

Voice Activity Detection based on Inverse Normalized Noise Likelihood Estimation

Tomas Dekens⁽¹⁾, Mike Demol⁽¹⁾, Werner Verhelst⁽¹⁾ and Frédéric Beaugendre⁽²⁾

⁽¹⁾ Vrije Universiteit Brussel, dept. ETRO-DSSP, Pleinlaan 2, B-1050 Brussels, Belgium
{tdeken, midemol, wverhels}@etro.vub.ac.be <http://www.etro.vub.ac.be/Research/DSSP/dssp.htm>

⁽²⁾ Voice Insight nv, J. Wybranlaan 40, B-1070 Brussels, Belgium, frederic.beaugendre@voice-insight.com <http://www.voice-insight.com/>

Abstract—In this paper we develop a voice activity detection algorithm based on the likelihood that only noise is present in the current signal frame. For this we exploit the fact that the Fourier coefficients of most noise processes can be modeled as statistically independent Gaussian random variables. We also give an overview of different voice activity detectors previously described in the literature and compare their results to the ones obtained with the voice activity detector we propose in this paper. According to our tests, at high speech detection probabilities, the proposed algorithm shows results that are comparable to or better than the other voice activity detectors we consider, while the simplicity of the algorithm ensures low computational complexity.

Key Words—Noise estimation, Speech enhancement, Voice activity detection.

I. INTRODUCTION

Voice Activity Detection (VAD) is important in many different fields of speech processing. In a lot of speech applications it is essential to know whether the speaker is talking or not. Most of the time the goal is to only retain the frames that contain speech, e.g. for variable-rate speech coding or automatic speech recognition. In the case of noise suppression, however, the frames that contain no speech are used to estimate the noise spectrum, which is needed to suppress the noise in all frames. This is especially important in the case of time varying noise, where a VAD is used to update the noise in non-speech frames to keep track of its time varying characteristics. There also exist techniques that constantly update the noise spectrum [1], [2], [3], [4]. These algorithms are generally faster in keeping track of the noise spectrum, but usually some energy of the speech signal is also picked up in the noise estimate, which leads to a degradation of the enhanced signal [5]. So for most noise suppression applications it is better to use a VAD.

Already a lot of research has been done on VAD, and in this paper we examine some of the previously developed voice activity detectors [5], [6] and we propose a rather simple alternative to these algorithms, i.e. the Inverse Normalized Noise Likelihood (INNLL) VAD.

In the following sections II and III, we briefly describe the different VAD methods considered and we introduce the INNLL VAD. In section IV we describe the data we used for testing the voice activity detectors and we discuss the results. Finally, in section V some conclusions are drawn.

II. VAD METHODS CONSIDERED

A. Long-term spectral divergence

The Long-Term Spectral Divergence (LTSD) method [5] compares the Long Term Spectral Envelope (LTSE) of the signal to the estimated noise short time amplitude spectrum in order to decide whether speech is present in the current frame.

Let $Y(m,k)$ be the short-time Fourier transform (STFT) of a noisy signal $y(n)$, with m the frame number and k the frequency bin. The LTSE of order N is then defined as follows:

$$LTSE(m, k) = \max \left\{ |Y(m + j, k)| \right\}_{j=-N}^{j=+N} \quad (1)$$

The LTSD is defined as the deviation of the LTSE with respect to the estimated noise short-time (ST) magnitude spectrum $N_A(m,k)$:

$$LTSD(m) = 10 \log_{10} \left(\frac{1}{N_{fft}} \sum_{k=0}^{N_{fft}-1} \frac{LTSE^2(m, k)}{N_A^2(m, k)} \right) \quad (2)$$

where N_{fft} is the number of FFT points used.

