	0.05	0.57	3.03
Gaussian	**Average**	**90.41**	**0.60**	**5.54**	**2.70**	**0.75**	**94.72**	**0.33**	**2.29**	**0.34**	**2.32**
Babble	0.00dB	63.78	0.37	5.13	27.17	3.55	66.39	0.39	4.87	17.75	10.60
Babble	5.00dB	68.23	0.20	1.71	25.04	4.81	69.79	0.19	1.58	17.54	10.90
Babble	10.00dB	72.94	0.11	0.60	21.44	4.91	70.93	0.07	0.34	17.56	11.10
Babble	15.00dB	80.71	0.07	0.43	15.02	3.76	72.37	0.04	0.07	17.02	10.50
Babble	20.00dB	90.40	0.06	0.59	6.70	2.25	78.37	0.02	0.04	13.69	7.88
Babble	25.00dB	95.78	0.06	1.08	1.65	1.43	88.03	0.02	0.03	7.74	4.18
Babble	**Average**	**78.64**	**0.15**	**1.59**	**16.17**	**3.45**	**74.31**	**0.12**	**1.16**	**15.22**	**9.19**
Vehicle	0.00dB	97.08	0.08	0.76	0.69	1.39	94.95	0.02	0.01	1.29	3.73
Vehicle	5.00dB	97.41	0.07	0.88	0.51	1.13	95.61	0.02	0.01	0.94	3.42
Vehicle	10.00dB	97.42	0.07	0.98	0.50	1.04	96.08	0.02	0.01	0.89	2.99
Vehicle	15.00dB	97.05	0.07	1.44	0.49	0.96	96.49	0.02	0.03	0.89	2.57
Vehicle	20.00dB	96.26	0.07	2.44	0.48	0.76	96.81	0.02	0.06	0.89	2.21
Vehicle	25.00dB	95.66	0.07	3.24	0.47	0.56	96.92	0.02	0.08	0.92	2.05
Vehicle	**Average**	**96.81**	**0.07**	**1.62**	**0.52**	**0.97**	**96.15**	**0.02**	**0.04**	**0.97**	**2.83**
Average	Average	88.62	0.27	2.92	6.46	1.72	88.39	0.16	1.16	5.51	4.78

ENVIRONMENT		ITU-T G729B VAD					PROPOSED SCHEME				
Noise	*SNR*	*Correct*	*FEC*	*MSC*	*NDS*	*OVER*	*Correct*	*FEC*	*MSC*	*NDS*	*OVER*
Gaussian	0.00dB	66.66	2.12	26.27	4.95	0.00	90.22	1.23	8.22	0.00	0.33
Gaussian	5.00dB	73.94	1.21	19.89	4.95	0.01	96.03	0.69	2.38	0.01	0.88
Gaussian	10.00dB	79.81	0.80	14.41	4.95	0.03	97.38	0.40	0.56	0.04	1.62
Gaussian	15.00dB	84.53	0.54	9.88	4.96	0.08	97.20	0.23	0.14	0.09	2.35
Gaussian	20.00dB	87.94	0.36	6.53	4.97	0.21	96.83	0.17	0.09	0.17	2.73
Gaussian	25.00dB	90.71	0.23	4.23	4.34	0.48	96.90	0.22	0.17	0.18	2.54
Gaussian	**Average**	**80.60**	**0.88**	**13.54**	**4.85**	**0.13**	**95.76**	**0.49**	**1.93**	**0.08**	**1.74**
Babble	0.00dB	59.37	0.92	19.53	19.83	0.35	69.05	0.52	4.19	21.44	4.80
Babble	5.00dB	64.40	0.65	15.41	19.24	0.30	75.93	0.37	1.72	17.56	4.41
Babble	10.00dB	69.13	0.46	11.40	18.71	0.30	84.70	0.33	0.89	10.98	3.10
Babble	15.00dB	72.84	0.30	8.00	18.52	0.33	92.80	0.33	0.55	4.17	2.15
Babble	20.00dB	76.12	0.20	5.25	17.98	0.44	96.55	0.31	0.37	0.78	1.99
Babble	25.00dB	80.50	0.14	3.54	15.17	0.65	97.27	0.28	0.26	0.13	2.06
Babble	**Average**	**70.39**	**0.45**	**10.52**	**18.24**	**0.39**	**86.05**	**0.36**	**1.33**	**9.18**	**3.08**
Vehicle	0.00dB	64.35	0.20	6.27	28.71	0.47	96.75	0.43	1.03	0.17	1.61
Vehicle	5.00dB	68.14	0.17	4.54	26.72	0.44	96.91	0.41	0.89	0.11	1.68
Vehicle	10.00dB	71.87	0.12	3.22	24.25	0.55	97.04	0.39	0.78	0.09	1.70
Vehicle	15.00dB	77.82	0.10	2.49	18.85	0.74	97.21	0.37	0.59	0.08	1.74
Vehicle	20.00dB	87.65	0.13	2.50	8.79	0.92	97.33	0.33	0.41	0.08	1.85
Vehicle	25.00dB	94.19	0.16	2.90	1.82	0.93	97.34	0.30	0.29	0.09	1.99
Vehicle	**Average**	**77.34**	**0.15**	**3.65**	**18.19**	**0.67**	**97.10**	**0.37**	**0.67**	**0.11**	**1.76**
Average	Average	76.11	0.49	9.24	13.76	0.40	92.97	0.41	1.31	3.12	2.19

