AN APPROACH TO AN OPTIMIZED VOICE-ACTIVITY DETECTOR FOR NOISY SPEECH SIGNALS

Henning Puder and Oliver Soffke
Signal Theory, Darmstadt University of Technology
Merckstr. 25, D–64283 Darmstadt, Germany
Tel.: +49 (0)6151 16 2815, Fax: +49 (0)6151 16 3778
e-mail: henning.puder@ieee.org

ABSTRACT

In this paper, we present a voice activity detection (VAD) algorithm for noisy speech which is necessary for many applications such as source coding or speech enhancement methods. The proposed algorithm is characterized by a large conditional detection probability required for noise reduction and shows considerable improvement compared to known methods. The detection procedure is based on the output of an adaptive prediction error filter presented in [2]. Our approach utilizes this prediction error signal to generate a highly reliable voice activity detection with the described procedures. For comparing the results of different methods for voice activity detection, we apply the receiver operating characteristic (ROC). This ROC allows to judge the VAD quantitatively. Additionally, we utilize the ROC to motivate and formulate a new procedure to adapt the VAD threshold automatically to the prevailing SNR which minimizes the required heuristic parameter settings.

1 INTRODUCTION

Modern telecommunication technology is present in many areas of everyday life. The growing demand on bandwidth and better signal quality as well as the increasing number of users is the motivation to develop and realize new ideas: For example, algorithms are worked out to compress, denoise or reconstruct signals.

Many of these applications require voice activity detectors. Their task is to precisely detect, often based on a noisy signal, the time instances when speech is present.

One application of VAD algorithms is the enhancement of noisy speech, i.e. the suppression of background noise while preserving the natural sound of speech. Spectral subtraction based algorithms are mainly applied for this task. The basic idea of these algorithms is to decompose the noisy speech into its spectral components and to weigh these components according to their individual signal-to-noise ratio (SNR). This procedure assumes that the noise power of each spectral component, i.e. the power spectral density (PSD) of the noise, is known. For one-microphone solutions, this noise PSD has to be estimated based on the noisy speech. Assuming that the noise characteristics do not change rapidly, an estimation of these quantities in speech pauses is sufficient. To release or freeze this estimation process, a voice activity detector is necessary.

In this paper, we will present such a voice activity detector which fulfills the requirements of noise reduction namely a high recognition rate. To assess the algorithm, we utilize a receiver operating characteristic (ROC) which is adapted to the VAD. With this criteria, we are also able to optimize the results and describe an algorithm which guarantees optimal results independent of the SNR.

The paper is organized as follows: First, in section 2, we present the basic ideas of the algorithm, before going into details in section 3. Section 4 is dedicated to the evaluation of the algorithm with the help of the receiver operating characteristic. Based on this knowledge, a threshold adaptation is developed in section 5.

2 THE BASIC PRINCIPLE OF THE VOICE-ACTIVITY-DETECTION-ALGORITHM

The basic principle of the detector is presented in Fig. 1.

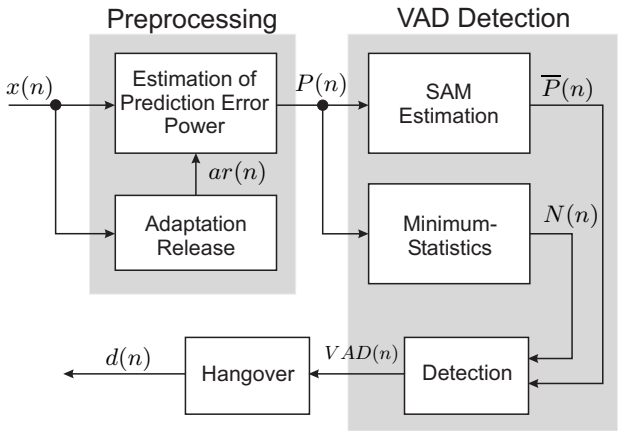

Figure 1: Block diagram of the proposed voice activity detector

The proposed voice activity detector, as presented in detail in the following section, may be decomposed into three main blocks according to Fig. 1.