This LTSD feature is then compared to a threshold to decide if speech is present in the current frame m . If no speech is detected, it is assumed frame m only consists of noise, thus we can update the noise spectrum estimate according to the following equation:

$$N_A(m + 1, k) = \alpha N_A(m, k) + (1 - \alpha) N_K(m, k) \quad (3)$$

where

$$N_K(m, k) = \frac{1}{2K+1} \sum_{j=-K}^K |Y(m+j, k)| \quad (4)$$

If speech is detected the current noise spectrum estimate will be kept:

$$N_A(m+1, k) = N_A(m, k) \quad (5)$$

B. Statistical likelihood ratio

The VAD algorithm proposed in [6] uses the Statistical Likelihood Ratio (SLR). Here it is assumed that the different frequency components of the STFT of a (short term) stationary speech process x_n and noise process d_n , denoted X_k and D_k respectively (we will sometimes omit the frame index from now on), are independent Gaussian random variables. Furthermore it is assumed that the two processes are uncorrelated such that we can write the SLR for each frequency component k as:

$$\Lambda_k = \frac{p(Y_k | H_1)}{p(Y_k | H_0)} = \frac{1}{1 + \xi_k} \exp\left\{ \frac{\gamma_k \xi_k}{1 + \xi_k} \right\} \quad (6)$$

where H_0 and H_1 are the hypotheses of speech absence and speech presence respectively, and $p(\cdot)$ represents the probability density function (PDF) of Y_k . $\xi_k = \lambda_x(k)/\lambda_N(k)$ and $\gamma_k = |Y_k|^2/\lambda_N(k)$ are called the *a priori* and *a posteriori* signal to noise ratio (SNR) respectively, where $\lambda_x(k) = E\left[|X_k|^2\right]$ and $\lambda_N(k) = E\left[|D_k|^2\right]$.

The logarithm of the global SLR can then be compared to a certain threshold η :

$$\log \Lambda = \frac{1}{N_{fft}} \sum_{k=0}^{N_{fft}-1} \log \Lambda_k \gg \eta \quad (7)$$

If no speech is detected the noise ST power spectrum estimate can be updated:

$$N(m+1, k) = \alpha N(m, k) + (1-\alpha) |Y(m, k)|^2 \quad (8)$$

In the case that speech is detected in the current frame, the estimate is left unchanged:

$$N(m+1, k) = N(m, k) \quad (9)$$

To calculate the *a posteriori* SNR γ_k we can use the STFT of the noisy signal $y(n)$ and the estimate of the ST noise power spectrum N_k . For the *a priori* SNR ξ_k we also need an estimate

of the speech ST power spectrum. To overcome this problem it was proposed in [6] to use the Decision-Directed (DD) *a priori* SNR estimation method [7] in combination with a noise suppression algorithm. The HMM based hangover scheme, described in [6], was also implemented.

III. INVERSE NORMALIZED NOISE LIKELIHOOD

The SLR algorithm requires knowledge of the speech power spectrum; this implies that it is necessary to execute some noise suppression on the noisy signal. Also, theoretically, the speech power spectrum used in Eq.6 should be the speech spectrum under speech presence. The DD *a priori* SNR estimator, however, estimates the *a priori* SNR independent of the fact of speech presence or absence. The method we introduce in this paper is based on the Statistical Likelihood Ratio method, but it does not require knowledge of the speech power spectrum. We call this method the Inverse Normalized Noise Likelihood VAD. Our method also relies on the fact that the different frequency components of the STFT of a noise process are independent Gaussian random variables. From this it follows that the components of the amplitude of this STFT follow a Rayleigh distribution. So in the case that speech is absent (H_0) we can write for the PDF of the amplitude of the k^{th} frequency component of the STFT $Y(k)$ of the noisy signal $y(n)$:

$$p(|Y(k)| | H_0) = \frac{2|Y(k)|}{\lambda_N(k)} \exp\left(-\frac{|Y(k)|^2}{\lambda_N(k)}\right) \quad (10)$$

We can use this PDF as a noise likelihood function. We will take the inverse of this PDF in order to get a value which increases with higher speech probabilities. Because in Eq.10 we will use an estimate of the noise ST power spectrum, which is updated frequently, in different frames for each frequency component a different shape of PDF will be used. This would be problematic for having a standard reference for the global likelihood. To solve this we normalize the PDF of each component so that its maximum w.r.t. $|Y(k)|$ is equal to one and afterwards take the inverse of the obtained function:

$$\sigma_k = \frac{1}{\text{Norm}(k) p(|Y(k)| | H_0)} \quad (11)$$

where

$$\text{Norm}(k) = \frac{1}{\frac{2}{\sqrt{2\lambda_N(k)}} \exp(-1/2)} \quad (12)$$

The logarithm of the global INNL feature is then compared to a threshold:

