VOICE ACTIVITY DETECTION IN THE DFT DOMAIN BASED ON A PARAMETRIC NOISE MODEL

Colin Breithaupt and Rainer Martin

{colin.breithaupt,rainer.martin}@rub.de
Ruhr-University of Bochum, Institute of Communications Acoustics (IKA)
44780 Bochum, Germany

ABSTRACT

We present a robust voice activity detection (VAD) algorithm which is based on the statistics of the coefficients of the discrete Fourier transform (DFT) derived from short signal segments. This algorithm uses a common parametric noise probability density function (PDF) in all frequency bins. The noise model is based on a Rayleigh inverse Gaussian distribution and adapted to the statistics of the noise during speech-absence. As only the current and past signal frames are analysed, the detection is causal and no additional delay is introduced. A framework for protecting low energy syllables at the end of utterances is also described.

1. INTRODUCTION

The detection of speech presence is crucial for many algorithms for speech enhancement or speech recognition. For framewise processing these algorithms can be optimised if knowledge of speech presence is available. Using the VAD decisions these algorithms can use different strategies during speech activity and speech pauses. It is, however, not a trivial task to design a VAD algorithm such that it will operate reliably in high levels fo noise, especially when the noise PDF is non-Gaussian and not known a priori. There exist different ways to increase the robustness of the VAD in the presence of noise which will be discussed below.

One well known solution is a likelihood ratio test [1, 2] in the spectral domain. As this test needs a statistical model for the sum of the spectral coefficients of speech and noise, it is not straightforward to incorporate non-Gaussian models for speech or noise. However, the assumption of a Gaussian noise PDF leads to a high rate of misclassification of non-speech frames in case of babble noise or similar non-Gaussian noises.

[3] uses a noise model only and does not rely on the speech statistics. As a Gaussian model is used for the noise, babble noise is still difficult to handle with that algorithm. In [4] the order statistics of subband energies for several consecutive frames is considered. This approach does not make explicit use of a probability density model.

In this paper we derive a VAD algorithm which relies only on a statistical model of non-speech frames as in [3]. A parametric PDF is used to adapt this model to the noise. The resulting detection criterion is used in conjunction with a novel criterion based on frame energies. Like in [2, 3] we also introduce a finite state machine that helps to protect low energy syllables at the end of words. In [4] this kind of protection can be avoided as the detector is non-causal.

The rest of this paper is organised as follows. In Section 2 an overview of the framewise signal processing in the spectral domain is given. Section 3 introduces the two criteria used for speech detection in noisy environments. The framework for protecting low energy syllables and reducing the false alarm rate due to single frame detection errors is described in Section 4. Experimental results are given in Section 5.

2. SIGNAL PROCESSING IN THE DFT DOMAIN

The observed noisy speech signal $y(t)$, where $t \in \mathbb{Z}$ is the discrete time index, is segmented into frames of length K with a frame shift $L = K/2$ and weighted by a Hann window $h_{hann}(t)$. The weighted frames are transformed by the DFT resulting in the observed spectrum

$$\begin{aligned} Y(k,l) &= \sum_{\tau=0}^{K-1} h_{hann}(\tau)\; y(lL+\tau)\; \mathrm{e}^{-j2\pi k\tau/K} \\ &= S(k,l) + N(k,l)\,, \end{aligned} \tag{1}$$

The DFT coefficients $Y(k,l) = S(k,l) + N(k,l)$ of the noisy signal in frame $l \in \mathbb{Z}$ and frequency bin $k = 0 \ldots K-1$ are assumed to be the sum of the clean speech coefficients $S(k,l)$ and the noise coefficients $N(k,l)$. For the sake of simplicity we will leave out the frequency and frame indices, k and l, whenever possible.

For the detection of speech frames we need an estimate $\widehat{S}$ of the clean speech spectral coefficients given the noisy observation Y. This estimate is calculated by the Wiener filter rule

$$\widehat{S} = \frac{\widehat{\xi}}{1+\widehat{\xi}}\, Y = G(\widehat{\xi}) \cdot Y, \tag{2}$$

where G denotes the Wiener filter gain and estimated quantities such as the estimated *a priori* SNR $\widehat{\xi}$ are marked with the hat symbol.

