On Noise Robust Voice Activity Detection
Tomas Dekens and Werner Verhelst
Interdisciplinary Institute for Broadband Technology – IBBT, Vrije Universiteit Brussel, department ETRO-DSSP, Belgium
tdekens@etro.vub.ac.be, wverhelst@etro.vub.ac.be

Abstract
In this paper, we show that the performance of voice activity detection algorithms (VAD) can be highly dependent on the type of background noise and we introduce a new VAD algorithm that is based on relative energy measurements in different frequency bands. The obtained experimental results are compared to the results obtained with two other spectrum-based VADs and it is concluded that a VAD, configured to use around 3 frequency bands can cope best with a large variety of background sounds.

Index Terms: voice activity detection, speech and audio segmentation and classification

1. Introduction
A voice activity detector is a sound classification unit that expects a noisy speech signal as input and then decides which parts of this signal contain speech and which parts don’t. This information could be of use in several speech processing applications. E.g., in speech coding or automatic speech recognition, where its goal is to only retain frames that contain speech. In noise suppression it is important to know which parts of the signal contain no speech at all, as these parts can be used to estimate the noise characteristics, which are needed if the noise has to be filtered out of the signal.

Our previous studies involving VAD led us to believe that most VAD algorithms described in the literature are greatly affected by the type of background noise they have to deal with. This should not come as a surprise, as a great deal of these algorithms relies on the dissimilarity between the noise sounds and speech to make the VAD decision. This conjecture is the motivation for the experiments we describe in this paper. We will examine some noise types that exhibit certain characteristics, which hamper the VAD process each in its own way. We propose an energy based VAD algorithm that we configure in 3 different ways, such as to determine the influence of the considered noise types on the different VAD mechanisms.

In section 2 of this paper we describe the energy based VAD algorithm and explain the different configurations we used. In section 3 we elaborate on the experiments that were conducted and we show some test results. Finally, in section 4 some conclusions are drawn.

2. Energy based VAD
We developed an energy based voice activity detector, that we named eVAD. The feature used in the eVAD algorithm is the smoothed energy, contained in the frequency region of interest. Let $Y(m,k)$ be the short-time Fourier transform (STFT) of the input signal $y(t)$, with $m$ the frame number and $k$ the frequency index. The smoothed energy is then calculated as:

$$E(m,\hat{k}) = \text{mean} \left\{ \frac{2}{N_{fft}} \sum_{j=0}^{N_{fft}/2} Y'(m+j,\hat{k}) \right\}_{j=1}^{N_{fft}}$$  \hspace{1cm} (1)

$$0 \leq \hat{k}, k \leq \frac{N_{fft}}{2}$$

Where $Y(m,k)$ is the same as $Y(m,\hat{k})$, except when $k$ corresponds to DC ($k=0$) or half the sampling frequency ($k=N_{fft}/2$), then $Y(m,k) = Y(m,\hat{k})/\sqrt{2}$ to ensure that the energy at these frequency bins is only counted once. One can see the parameter $N$ in (1) can control the extent to which the feature is smoothed. The index $\hat{k}$ represents the frequency range $[k_1,k_2]$ that is used to calculate the energy $E$.

During an initialization phase the first frames of the input signal are used to calculate the noise energy in the considered frequency region using (1); the mean of these noise frame energies gives us the initial estimation of the smoothed noise energy $E_{\text{Noise}}(m,\hat{k})$. Note that in our experiments we ensured that each signal does indeed start with a noise-only part. Next, the smoothed energy is calculated for each input frame $m$. This energy is then divided by $E_{\text{Noise}}$ and the logarithm is taken:

$$E_{\text{Ratio}}(m,\hat{k}) = 10 \log_{10} \frac{E(m,\hat{k})}{E_{\text{Noise}}(m,\hat{k})}$$  \hspace{1cm} (2)

This ratio is an indication of the difference the signal and the expected background noise exhibit in terms of energy content in the considered frequency region, at the current time instant $m$. A large difference indicates that, besides noise, other signals are present and could be a speech activity cue. In the next paragraphs we will describe how we use the energy ratios (2) corresponding to certain frequency ranges $\hat{k}$ in different configurations of eVAD.

2.1. Configuration 1: One band
The first configuration of eVAD uses one frequency band $\hat{k}$ to calculate the $E_{\text{Ratio}}$ according to (2). If it is known that speech will only cover a fraction of the full signal frequency range or if the noise energy can be expected to be small compared to the signal energy in a certain frequency band, this frequency region can be selected. However, in our experiments we will assume the noise characteristics are unknown and the full frequency band, i.e., $k_1=0$, $k_2= N_{fft}/2$, will be used.