First, a preprocessing unit generates a signal that is distinguished by a possibly large ratio of signal power during speech activity compared to speech pauses. This preprocessing unit mainly consists of a prediction error filter which is adapted to the PSD of the background noise as proposed by the GSM-VAD [1, 2]. However, to determine the time instances when the predictor is adapted, we propose a computationally much more efficient algorithm compared to the GSM proposal.

In a second step, this preprocessed signal is utilized to determine a reference signal and a threshold which are employed for a preliminary voice activity detection in the third step. Finally, a hangover algorithm avoids missed voice detection during speech sections with low excitation power.

3 DETAILED DESCRIPTION OF THE VOICE-ACTIVITY-DETECTION-ALGORITHM

3.1 Preprocessing

The aim of the preprocessing unit is to generate a signal at the output of the prediction error filter that allows the best distinction between speech activity and speech pauses. The components of this unit are depicted in Fig. 2.

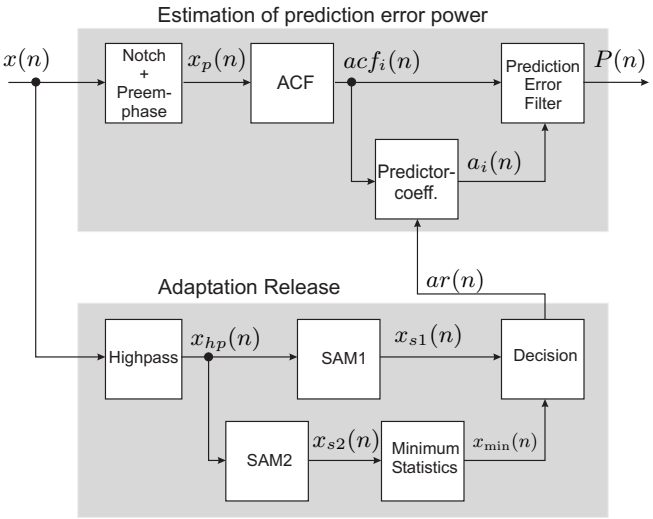

Figure 2: Subunits of the preprocessing algorithm

3.1.1 Estimation of the Prediction Error Signal Power

For the determination of the prediction error signal power, the input signal is first filtered with a zero-Hz-Notch filter to generate a signal with zero mean:

$$x_{oc}(n) = x(n) - x(n-1) + \alpha \cdot x_{oc}(n-1)\,,\ \alpha = 0.999\,. \quad (1)$$

In the next step, a preemphase filter amplifies the high frequency components of the signal which generally exhibit a larger SNR than the low frequency components when car noise is present

$$x_p(n) \quad = \quad x_{oc}(n) - \beta \cdot x_{oc}(n-1)\,,\ \beta = 0.86\,. \quad (2)$$

This signal is decomposed in overlapping blocks of length $N = 128$ with an overlap of $N - M = 64$ values, where M denotes the subsampling rate. The following steps, necessary to determine the prediction error signal power are performed in blocks.

First the autocorrelation function is calculated:

$$acf_i(n) = \sum_{k=nM+i}^{nM+N-1} x_p(k) \cdot x_p(k-i)\,,\ i = 0 \ldots L\,, \quad (3)$$

and the mean value of the last $F = 4$ blocks is determined by:

$$\overline{acf}_i(n) = \sum_{k=0}^{F-1} acf_i(n-k)\,,\ i = 0 \ldots L\,. \quad (4)$$