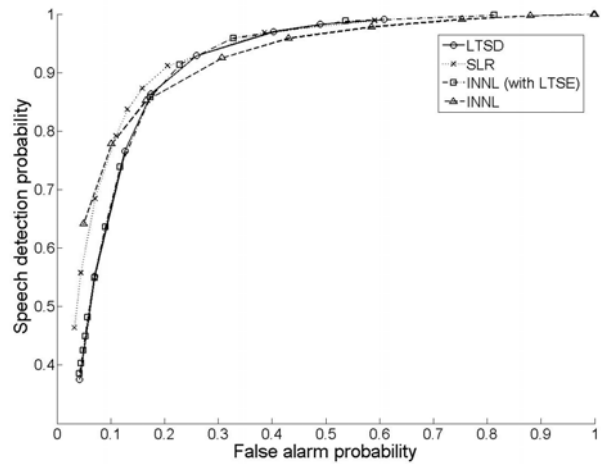
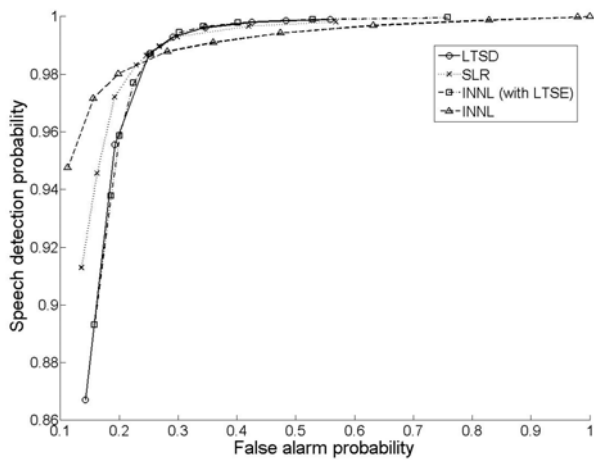


Fig. 1. Speech detection probability vs False alarm probability. (babble noise). SNR = 20dB (left), SNR = 10dB(right)

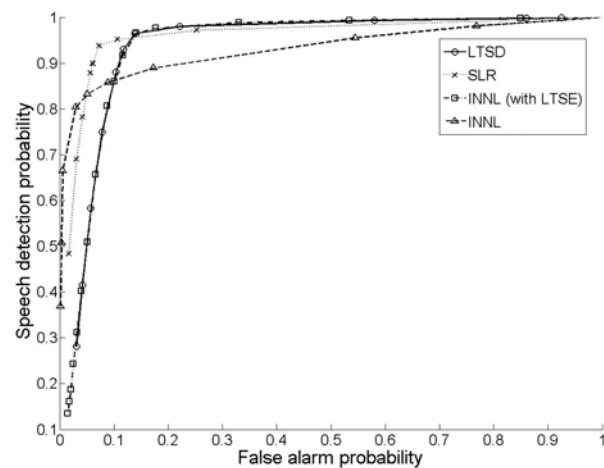
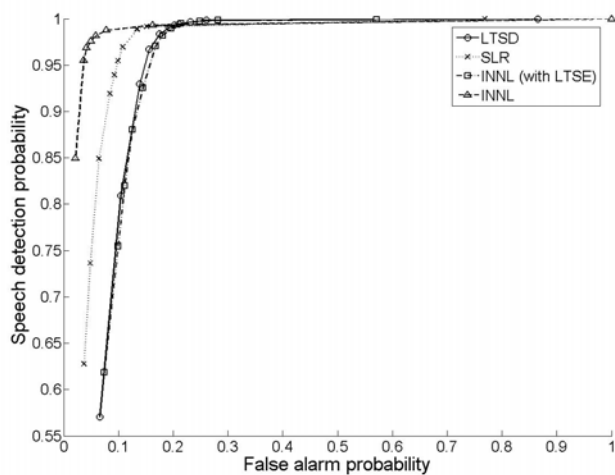


Fig. 2. Speech detection probability vs False alarm probability. (white noise) SNR = 20dB (left), SNR = 0dB(right)

$$\log \sigma = \frac{1}{Nfft} \sum_{k=0}^{Nfft-1} \log \sigma_k \ll \eta \quad (13)$$

In order to smooth the resulting INNL function we can use $\max\{|Y(m+j,k)|\}_{j=-N}^{j=+N}$ instead of $|Y(m,k)|$ in Eq.11, which similar to the LTSE, defined in [5].