The calculated energy ratio $E_{\text{Ratio}}$ is compared to a certain threshold. When the ratio is smaller than this threshold the frame is considered to contain only noise and the noise energy is updated:

$$E_{\text{Noise}}(m+1,\hat{k}) = \alpha E_{\text{Noise}}(m,\hat{k}) + (1-\alpha)E(m,\hat{k})$$  \hspace{1cm} (3)

$$0 \leq \alpha \leq 1$$

If the ratio is larger than the threshold, speech is detected and the current noise energy estimate will be kept:
According to the expected SNR of the input signals, an appropriate threshold value can be selected. The algorithm, however, also gives the possibility to work with a self-adaptive threshold. For this an on-line SNR estimation is performed. This SNR estimation requires a speech (actually speech + noise) energy estimate \( E_{\text{signal}}(m) \). If speech is detected in a certain frame, this estimate is updated in a similar fashion as the noise energy estimate:

\[
E_{\text{signal}}(m+1) = \alpha E_{\text{signal}}(m,k) + (1-\alpha)E(m,k) \tag{5}
\]

If a frame is classified as a noise only frame, the estimate is left unchanged. A certain percentage of the logarithm of the ratio

\[
E_{\text{signal}}(m,k)/E_{\text{noise}}(m,k) \tag{6}
\]

is then used as a threshold value.

Low energy phonemes such as /s/ or /l/ are hard to detect with an energy based VAD. If these phonemes occur in the middle of a speech fragment this doesn’t pose a problem. The surrounding phonemes will ensure the energy feature curve won’t drop to too low values. One can also select a minimum duration that a detected pause should last before it is actually classified as a pause. This can eliminate pauses that would be detected due to a short unwanted drop in the energy curve. On the other hand, very often a speech fragment starts or ends with such a low energy phoneme and this can not be dealt with by selecting a minimum pause length. That is why a detected speech portion is extended in time (front and back) by a certain amount, i.e. the regions before and after a detected speech fragment will be classified as speech regions as well. Because it is rather uncertain whether these regions are speech or not, they are not used to update the estimated noise or signal energies. To cope with short bursts of high energy like e.g. clicks, a minimum speech length can also be selected. If a speech region is detected that is shorter than this minimum length, it is classified as noise, but it is not used to update the noise energy.

Besides the relative energy ratio (2) it is also useful to use an absolute power measure. This can make the VAD deaf to signals whose power is below a certain value. So if

\[
\frac{1}{N_{\text{win}}} \sum_{n=0}^{N_{\text{win}}-1} |\text{win}(n) y(n + Sm)|^2 < \delta \tag{7}
\]

doesn’t contain speech, the current estimates are kept unchanged. If the current frame is classified as a noise frame, the different noise energies \( E_{\text{noise}}(m,k), k=k_1, \ldots, k_2 \) are updated using (3). In the other case the current estimates are left unchanged. (7) can be used to avoid speech detections if the absolute power of a frame is too low. We will call this configuration eVAD(W bands/mean).

Note that this configuration would lead exactly to the LTSD VAD if the maximum would have been used instead of the mean in (8) and if the logarithm would be taken after computing the mean in (10).

### 2.3. Configuration 3: W bands

The approach taken in the third configuration lays somewhere between those taken in the first two configurations. Here the useful frequency region is subdivided into \( W \) frequency bands \( k_1, k_2, \ldots, k_W \). We discriminate two ways to employ the \( W \) corresponding energy ratios \( E_{\text{ratio}}(m,k) \) that will be calculated as indications of speech activity.

#### 2.3.1. L detections

The first method of speech detection compares each of the \( W \) energy ratios to a certain threshold. If a number of ratios \( L \) or more are higher than their corresponding threshold, the frame is classified as a frame that contains speech. If the number of energy ratios that exceed their threshold value is lower than \( L \), the frame is considered to contain only noise and the \( W \) different noise energies \( E_{\text{noise}}(m,k), k=k_1, \ldots, k_W \) are updated using (3). If self-adaptive thresholds are desired, (6) can be used to calculate in each frame the threshold used in each frequency band. We will refer to this method as eVAD(W bands/L detections).