Whenever the adaptation of the prediction error filter is released (s. Sec. 3.1.2), i.e. the current signal block contains with high probability only noise and no speech, the coefficients of the prediction error filter are refreshed according to the Yule-Walker equation:

$$\mathbf{a}(n) \quad = \quad \mathbf{s}_{xx}^{-1}(n)\,\mathbf{q}(n), \quad (5)$$

$$\begin{aligned} \text{with:} \quad \mathbf{a}(n) &= [a_1(n), \cdots, a_L(n)]^T, \\ \mathbf{s}_{xx} &= \begin{bmatrix} \overline{acf}_0(n) & \cdots & \overline{acf}_{L-1}(n) \\ \vdots & \ddots & \vdots \\ \overline{acf}_{L-1}(n) & \cdots & \overline{acf}_0(n) \end{bmatrix}, \\ \mathbf{q}(n) &= [\overline{acf}_1(n), \cdots, \overline{acf}_L(n)]^T. \end{aligned}$$

The equation may be solved computational efficiently with the help of the Levinson-Durbin algorithm.

In the following, the prediction coefficients are utilized to determine the power of the prediction error signal:

$$P(n) = \left(x_p(n) - \sum_{i=1}^{L} a_i(n)\, x_p(n-i) \right)^2 . \quad (6)$$

Making use of the mean autocorrelation of the signal and the correlation of the prediction coefficients

$$r_i(n) = \sum_{k=0}^{L-i} a_k(n)\, a_{k+i}(n)\,,\ a_0 = -1\,,\ i = 0 \ldots L\,, \quad (7)$$

the power of the prediction error signal can also be determined by:

$$P(n) = r_0(n)\, \overline{acf}_0(n) + 2 \sum_{k=1}^{L} r_k(n)\, \overline{acf}_k(n), \quad (8)$$

with a prediction order of $L = 8$ according to the GSM voice activity detector [2].

3.1.2 Adaptation Release

As described earlier, the adaptation of the prediction coefficients is only released if pure noise and no speech is present with a high probability.

In contrast to the GSM voice activity detector [2], the release is determined based on a simple, only power based algorithm, named 'classical' VAD from now on. Its advantages are its low computational demands combined with its robustness. It functions as follows:

First, the input signal is filtered with a low-order IIR highpass having a cutoff frequency of 500 Hz to remove the low frequency components which generally show a low SNR. The output $x_{hp}(n)$ is then processed with two SAM (Short term average magnitude) estimators. The first determines the detection reference $x_{s1}(n)$ of the classical speech activity detector while the second determines the signal $x_{s2}(n)$ which is further processed with the minimum statistics algorithm [3] to obtain the threshold $x_{\min}(n)$.

An SAM estimator, generally defined as

$$\begin{aligned} x_s(n) &= \begin{cases} y_f & : \ |x_{hp}(n)| < x_s(n-1) \\ y_r & : \ |x_{hp}(n)| \geq x_s(n-1)\,, \end{cases} \\ \text{with:} \quad & y_f = \alpha_f \cdot x_s(n-1) + (1-\alpha_f) \cdot |x_{hp}(n)|\,, \\ & y_r = \alpha_r \cdot x_s(n-1) + (1-\alpha_r) \cdot |x_{hp}(n)|\,, \end{aligned}$$

offers the possibility to smooth a signal while tracking raising and falling signal magnitudes with different rates depending on the choice of α_f and α_r. With $\alpha_{r,1} < \alpha_{f,1}$ chosen for the first estimator, it is possible to track raising signal slopes faster than falling slopes. The output, which is utilized as detection reference, thus allows a good tracking of speech activity from the beginning. To tide over low power speech sections, the falling smoothing constant is chosen larger.

We utilized the following values: $\alpha_{f,1}$=0.999, $\alpha_{r,1}$=0.95, $\alpha_{f,2}$=0.995, $\alpha_{r,2}$=0.995.