The true *a priori* SNR ξ is defined as $\xi = P_s/P_n$, where $P_s(k) = E\left\{|S(k)|^2\right\}$ and $P_n(k) = E\left\{|N(k)|^2\right\}$ are the speech power and the noise power in frequency bin k, respectively. We use the *decision-directed* approach [1] to obtain an estimate $\widehat{\xi}$ of the *a priori* SNR. The estimate $\widehat{P}_n$ of the noise power is calculated as the empirical mean by recursive averaging during speech pauses. The mean of $|N|^2$ is an estimator of the noise power P_n that is independent of the statistical distribution of $|N|^2$ and is therefore suitable for different noise types.

3. SPEECH DETECTION CRITERIA

3.1. Outlier count

We now describe a speech detection criterion that solely relies on knowledge of the statistical behaviour of the *a posteriori* SNR $\gamma(k) = |Y(k)|^2/\widehat{P}_n(k)$ during speech absence. For each frequency bin k, the observation of $\gamma(k)$ is treated as a random variable. These random variables are assumed to be independent, identically distributed. An observation $\gamma(k) \geq \gamma_{th}$ is called an outlier, whereby a threshold $\gamma_{th} > E\{\gamma\} = 1$ (speech absence) is chosen. If the number of outliers in a frame exceeds a threshold, that frame is very unlikely to contain noise only and the presence of speech is assumed.
For a frame l we count the number of outliers in a subset of frequency bins as

$$n(\gamma_{th}) = \sum_{k\in\mathbb{K}} \left[\gamma(k) \geq \gamma_{th}\right], \tag{3}$$

whereby the operator $[a \geq b]$ returns 1, if the statement is true, otherwise it gives 0. Note that $n(\gamma_{th})$ is a non-negative integer. We regard only that subset

$$\mathbb{K} = \{k_l, k_l + \Delta k, \ldots, k_h - \Delta k, k_h\}$$

of frequency bins where γ significantly changes in the presence of speech. The lower and upper limit are chosen to give $k_l f_s/K \approx 200$Hz and $k_h f_s/K \approx 3.5$kHz, whereby f_s is the sampling frequency. As the bins need to be statistically independent they are chosen at a distance $\Delta k = 3$, which corresponds to the spectral width of the main lobe of the windowing function $h_{hann}(\tau)$.
In the following, we derive the probability of observing a certain number of outliers in a frame that contains noise only. For a single frequency bin k the probability $\mathrm{P}(\gamma_{th})$ to observe $\gamma(k) \leq \gamma_{th}$ is given as

$$\mathrm{P}(\gamma_{th}) = \Pr\{\gamma < \gamma_{th}\} = \int_0^{\gamma_{th}} p\left(\gamma\right) d\gamma. \tag{4}$$

The PDF $p\left(\gamma\right)$ describes the statistical behaviour of γ in the case of speech absence and thus depends on the current noise type. We use a parametric PDF that is adapted at the beginning of each utterance as explained in Section 3.2. If the number of indices contained in $\mathbb{K}$ is $\mathrm{N} = |\mathbb{K}|$, the probability $\Pr\{n(\gamma_{th}) \geq n_0\}$ of observing n_0 or more outliers in a frame in case of speech absence is

$$\Pr\{n(\gamma_{th}) \geq n_0\} = \sum_{n=n_0}^{\infty} \binom{\mathrm{N}}{n} \left(1 - \mathrm{P}(\gamma_{th})\right)^n \left(\mathrm{P}(\gamma_{th})\right)^{\mathrm{N}-n}, \tag{5}$$

if the random variables $\gamma(k)$ are statistically independent for different k.
For speech detection we define the maximum probability P_{th} of observing n_0 or more outliers in a non-speech frame. From this the threshold n_0 follows to be the smallest value that gives

$$\Pr\{n(\gamma_{th}) \geq n_0\} \leq P_{th}. \tag{6}$$

We decide for speech presence ($\mathcal{H}_{1a}$) in the current frame in the case of $n(\gamma_{th}) \geq n_0$, otherwise speech absence ($\mathcal{H}_{0a}$) is assumed. The first criterion $\mathrm{H}_a(l)$ for speech detection in frame l is

$$\mathrm{H}_a(l) = \begin{cases} \mathcal{H}_{1a} & \text{if } n(\gamma_{th}) \geq n_0 \\ \mathcal{H}_{0a} & \text{else} \end{cases} \tag{7}$$