#### 2.3.2. Mean

The second method entails calculating the mean of the \( W \) different energy ratios:

\[
E_{\text{ratio}_{\text{mean}}} = \frac{1}{W} \sum_{k=k_1}^{k_W} E_{\text{ratio}}(m,k) \tag{8}
\]

This value is compared to one threshold. Again, if the frame is classified as a noise-only frame, the \( W \) different noise energies \( E_{\text{noise}}(m,k), k=k_1, k_2, \ldots, k_W \) are updated using (3). If the frame is expected to contain speech, the current estimates are kept (4). We will call this method eVAD(W bands/mean).

In both methods (7) will still be used to overrule speech detections if the absolute power of a frame is below a certain value.
3. Experiments

In order to evaluate the different configurations of the eVAD algorithm, we used the algorithm to detect pauses in the Dutch subset of a test database [1]. During the recording of this database four different (Dutch) speakers uttered 8 different sentences, which resulted in 32 sound files. The mean duration of the utterances is 37 seconds. On average, 74.59% of such an utterance contains speech; the other 25.41% consists of pauses. The sampling frequency of the sound files is 16kHz. The speech files were manually labeled in order to have a reference for the speech-pause detection. Only pauses of 200ms or longer were labeled as a pause, since it is not our intention to detect e.g. inter-word pauses [1].

Besides the eVAD algorithms, the LSTD algorithm [2] and the SLR algorithm [3] were also evaluated in the experiments. During the tests, different kinds of noises were added to the clean speech at different SNRs. Subsequently, the three VAD algorithms were applied on the 32 utterances using 32ms long speech frames with 50% overlap and different threshold values. For each threshold value the speech detection probability (SDP) was calculated by dividing the number of frames that were correctly classified as speech by the total number of speech frames, and the false alarm probability (FAP) was calculated as the number of pause frames that were classified as speech frames divided by the total number of pause frames. Note that the SDP and FAP are exactly the same as, respectively, the sensitivity and one minus the specificity of the binary classification.

The smoothing parameter $N$ in (1) was for every configuration set to 6. This leads to a low-pass smoothing, with a rise and fall time of about 160ms. The weight parameter $\alpha$ in (3) was set to 0.95. Since we want to get an impression of how well the features used in the different algorithms and configurations are a representation of speech presence or absence, we used no extension of detected speech fragments, nor any absolute power or speech duration constraints.

In configuration 1 and 2 we used the full frequency band, meaning that in the case of configuration 2, we work with $\sqrt{\text{filter}} / 2 + 1$ different bands. In configuration 3, three frequency bands are utilized: (0-500) Hz, (500-2000) Hz and (2000-8000) Hz. In the case where each of the three energy ratios is compared to a threshold, the same threshold was used for the three different bands. This was done in order not to overcomplicate the SDP-FAP relation and as a result make it possible to compare the results to those obtained with the other methods. Note, however, that the best performance can probably be seen when the threshold values are not identical. The amount of ratios $L$ that need to be higher than their threshold was set to 1 (leading to the name eVAD(3bands/1detection) seen in Figures 1 and 2).

Some results of the experiments can be seen in Figures 1 and 2. Each point on the curves displayed in these figures shows the FAP and SDP values that were obtained in the case of a certain noise type and with a certain threshold value. An obvious noise property that can upset most VAD mechanisms is non-stationarity. For this reason we used a non-stationary noise type we called multi-noise. This is nothing more than noise whose characteristics change drastically every 5 seconds and was produced by repeatedly concatenating 5 second fragments from 5 different normalized noise files (pink noise, jet fighter noise, tank noise, factory noise, car noise). It is clear that VADs that compare the spectrum of the signal to that of the background noise will suffer most from sudden

![Figure 1: Speech detection probability vs False alarm probability. Left: 15dB SNR multi-noise, right: 0dB SNR babble noise](image1)

![Figure 2: Speech detection probability vs False alarm probability. Left:-10dB SNR tank noise, right:-10dB SNR white noise](image2)
changes in the background noise spectrum. The background noise energy however stays more or less constant, which implies that a VAD that analyzes the global energy should still be able to attain good classification scores, as long as the overall SNR remains high enough (we kept the SNR stable at 15dB in our experiments). This hypothesis is proven by what can be seen in the left hand panel of Figure 1, where the results are displayed for multi-noise at a 15dB SNR. The eVAD(1 band) attains a high SDP at low FAP values. LTSD, SLR and eVAD(all bins) are three methods that rely strongly on changes in the spectral curve; that is why for these methods a high FAP needs to be tolerated if high SDP values are desired. The two eVAD(3 bands) methods possess a spectral sensitivity that lays between those of the other two families of methods. When noise with changing characteristics is present, the energy of the noise changes in every band but not to the same extent as the energy at the frequency bin level. As expected from all this, the SDP-FAP curves of the eVAD(3 bands) methods indicate the second best performances.