For the second estimator, the smoothing constants are chosen identically $\alpha_{r,2} = \alpha_{f,2} < \alpha_{f,1}$ in order to retrack the average noise magnitudes after speech activity relatively fast. The signal $x_{s2}(n)$ is further processed by the minimum statistics which delivers a continuous estimate $x_{\min}(n)$ of the average noise magnitude. Multiplied by a factor $v = 3.0$ which equalizes the bias of the first SAM estimator, the threshold is obtained. Whenever the detection reference signal is smaller than the threshold, the adaptation of the prediction coefficients is released:

$$ar(n) \quad = \quad \begin{cases} 1 & : \quad x_{s1}(nM) < v \cdot x_{\min}(nM) \\ 0 & : \quad \text{else} \end{cases} . \quad (9)$$

The adaptation release is required in the subsampled rate only. To economize computational power, the minimum statistics algorithm may also be driven in the subsampled rate. The factor v is chosen such as the adaptation is only released during reliably detected speech pauses.

3.2 Detection of Voice Activity

The real voice activity detection is now based on the prediction error signal power $P(n)$ which is determined for every signal block. The reference signal of the detection is calculated by an SAM smoothing of the power $P(n)$ comparable to the classical detector:

$$\overline{P}(n) \quad = \quad \begin{cases} \overline{P}_f(n) & : \quad P(n) < \overline{P}(n-1) \\ \overline{P}_r(n) & : \quad P(n) \geq \overline{P}(n-1) \end{cases} \quad (10)$$

$$\text{with:} \qquad \overline{P}_f(n) = \beta_f \cdot \overline{P}(n-1) + (1-\beta_f)\, P(n)$$
$$\overline{P}_r(n) = \beta_r \cdot \overline{P}(n-1) + (1-\beta_r)\, P(n)\,,$$

where the falling smoothing constant $\beta_f = 0.7$ is chosen larger than the raising constant $\beta_r \leq 0.3$.

As $P(n)$ is already a smoothed magnitude due to the smoothing of the autocorrelation function, it can be directly used as the input for the Minimum-Statistics:

$$N(n) = \text{MINSTAT}\,\{P(n)\}\,. \quad (11)$$

Finally a preliminary decision $VAD(n)$ is determined as follows:

$$VAD(n) = \begin{cases} 1 & : \quad \overline{P}(n) \geq b \cdot N(n) \\ 0 & : \quad \text{else} \end{cases} . \quad (12)$$

The factor b, used to raise the minimum is chosen as a fixed value for the present. In section 5, it is shown that with an adaptive factor, further enhancement of the algorithm can be achieved.

3.3 The Hangover Algorithm

When high background noise is present, it is nearly impossible – even with an optimal adapted prediction error filter – to distinguish low power speech sections from noise. Thus, the probability is very high that these sections are accidentally considered as pure noise. A first approach to avoid this, is to increase the falling smoothing constant β_f when calculating the reference signal. The reference signal is then decreasing slower which lengthens the detected speech sections. Nevertheless, this procedure suffers from the disadvantage that the length by which the detected speech sections are increased becomes dependent on the SNR: The larger the SNR, the more time is necessary before the reference signal falls below the threshold at the end of each speech section. Therefore, instead of increasing β_f, the following hangover-algorithm is utilized:

Every detected speech activity section is held for a certain time $T_{\max}$ before switching to speech pauses again. However, to avoid that false detections are lengthened undesirably long, the duration for which speech activity is held, is at most doubled:

$$T_{hold} = \min(T_{\max}, T_{speech}). \quad (13)$$

where T_{speech} is the length of the last detected speech section. We obtained the best results with $T_{\max} = 0.2s$.

This procedure compromises well the two requirement to completely detect speech activity sections and to limit false detections as much as possible. The final VAD signal is then denoted as $d(n)$.

4 EVALUATION OF VAD ALGORITHMS

In order to evaluate and compare algorithms for voice activity detection, it is necessary to consider both, the conditional detection probability (P_d) and the conditional false alarm probability (P_f). Here the value P_d is the probability that speech is detected at the condition that speech is present and P_f the probability that speech is detected when no speech is present. For speech activity detectors these probabilities are strictly related to the threshold, the reference signal is compared with. For a high threshold these two probabilities are small and for a low threshold both are high.