The parameter P_{th} thus sets the false alarm rate.

3.2. Statistical model $p\left(\gamma\right)$

In order to model the statistical behaviour of γ in case of speech absence a parametrical PDF $p\left(\gamma\right)$ is adapted within the first K_0 frames of an utterance. For the PDF the Rayleigh-Inverse-Gaussian as introduced in [5] is used. It is given as

$$p\left(\gamma\right) = \sqrt{\frac{2}{\pi}}\, \alpha^{3/2}\, \delta \exp(\delta|\alpha|) \times \frac{\gamma}{(\delta^2+\gamma^2)^{3/4}} \mathrm{K}_{3/2}(\alpha\sqrt{\delta^2+\gamma^2}). \tag{8}$$

$\mathrm{K}_{3/2}(\cdot)$ is the modified Bessel function of the second kind. The shape parameter α determines the heavy-tailedness of the distribution and is used here to model the different noise types. The scale parameter δ is determined by the variance of γ. α and δ are estimated by the expectation maximisation algorithm given in [5]. The $\mathrm{N} \cdot K_0$ values $\gamma(k)$ from the frequency bins $k \in \mathbb{K}$ of the first K_0 frames are used for this estimate.

3.3. Energy of filtered frames

In the following a second criterion for speech detection is introduced. Whenever the SNR is small, low energy spectral components of speech are increasingly covered by noise. In this case, the detection based on the criterion in section 3.1 becomes unreliable. In order to check if speech has been missed in the last frame, the energy of

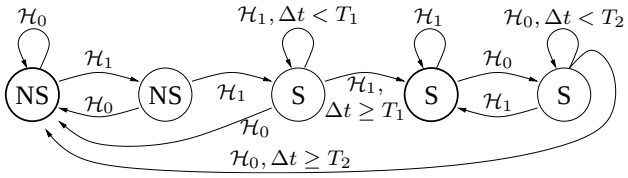

Figure 1: The automaton changes states depending on the framewise decision $\mathrm{H}(l)$ from equation (12). The labels of the states show, whether the current frame is classified as speech (S) or non-speech (NS). The time Δt that passed since a state has been entered is also considered.

the preceeding filtered frame $\mathcal{E}_f(l-1)$ is compared to the expected frame energy $\mathcal{E}_n(l-1)$ in case of speech absence. The frame energies are computed as

$$\mathcal{E}_f(l-1) = \sum_{k\in\mathbb{K}} |G(k,l-1)\ Y(k,l-1)\,|^2, \quad (9)$$

$$\mathcal{E}_n(l-1) = \sum_{k\in\mathbb{K}} |G(k,l-1)\,|^2\ \widehat{P}_n(k,l-1)\,. \quad (10)$$

Note that these quantities do not reflect true frame energies as only the subset $\mathbb{K}$ is used. As speech energy is concentrated in bins below k_h, the detection becomes easier, if frequencies above k_h are not considered.
The gain function G emphasises frequency bins with high SNR in the preceeding frames. These bins are likely to contain speech in frame $l-1$. Thus, by using the gain function G in equations (9) and (10), the sensitivity to speech components is increased.
We can now define the second detection criterion as

$$\mathrm{H}_b(l) = \begin{cases} \mathcal{H}_{1b} & \text{if } \mathcal{E}_f(l-1) > \beta\,\mathcal{E}_n(l-1) \\ \mathcal{H}_{0b} & \text{else} \end{cases}. \quad (11)$$

The factor $\beta > 1$ depends on the shape parameter α in order to reflect the variability of frame energies of different noise types.

