Entropy Based Voice Activity Detection in Very Noisy Conditions

Philippe Renevey† *and Andrzej Drygajlo*‡

† Swiss Center for Electronics and Microtechnology, Neuchâtel, Switzerland
‡ Swiss Federal Institute of Technology, Lausanne, Switzerland
philippe.renevey@csem.ch, andrzej.drygajlo@epfl.ch

Abstract

This paper addresses the problem of robust voice activity detection (VAD) capable for working at very low signal-to-noise ratios (SNR<10dB). A new algorithm that we propose is based on entropy estimation measures of the time-frequency magnitude spectrum. The problem of the estimation of the distribution of noise in detected non-speech segments of analysed signal is also presented. It is shown that the new entropy based VAD significantly outperforms the commonly used energy-based algorithms in all (stationary, non-stationary, white and coloured) noise conditions at SNRs from 10 dB down to -10 dB and below. One of the main advantages of the method proposed in this paper is that it is not very sensitive to the changing level of noise.

1. Introduction

When speech processing systems are designed to operate in noisy conditions, an estimate of the noise spectra is often required to compensate for the influence of the noise in the system. In a single channel approach, the noise has to be estimated from the noisy speech signal. One possible approach is to segment the noisy signal into two classes: *speech segments* and *non-speech segments*. The noise can be estimated during non-speech segments. Such an approach requires that noise is stationary or slowly varying.

In this paper we present a new method for the detection of the speeech segments based on the entropy of the signal spectrum. This approach offers the advantage to be more robust to changes in the signal to noise ratio (SNR) than other classical methods based on energy thresholding and allows an easy way to calculate a detection threshold. We first present an energy based voice activity detection (VAD) that is used for comparison with the proposed method. Then we describe the entropy based VAD and we compare the performance of the two approaches in estimation of the distribution of the noise in the spectral domain.

2. Voice activity detection

Voice activity detectors (VAD) are designed to divide the speech signal into *speech segments* and *non-speech segments*. Non-speech segments are pre-utterance, post-utterance and between words silences. Algorithms to detect automatically non-speech segments are necessary for a wide range of applications like speech coding, speech recognition, speech enhancement, etc.

In the case of estimation of noise characteristics during non-speech segments, VADs have to adapt to the changes of the noise characteristics [1–3]. Robustness against noise variations is difficult to obtain. Unvoiced segments of the speech signal are more difficult to detect than voiced segments, because they are more similar to the noise and the SNR is generally lower in unvoiced than in voiced segments.

Voice activity detection is composed of two stages:

- *Parameter extraction:* Relevant parameters are extracted from the speech signal. In order to allow a good detection of the speech regions, the parameters chosen have to show a discriminative variation between speech and non speech segments.
- *Thresholding:* A threshold is applied to the extracted parameter in order to divide the speech signal between speech and non-speech segments. This threshold can be fixed or adaptive.

A VAD for every noisy conditions should include a parameter extraction stage robust against a wide variety of noises and a thresholding methods that can adapt to the changes of the noise conditions.

3. Detection based on energy thresholding

Many of the VAD algorithms are based on energy estimation measures. These algorithms have the advantages that they are simple and that no assumptions are needed about the noise characteristics. However, energy based algorithms are sensitive to noise and therefore, variations in the noise statistics can reduce the algorithm performance. In noisy conditions these algorithms obtain poor results in the detection of low energy sections like unvoiced segments of speech because such segments are masked by the noise.

The short time energy of the mth frame of length N is defined as:

$$E(m) = \sum_{n=m\cdot N}^{m\cdot N+N-1} y^2(n) \qquad (1)$$

where n is the time index.

The problem of energy based detection is to estimate online the levels of noise and speech energy and to compute a decision threshold. Generally, two thresholds are used to form a hysteresis, to avoid switches when the energy level is near to the threshold. The algorithm presented uses an adaptive calculation of the noise level. The noise level is estimated using a sliding mean calculation. It is defined as:

$$E_{noise}(m) = \lambda_1 E_{noise}(m-1) + (1-\lambda_1)E(m) \qquad (2)$$

in speech segments and

$$E_{noise}(m) = \lambda_2 E_{noise}(m-1) + (1-\lambda_2)E(m) \qquad (3)$$

in noise segments.

$\lambda_1 \in [0.85, 0.95]$ and $\lambda_2 \in [0.98, 0.999]$ are the adaptation factors for noise and speech segments, respectively. The λ_1 and

Figure 1: Short-time energy and adaptive detection threshold for *"three four five"* in white Gaussian noise at 3dB SNR.