In order to evaluate the potential of a VAD and to compare it with others, a method known by radar detection may successfully be applied: the receiver operating characteristic (ROC). This characteristic is obtained by varying a parameter of the decision unit and plot the conditional detection probability as a function of the conditional false alarm probability for every parameter value. The threshold is mostly utilized for the variable parameter. In our case the parameter is the factor b by which the threshold is raised.

Increasing the threshold (or the factor b) beginning with zero, we obtain a graph that starts in the upper right and ends in the lower left corner. The optimum is given in the upper left corner equivalent to 100 % detection 0 % false alarm probability. The closer the graph reaches this optimum, the larger is the potential of the detector.

An example of such ROCs for the proposed VAD is given in Fig. 3. The two graphs are obtained for speech signals with

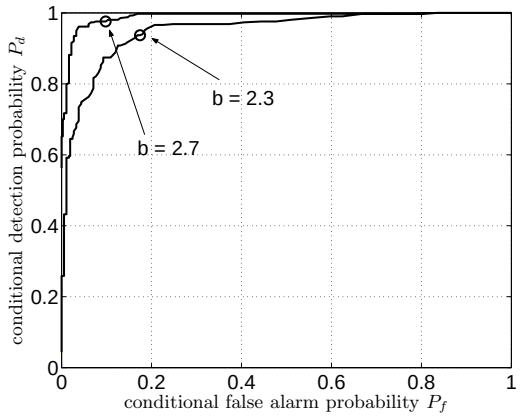

Figure 3: Two ROC graphs obtained with the proposed VAD for noisy speech signals with different SNR

different SNR. The optimal operating point is indicated with the corresponding factor b. The optimization criteria for the operating point is adapted to the requirements of noise

reduction: A high detection probability is more important than a low false alarm probability. Two properties become obvious when evaluating the results of Fig. 3:

- The larger the SNR of the noisy speech signal, the closer are the ROC graphs to the optimal upper left corner.
- The optimal values of the factor b depend on the signal to noise ratio.

5 OPTIMIZED THRESHOLD CALCULATION

The second item of the just mentioned properties shows that it is possible to determine the potential of the detectors in principle with a ROC. Nevertheless, the optimal operating point has still to be fixed. As one can observe in Fig. 3 the factor b should be increased with increasing SNR.

We decided therefore to design a procedure that determines the factor b in a linear dependency of the SNR, i.e. the quotient of the signal power and the noise power. For the signal power we utilized short term maxima values of the signal calculated with an SAM estimator that follows signal maxima by the choice of the smoothing constants to $\gamma_r = 0$ und $0 \ll \gamma_f < 1$. Additionally, the values are updated only during speech activity.

$$x_{\max}(n) = \begin{cases} x_{\max}(n-1) & : \quad VAD(n) = 0 \\ \gamma_f \, x_{\max}(n) + (1-\gamma_f)\,|x_{hp}(n)| & : \\ \quad VAD(n) = 1 \wedge |x_{hp}(n)| \leq x_{\max}(n-1) & \\ |x_{hp}(n)| & : \quad \text{else} \end{cases} \tag{14}$$

For the input of the SAM estimator, the signal $x_{hp}(n)$ is utilized which was filtered by a 500 Hz highpass (s. Sec. 3). Here the prediction error signal power $P(n)$ offers no advantages. On the contrary, the amplification of the high frequencies by the prediction error filter results in very high values of $P(n)$ for fricatives which are not representative for the signal power.

For the estimate of the noise power, the equivalent value $x_{\min}$ is utilized (s. Sec. 3) which is also based on $x_{hp}(n)$.

With these values we determine the adaptive threshold factor according to

$$b(n) = b_{min} + u \cdot \frac{x_{\max}(nM)}{x_{\min}(nM)} \tag{15}$$

and limit the factor $b(n)$, which is also subsampled, to a range between b_{min}=2 and b_{max}=10. By applying heuristic optimization which is based on many speech signals, we found an optimal value $u = 0.06$.