decisions generated every 20 ms from samplewise hand-labeled data. The G729B VAD and the proposed scheme were both compared to a set of ideal decisions generated every 10 ms from samplewise hand-labeled data.

Sarikaya and Hansen [15] noted that the choice of test data is important in VAD evaluation, since it is simple to optimize a particular scheme for a small test set. Here, we have thus used the entire TIMIT TEST corpus, consisting of 168 individual speakers each speaking 10 sentences. Sentences were concatenated in sets of four and silence was inserted between sentences. The amount of silence between sentences was randomly chosen; however, the total amount of silence in the set was constrained to be 60% of samples, this included silence inherent in the sentences such as pauses. The resulting set thus consisted of 186 min of speech data, of which 40% was active samples, which is the amount of speech activity in a typical telephone conversation [3]. This data consisted of 1680 different spoken sentences, encompassing all phones and eight different dialects as defined in the TIMIT set.

The whole set was samplewise hand labeled from phone transcriptions. A frame of speech was considered active if the majority of the frame was active as based on the samplewise hand labeled data. Several noise environments were artificially added to the test set at varying SNRs. The noise used was taken from the NOISEX-92 database and consisted of babble, Gaussian, and vehicle noise. Including different noise environments and SNRs, the entire testing set consisted of 46.5 h of noisy speech. Results are presented in Table II.

VAD performance comparison is complicated and should be considered carefully. Ideally, a VAD should maximize the correct rate, and minimize all errors. However failing this, the affect different types of errors have on the discontinuous speech signal (speech signal with nonspeech periods removed and comfort noise inserted) should be considered. The purpose of a VAD in the context of a telephone conversation is to enable data savings by not transmitting nonspeech periods, while maintaining speech quality. Speech quality should be of utmost importance. Therefore, is it important to note the affect that each of the different errors have on speech quality.

Clipping errors such as FEC and MSC are a worst case scenario in terms of speech quality degradation. They remove portions of speech from the original signal, resulting in reduced speech intelligibility. These errors should thus be avoided at all costs.

In contrast, insertion errors such as NDS and OVER do not have any affect on speech quality. They do however result in reduced effectiveness of the VAD scheme. These errors are therefore secondary to clipping errors, which reduce speech intelligibility. A relationship between these two errors, for example, one insertion error equals two clipping errors, is difficult to determine. Here we will use the broad notion that clipping errors are less desirable than insertion errors.

Examining Table II we note some interesting points.

- Both the ETSI schemes have very close average correct detection rates, with 0.23% separating the two.
- Excluding the OVER metric, the ITU-T G729 scheme exhibited the worst average results over the test set.
- Generally the largest contributor to total error was insertion errors rather than clipping errors.
- Clipping errors were generally worst in the Gaussian noise environment.

In more detail we see the proposed scheme exhibited the highest average correct detection rate. It was more than 4% higher than the ETSI schemes and 16% higher than the ITU-T G729 scheme. This is a good result; however, it is important to investigate the sources of error, since clipping errors are less desirable than insertion errors.

As noted earlier clipping errors are a worst case scenario for VAD schemes and should be avoided. The ETSI AMR VAD Option 2 scheme exhibited the lowest average clipping (FEC+MSC) over the entire test set. This was closely followed by the proposed scheme which lagged behind by 0.4%. Clipping errors play a large role in speech intelligibility as mentioned earlier, and high FEC can result in unintelligible speech. The proposed scheme had an average of 18.5× higher FEC in the vehicle noise environment as compared to the ETSI AMR VAD Option 2, which can have a derogatory affect on speech intelligibility in discontinuous transmission applications. This is possibly due to the fact that the proposed scheme is optimized for insertion errors, rather than clipping errors. The results highlight the possibility that the authors of the ETSI schemes were in contrast attempting to optimize for clipping errors such as these, since they play a large role in speech intelligibility, and the schemes are primarily for mobile telephony applications.

Insertion errors should be minimized to ensure the VAD is as effective as possible. The proposed scheme had the lowest average NDS over the entire test set. The proposed scheme also had a low average OVER rate, coming in at more than 50% better than the ETSI AMR VAD Option 2 scheme. This is an interesting result, since the proposed scheme is optimized for insertion errors. This complements the previous result, and again highlights the possibility that the ETSI VAD schemes are optimized for clipping errors.