Depending on the result of the VAD we can again update the estimate of the noise power spectrum, which we will use in Eq.11, according to Eq.8 and Eq.9.

IV. EXPERIMENTS AND RESULTS

A. Testdata

In order to test the different VAD algorithms, we applied them on the Dutch subset of a test database [8]. Four different speakers were asked to utter 8 different sentences in a fast, normal and slow fashion, which resulted in 96 sound files. The mean of the duration of the utterances is 28 seconds. On average, 74.36% of such an utterance contains speech. The sampling frequency of the sound files is 16kHz.

The speech files were manually labeled to have a reference

for the speech-pause detection. White and babble noise was added to the clean speech at different SNRs. Then, for each SNR, the three VAD algorithms were applied on the 96 utterances using 32ms long speech frames with 50% overlap. The speech detection probability was then calculated by dividing the number of frames that were correctly classified as speech by the total number of speech frames. The false alarm probability is defined as the number of pause frames that were classified as speech frames divided by the total number of pause frames.

B. Results

Fig.1 and Fig.2 show the results obtained for white and babble noise at SNRs of 20dB and 0dB in the case of white noise, and 20dB and 10dB in the case of babble noise. In the region of low false alarm probabilities the SLR and the INNL method show better speech detection probabilities than the other two algorithms. This is due to the fact that the features that they compare to the threshold in order to make the decision of speech absence or presence are not smoothed. As a result, at the onset or offset of speech, the adjacent pause frames will probably not be classified as speech. (Note that this effect is less pronounced with the SLR method because it

uses a HMM hangover scheme [6], which can cause false speech detections at speech offset regions). The LTSD and INNL where the LTSE is used in Eq.11, on the other hand, do use a smoothed feature and will usually classify some adjacent pause frames as speech frames. If we want a low false alarm probability with these methods, the threshold has to be set so high that the feature value of a lot of speech frames will fall below this threshold, hence the worse speech detection probabilities we get with these algorithms. In the case that a higher false alarm rate is tolerated, however, the smoothed feature VAD techniques show better speech detection probabilities. Here, the unsmoothed INNL and SLR (to a lower extent because of the HMM hangover scheme) can fall below the threshold for speech frames, while the high surrounding values ensure a more constant high feature value in the case the LTSE is used. As for many applications that require a VAD algorithm a high speech detection probability is very important, these applications will probably work in the region where the INNL with LTSE shows the best results of all described methods. The main asset of the INNL method, however, lies in its simplicity compared to the SLR method. In the case a VAD is needed while no noise suppression is required, using the INNL method can significantly reduce the required computation times.

V. CONCLUSION

In this paper we introduced the INNL VAD algorithm. At high speech detection probabilities, the algorithm shows results that are better than or comparable to those of the SLR and LTSD methods. The INNL algorithm seems to be a very good alternative to the SLR method, especially if no noise suppression is required for the application, since this also reduces the computational complexity.

ACKNOWLEDGMENT

Part of the research reported on in this paper was performed in the context of the EU-FP6 project SAFIR (IST-507427). Mike Demol enjoys a PhD scholarship from the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT/SB/51463).

REFERENCES

- [1] I. Cohen., 2002. "Noise estimation by minima controlled recursive averaging for robust speech enhancement", *IEEE Signal Process. Lett.* 9 (1), 12–15.
- [2] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging", *IEEE Trans. Speech and Audio Processing*, Vol. 11, No. 5, Sep. 2003, pp. 466-475.
- [3] G. Doblinger, 1995, "Computationally efficient speech enhancement by spectral minima tracking in subbands", *Proc. Eurospeech 2*, 1513–1516.
- [4] S. Rangachari, P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments", *Speech Communication* 48 (2006), pp. 220–231.
- [5] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, A. Rubio, "Efficient voice activity detection algorithms using long-term speech information", *Speech Communication* 42 (2004), pp. 271-287.
- [6] 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.
- [7] 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. ASSP-32, no. 6, pp. 1109-1121, 1984.
- [8] M. Demol, W. Verhelst, and P. Verhoeve, "A Study of Speech Pauses for Multilingual Time-Scaling Applications", in *proc. ISCA-ITRW Multiling 2006*, Stellenbosch, South Africa, April 9-11, 2006.