A test with instationary noise finally confirmed the good choice of $b(n)$: The conditional probabilities P_d and P_f, we obtain with the adaptive factor $b(n)$, are located above the ROC graph in the P_d-P_f diagram. The reason is that every P_d-P_f couple of the ROC graph is determined with a fixed factor. As the noise power of the instationary noise varies, only an adaptive factor $b(n)$ can guarantee that the detector is working in the optimal operation point all the time.

6 RESULTS AND CONCLUSIONS

To evaluate the potential of the proposed VAD, its receiver operating characteristic is depicted in Fig. 4 in comparison to the classical VAD (the signal $ar(n)$) and the GSM detector [2]. Additionally, the operating point chosen by the GSM algorithm is marked. Hereby, it is obvious that the GSM detector is optimized for a low false alarm probability. We can also observe that the potential of the GSM detector is higher than for the classical VAD. However, our proposed algorithm outperforms the others significantly.

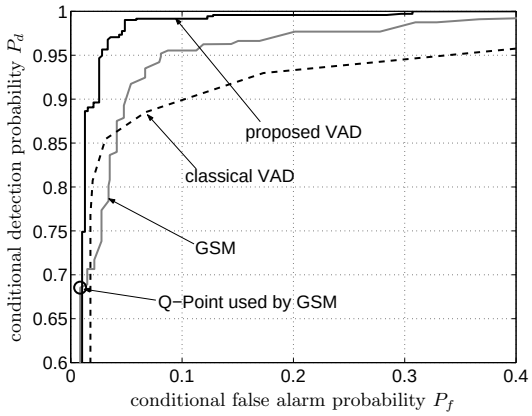

Figure 4: Comparison of the ROCs of the classical (dashed), the GSM (grey), and our proposed VAD (black)

With the comparisons of above's ROCs, the general potential of the different VADs could be evaluated. In the following we will show that also the adaptation of the factor $b(n)$ corresponding to the choice of the optimal operating point is working well. In Fig. 5 different ROCs are shown corresponding to different noise levels. The average factors $b(n)$ and the operating points are marked (◯) and prove to be almost optimal.

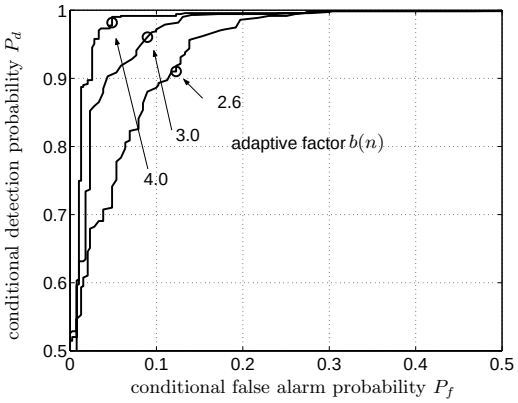

Figure 5: ROCs of the proposed VAD algorithm for signals with different SNRs. The chosen operation points and the corresponding mean values of the adaptive threshold factors $b(n)$ are marked.

We finish the paper with some concluding remarks: A voice activity detector was presented based on the prediction error power which is utilized as reference signal and allows an optimal distinction of noise and speech. In combination with additional processing units, e.g., the Hangover algorithm and the adaptive choice of the threshold factor $b(n)$ an algorithm is developed which outperforms other solutions known so far.

References

[1] ETS 300 961: Digital cellular telecommunications system (Phase 2+); Full rate speech; Transcoding (GSM 06.10 version 5.1.1), 2nd Edition

[2] ETS 300 580-6: Digital cellular telecommunications system (Phase 2); Full rate speech; Part6: Voice Activity Detection (VAD) for full rate speech traffic channels (GSM 06.32 version 4.3.1), 4th Edition

[3] R. Martin: *Spectral Subtraction Based on Minimum Statistics,* in Proc. Seventh European Signal Processing Conference, pp. 1182-1185 (1994)