It is interesting to study the babble noise test at this point. We see that the majority of error is due to insertion errors. We also see that the proposed scheme performs well in this environment, offering quite a large reduction in insertion errors. This is due to the fact that the scheme is optimized for insertion errors, in contrast to the baseline schemes that appear to be optimized for clipping errors.

A further interesting result is the constant NDS rate exhibited by the G729B VAD in the Gaussian noise environment. This result implies that the VAD is insensitive to input level, this is a good characteristic for a VAD to have. In contrast we see all other schemes exhibited some form of level sensitivity in varying NDS rates with differing SNR. It is also interesting to see that all schemes exhibited a reducing OVER rate as the SNR became lower in the Gaussian noise environment. This indicates that the hangover schemes became more effective as the SNR decreased. It is interesting that this is the case for all schemes tested.

Considering the proposed scheme performance in the Gaussian noise environment, we see the MSC rate increases with falling SNR. This is because the assumption that the speech introduces a significant shift in mean begins to fail, thus resulting in false rejections during low-energy portions of speech. Further, examining the behavior of the proposed scheme in the babble noise environment, we see that the NDS rate is not constant as expected, and in fact becomes greater with falling SNR. This is because during the high-SNR periods, the estimated noise level $\hat{P}_{vv,k}(f_l)$ is lower than the minimum allowed noise level. The minimum is thus stored as the noise level and the SNR becomes biased toward -1 and the perceived variance is reduced. Finally, it is also an interesting point that the scheme has a much lower false-alarm rate than the 5% that was used in the implementation. This is because of the upper and lower constraints placed on the threshold, the averaging of the SNR and the fact the hangover scheme influences false alarms.

In summary, the proposed scheme presents a good alternative to standardized algorithms. It exhibits a consistent correct rate over a variety of noise environments and conditions. It has lower average NDS than all standardized algorithms over the test set and has low FEC and MSC while maintaining a low

OVER rate. These characteristics make it a simple and reliable choice for many VAD applications. Further, the scheme requires only low computation time and memory. In the implementation presented here, 19 16-point real FFTs are required with only marginal further additions, multiplications and divisions per 10 ms of signal. Memory requirements are kept low by utilizing exponential averages that require only the previous averaged value. This low-complexity approach further increases the suitability of the proposed scheme as an alternative to standardized algorithms.

VII. Conclusion and Future Work

A novel voice activity detection method based on a statistical decision mechanism has been presented. The proposed VAD incorporates a low-variance spectrum estimation technique and a method for determining an adaptive threshold based on noise statistics. These innovations result in a statistical test that is simple and elegant, whilst maintaining a high detection rate and a low error rate. A possible algorithm implementation has been outlined along with a state machine based hangover scheme. The implementation was tested and compared to current standardized VAD algorithms.

The simplicity of the proposed VAD coupled with the encouraging results, mathematical tractability and high detection consistency make it a good alternative to current schemes. The behavior of the VAD is easily altered by changing one meaningful parameter, and as such makes the VAD well suited to varying applications.

There is still significant room for further research in this field of practical importance. Further avenues of investigation include the following.

- Investigation of alternative or possibly arbitrary distributions for characterization of the distribution of the SNR measure during periods of noise.
- Determination of an optimal weighting scheme for the spectral bins used during calculation of the average SNR.
- Use of minimum statistics to estimate the expected noise PSD during both periods of speech and nonspeech activity [23].
- Investigation of an SNR based hangover scheme where the amount of hangover addition is determined by the estimated SNR.
- Investigation and comparison of different hangover schemes.
- Investigation of different low-variance spectrum estimation techniques.

Acknowledgment

The authors would like to acknowledge the invaluable help of S.-Y. Low for his editing. The authors would also like to thank the reviewers for their insight and helpful comments. In particular one comprehensive reviewers comments led to a better development of the core ideas.

References

[1] R. V. Prasad, A. Sangwan, H. S. Jamadagni, M. C. Chiranth, R. Sah, and V. Gaurav, "Comparison of voice activity detection algorithms for VoIP," in *Proc. Int. Symp. on Computers and Communications*, 2002, pp. 530–535.

[2] *Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi Rate (AMR); Speech Processing Functions; General Description*, 1998.

[3] F. Beritelli, S. Casale, and G. Ruggeri, "Performance evaluation and comparison of ITU-T/ETSI voice activity detectors," in *Proc. IEEE ICASSP'01*, vol. 3, Salt Lake City, UT, 2001, pp. 1425–1428.

[4] *Digital Cellular Telecommunications System (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi Rate (AMR) Speech Traffic Channels; General Description*, 1999.