3.4. Overall framewise decision

The two decisions (7) and (11) are combined to give the framewise decision. As the two detection criteria cover two different types of speech sounds, they are combined as

$$\mathrm{H}(l) = \begin{cases} \mathcal{H}_1 & \text{if } \mathrm{H}_a(l) = \mathcal{H}_{1a} \text{ OR } \mathrm{H}_b(l) = \mathcal{H}_{1b} \\ \mathcal{H}_0 & \text{else} \end{cases} \quad (12)$$

As the outputs $\mathrm{H}(l)$ for different frames are not linked, the results can change several times within a short time span in cases of low SNR. In order to adapt the rate of change in the output of the detector to time constants similar to those of natural speech, $\mathrm{H}(l)$ is embedded in a framework that is introduced in Section 4.

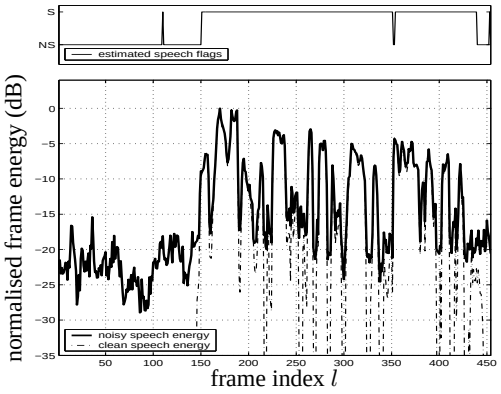

Figure 2: Frame energies of the noisy signal (solid thick line) and of the clean signal contained therein (dash-dotted line). The noise signal is babble noise at a segmental SNR of 5dB. The energies are normalised to the maximum frame energy within the utterance and displayed on a logarithmic scale. Above, the output of the finite state machine is shown for corresponding frames.

4. VAD FRAMEWORK

The detection mechanism described in Section 3 decides on a frame-by-frame basis. In [2] and [3] these decisions are embedded in a framework in order to prevent clipping of low energy syllables at the end of words that are not discovered by the frame-wise detection. This protection is generally achieved by a grace period that has to pass before a change in classification from speech frames to non-speech frames is made. In Figure 1 a finite state machine is shown that relates the framewise decisions $\mathrm{H}(l)$ to finally classify a frame as speech (S) or non-speech (NS).

In periods of speech absence the automaton is in the leftmost state of Figure 1. A single frame that is classified as speech ($\mathrm{H} = \mathcal{H}_1$) does not cause the automaton to signal speech presence. This compensates for single frame classification errors and reduces the false alarm rate effectively.

The grace period T_2 is only applied after continuous speech presence has been signaled by $\mathrm{H}(l)$ for a minimum duration T_1. If several successive non-speech frames are misclassified as speech, the error would be worsened, if the grace period would be applied immediately.

Figure 2 gives an example of the automaton's output for a noisy signal. The noise is highly non-stationary babble noise at an segmental SNR of 5dB. In frames 150 and 350 the delayed change from output "NS" to "S" is visible. For applications that do not depend on a causal processing a one frame look-ahead would reduce this delay. At frames 340 to 350 and 430 to 440 the speech energy is too low to result in frames that can be distinguished from non-speech frames. Here the detector profits from the grace period.

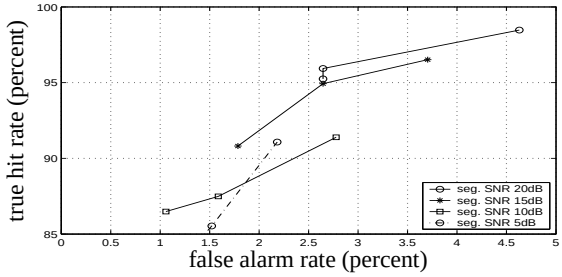

Figure 3: Percentage of speech frames correctly classified (true hit rate) against percentage of non-speech frames misclassified as speech (false alarm rate) for pink Gaussian noise at different segmental SNRs.