Most VAD algorithms do not struggle when stationary noise types are encountered, as long as the SNR stays within an acceptable range. When the SNR becomes too low, problems can arise. To investigate this, we added two high power stationary noise types to the clean speech signals: one where the noise energy is spread over the whole frequency range (white noise), and one where the noise energy content stays confined to a relatively narrow band (tank noise). The right hand panel of Figure 2 shows how high power tank noise (SNR = -10dB) affects the VAD algorithms. In this case the spectrum based VADs (LTSD, SLR, eVAD(all bins)) do a better job in distinguishing speech from noise. This can be explained as follows: tank noise only covers the lower end part of the spectrum (0-1500Hz). The rest of the spectrum is unaffected by this type of noise. This means that when speech starts, overall the spectrum still changes to a large extent. The energy however doesn’t, because of the high amount of energy inserted in the low part of the spectrum by the added noise. Again, the two eVAD(3 bands) methods seem to lead to the second best results. In this case, however, these results are certainly not as awful as the worst results that are obtained by eVAD(1 band). Although the global energy of the signal does not change much when speech starts, the energy in one of the bands used in eVAD(3 bands) (the highest one in this case) is altered to a large extent when speech emerges in the signal, leading to correct speech detections.

The left hand panel of Figure 2 shows the SDP-FAP curves that are obtained when high power white noise (SNR = -10dB) is added to the speech. It is obvious that looking at the global energy is not a good idea, as the change in energy when speech starts will almost not be noticeable. In this case however, looking at the global spectral shape is not a good idea either. The spectrum of the signal will be more or less flat at all times; when speech starts its energy at higher frequencies is so low compared to that of the noise that no spectral changes can be seen. In the very low frequency part of the spectrum (0-500 Hz) on the other hand some peaks can be seen that emerge from the flat spectrum. This is not enough to influence the global spectral shape features used in these algorithms much, hence their mediocre performances. This is, on the other hand, a sufficient change for the eVAD(3 bands) methods, since the energy content modification in the lowest band is relatively high. That is why we see much better results when these algorithms are employed.

The last noise type we consider is babble noise. Since the spectrum of babble noise is somewhat comparable to that of speech, the spectrum is altered to a lesser degree when speech starts than if a noise type with the same power but a completely non-speech like spectrum were present. This affects all spectrum based VADs. SLR, LTSD and eVAD(all bins) however suffer more from the non-stationarity of the noise. eVAD(3bands) and eVAD(1 band) are slightly less disturbed. This is reflected in the curves that are displayed in the right hand panel of Figure 1, which are obtained when babble noise is added at an SNR of 0dB.

Of note should be that selecting a more appropriate band instead of the full available band when using eVAD(1 band), LTSD, SLR or eVAD(all bins) could lead to much better results (e.g., by not using the low frequency part in eVAD(1 band) when tank noise is present, or by using exactly only the low frequency part in eVAD(1 band), LTSD, SLR or eVAD(all bins) when high level white noise is present). However, this demands that the noise characteristics are known beforehand, which generally speaking is not the case.

To summarize, methods such as LTSD, SLR or eVAD(all bins) are very sensitive to global spectral changes and are useful when high energy noise is present but highly concentrated around certain frequencies; they fail when the noise characteristics change over time. Looking at the global energy range. eVAD(1 band)) is a well suited technique when the (low energy) background noise has variable spectral properties, but inappropriate for use with frequency concentrated noise. eVAD(3 bands) certainly is the best option when a flat spectrum noise such as white noise is encountered. Also in the other cases we investigated, using 3 energy bands leads to acceptable results.

4. Conclusions

In this paper we described an energy based voice activity detection algorithm. We showed that certain configurations of this VAD can outperform conventional spectrum based VADs under certain noise conditions. A configuration where 3 frequency bands are used generally leads to the most satisfactory results when the noise characteristics are unknown. When the noise properties are known, the configurability of the described algorithm allows for the selection of an appropriate VAD mechanism.

5. Acknowledgements

Parts of the research reported on in this paper were performed in the context of the CATRIS (TBM-80662) and AMASS++ (SBO-060051) projects, that are both supported by the Flemish government agency for Innovation by Science and Technology (IWT), and EU-FP7 project ALIZ-E (ICT-248116).

6. References