[5] ITU, Coding of Speech at 8 kbit/s Using Conjugate Structure Algebraic Code—Excited Linear Prediction. Annex B: A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommend. V.70, International Telecommunication Union, 1996.

[6] F. Beritelli, S. Casale, and A. Cavallaro, "A robust vouce activity detector for wireless communications using soft computing," *IEEE J. Select. Areas Commun.*, vol. 16, no. 9, pp. 1818–1829, Dec. 1998.

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

[8] Y. D. Cho, K. Al-Naimi, and A. Kondoz, "Improved statistical voice activity detection based on a smoothed statistical likelihood ratio," in *Proc. IEEE ICASSP'01*, vol. 2, Salt Lake City, UT, 2001, pp. 737–740.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[10] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaptation," in *Proc. IEEE ICASSP'98*, vol. 1, Seattle, WA, 1998, pp. 365–368.

[11] S. M. Kay, *Fundamentals of Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[12] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[13] B. L. McKinley and G. H. Whipple, "Model based speech pause detection," in *Proc. IEEE ICASSP'97*, vol. 2, 1997, pp. 1179–1182.

[14] A. Davis and S. Nordholm, "A low complexity statistical voice activity detector with performance comparisons to ITU-T/ETSI voice activity detectors," in *Proc. Joint Int. Conf. Information., Communications, Signal Processing, Pacific Rim Conf. Mulitmedia*, vol. 1, 2003, pp. 119–123.

[15] R. Sarikaya and J. H. L. Hansen, "Robust speech activity detection in the presence of noise," in *Proc. 5th Int. Conf. Spoken Language Processing*, 1997, pp. 922–925.

[16] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*. Reading, MA: Addison-Wesley, 1994.

[17] S. Haykin, *Communication Systems*, 3rd ed. New York: Wiley, 1994.

[18] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 799–807, Nov. 2001.

[19] F. Beritelli, S. Casale, and G. Ruggeri, "A physcoacoustic auditory model to evaluate the performance of a voice activity detector," in *Proc. Int. Conf. Signal Processing*, vol. 2, Beijing, China, 2000, pp. 69–72.

[20] D. K. Freeman, C. B. Southcott, I. Boyd, and G. Cosier, "A voice activity detector for pan-European digital cellular mobile telephone service," in *Proc. IEEE ICASSP'89*, vol. 1, Glasgow, U.K., 1989, pp. 369–372.

[21] *Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi Rate (AMR) Speech; ANSI-C Code for AMR Speech Codec*, 1998.

[22] ITU, Coding of Speech at 8 kbit/s Using Conjugate Structure Algebraic Code—Excited Linear Prediction. Annex I: Reference Fixed-Point Implementation for Integrating G.729 CS-ACELP Speech Coding Main Body With Annexes B, D and E, Int. Telecommun. Union, 2000.

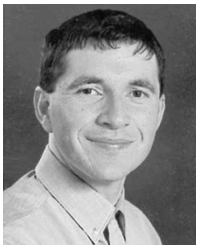
[23] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

Alan Davis (S'01) received the B.E. degree from Curtin University of Technology, Australia, in 2002. He is currently pursuing the Ph.D. degree at the Western Australian Telecommunications Research Institute (WATRI), a joint institute between Curtin University and the University of Western Australia.

His research interests include voice activity detection and array signal processing with applications to speech enhancement and automatic speech recognition.

Sven Nordholm (M'91–SM'04) received the Dipl.Eng. and Ph.D degrees from Lund University, Sweden, in 1983 and 1992, respectively.

From 1983 to 1986, he was a Development Engineer at GAMBRO, Sweden. He started his academic career at Lund University and subsequently co-founded the Department of Signal Processing at Blekinge Insitute of Technology, Sweden. He was appointed a Professor and Director of the Australian Telecommunications Research Institute in 1999. He is currently the Research Director for Signal Processing Laboratories in WATRI and the Research Executive for the Wireless Program ATcrc. His research interests are adaptive array processing, optimization methods, blind signal separation, equalization and filter design. He holds several patents and published more than 100 publications.



Roberto Togneri (M'89–SM'04) received the B.E. degree in 1985 and the Ph.D. degree in 1989 both from the University of Western Australia.

He joined the School of Electrical, Electronic and Computer Engineering at the University of Western Australia in 1988 as Senior Tutor, was appointed Lecturer in 1992 and Senior Lecturer in 1997. He has published over 20 papers and the book *Fundamentals of Information Theory and Coding Design*.

Dr. Togneri is a member of the Centre for Intelligent Information Processing Systems (CIIPS) and heads the Signals and Information Processing Group. His research activities include signal processing and feature extraction for speech signals, statistical and neural network models for speech recognition, and applications of spoken language technology. He was also the deputy-chair on the technical programme committee of the International Conference on Spoken Language Processing (ICSLP98).