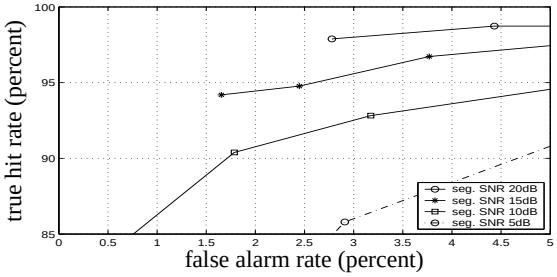

Figure 4: Percentage of speech frames correctly classified (true hit rate) against percentage of non-speech frames misclassified as speech (false alarm rate) for babble noise at different segmental SNRs.

5. EXPERIMENTAL RESULTS

For the experimental evaluation a set of 10 utterances from the TIMIT-database is used. Their sampling rate is $f_s = 16$kHz. The signals are processed in frames of length $K = 512$. Each clear utterance is preceeded by more than 2 seconds of silence in order to evaluate the detection in speech pauses. All frames that have a frame energy of at most -45dB below the maximum frame energy of the clear utterance are classified as speech. Speech pauses shorter than 200ms are also classified as speech as they belong to the so called *structural pauses* of speech. The noise signals are taken from the NOISEX92 database.

The integral resulting from (4) and (8) is solved numerically. For the estimation of the parameters α and δ the first $K_0 = 40$ frames of the noisy utterance are used. For the definition of an outlier we choose a threshold $\gamma_{th} = 4$. The threshold n_0 is found by calculating $\Pr\{n(\gamma_{th}) \geq n_0\}$ for increasing values of n_0 until the condition in equation (6) is violated. The calculation of n_0 in (6) is done for a range $P_{th} = 0.1\% \ldots 25\%$ to give different points of the receiver operator characteristics (ROC). The decision (11) based on the frame energies uses a constant $\beta = 7$. The grace period T_2 of the automaton is set to $T_2 = 2T_1 = 200$ms.

For the ROC in Figures 3 and 4 the false alarm rate is defined as the number of non-speech frames misclassified as speech divided by the total number of non-speech frames. The true hit rate is the number correctly detected speech frames divided by the total number of speech frames.

In Figure 3 the ROC is shown for pink noise at different segmental SNRs. As the detector depends on the three parameters P_{th}, β, and T_2, the curves do not cover a large range of possible false alarm rates, if only one of the parameters is changed. For the case of 20dB segmental SNR it can be seen that the false alarm rate cannot be reduced below 2.6% by varying P_{th}. For high SNR the grace period $T_2 = 200$ms covers non-speech frames at the end of the utterances. Making T_2 adaptive would make it possible to reduce the false alarm rate at high SNR.

The case of babble noise is shown in Figure 4. As the adaptive PDF $p(\gamma)$ reflects the changed statistics of this noise type, the false alarm rate increases only for high values of P_{th}. We found that $P_{th} = 2\%$ is a good compromise.

6. CONCLUSIONS

In this paper we have presented a VAD algorithm based on two criteria that relies on a statistical model of noise. Only the statistics of the noise is considered in the form of the *a posteriori* SNR in speech pauses. As a second detection criterion frame energies are considered. A finite state machine links the framewise decisions to compensate for single frame decision errors and speech clipping at the end of words.

In future work a reduction in the false alarm rate in case of high SNR is attempted by making the grace period T_2 adaptive. Using the shape paramter α the factor β in equation (11) can also be adapted to the variability of the noise.

This work is funded by the German Research Foundation **DFG**.

7. REFERENCES

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

[2] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, pp. 1–3, Jan. 1999.

[3] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 412–424, Mar. 2006.

[4] J. Ramírez, J. C. Segura, C. Benítez, Á. de la Torre, and A. Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1119–1129, Nov. 2005.

[5] T. Eltoft, "The Rician inverse gaussian distribution: A new model for non-Rayleigh signal amplitude statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1722–1735, Nov. 2005.