λ_2 in Eq. 2 define a low-pass filtering of the signal energy. The value of the decay defined by λ_1 is fixed according to the following constraints: it should be small to track noise variation, but greater than the speech variation to avoid the adaptation following the variation of the energy when speech is present. This leads to decays between 60 ms and 200 ms which corresponds to the value of λ_1 given previously, when the sampling period for the energy is 10 ms. λ_2 is fixed with similar constraints: the decay must be big enough to avoid tracking the variation of the speech energy, but small enough to adapt to variations in the background noise, which leads to values between 500 ms to 1 s. The noise and speech thresholds are defined as:

$$\begin{array}{ll} T_N(m) & = E_{noise}(m) + \delta_N \\ T_S(m) & = E_{noise}(m) + \delta_S \end{array} \quad (4)$$

where $T_N(m)$ is the *noise threshold* and $T_S(m)$ the *speech threshold*, $\delta_N \in [0.1, 0.4]$ and $\delta_S \in [0.5, 0.8]$ are additive constants used to determine the thresholds. When the energy is greater than the speech threshold, speech is detected and when the energy is lower than the noise threshold speech pause is detected. The use of two thresholds defines a hysteresis and reduces the problem of fast changes in the detection which are obtained if a single threshold is used.

Fig. 1 shows the short-time log-energy for the utterance *"three four five"* in white Gaussian noise at SNR = 3 dB and the calculated threshold.

This speech detection algorithm obtains satisfactory results under the conditions where the noise varies slower than the tracking capabilities of the algorithm and that the level of energy of the speech segments is higher than the noise level. When these conditions are not respected, the performance of the algorithm dramatically degrades.

4. Detection based on entropy of the magnitude spectrum

Energy based VADs provide good performance when the energy of the speech is significantly higher than the energy of the background noise. When the SNR is very low (*e.g.* smaller than 0 dB), the energy of the background noise is similar to that of speech and detection using an energy criterion obtains poor results. However, observation of spectrograms of very noisy signals shows that the speech regions are more "organized" than noise regions. An appropriate metric to measure the organization of the signal is Shannon's entropy [4].

Originally, the entropy was defined for information sources by Shannon [5]. It measures the average length of bit code per symbol under optimal coding and is defined as:

$$H(S) = -\sum_{i=1}^{N} P\left(s(i)\right) \log_2 \left(P\left(s(i)\right)\right) \quad (5)$$

where $S = [s(1), ..., s(i), ..., s(N)]$ represents a source of N symbols, $P\left(s(i)\right)$ is the probability of emission of symbol i. The entropy $H(S)$ maximal ($H(S) = -\log_2(\frac{1}{N})$) when all the symbols are equi-probable ($P\left(s(i)\right) = \frac{1}{N}, \quad \forall i$), and minimal ($H(S) = 0$) when one symbol has a probability of one and the others zero.

Figure 2: Entropy curve and decision thresholds for *"three four five"* in white Gaussian noise at 3 dB SNR.

Figure 3: Entropy curve for *"three four five"* in band-pass white Gaussian noise (750 and 1250 Hz) after whitening filter at 0 dB SNR.

The application of the concept of entropy to the speech detection problem is based on the assumption that the signal spectrum is more organized during speech segments than during noise segments. The measure of entropy is defined in the spectral energy domain as:

$$H(|\mathbf{Y}(\omega,t)|^2) = -\sum_{\omega=1}^{\Omega} P(|Y(\omega,t)|^2) \log\left(P(|Y(\omega,t)|^2)\right) \quad (6)$$

where $P(|Y(\omega,t)|^2) = \frac{|Y(\omega,t)|^2}{\sum_{\omega=1}^{\Omega} |Y(\omega,t)|^2}$ is the *probability* of the frequency band ω for the magnitude spectrum for frame t.

In this paper, the threshold is determined using the global statistics of the entropy of the signal. This distribution shows a bimodal distribution. Two Gaussian distributions are used to model this distribution and are determined using the expectation-maximization (EM) algorithm. The statistically optimal threshold is then determined using these Gaussian distributions as it was proposed in [6]. An adaptive method as presented in Eq. 2 can also be used for the entropy based method.

$H(|\mathbf{Y}(t)|^2)$ is maximum when Y is a white noise, $H(X) = log(\Omega)$, and minimum when it is a pure tone (sinusoid), $H(Y) = 0$. The dynamics of $H(.)$ is bounded by 0 and $log(\Omega)$ and under white noise, the entropy of the noise frame is not dependent upon the noise level and the threshold can be estimated *a priori*. Under this observation, the entropy based method is well suited for speech detection in white or quasi-white noises, but will perform poorly for colored noises. The detection of the speech segments is based on a thresholding of the entropy curve between noise and speech level. The advantage is that the threshold value does change only when the spectral nature of the noise changes, but not when the noise level changes. This allows the VAD to be robust to the change of the noise level.

Fig. 2 presents the entropy and the estimated threshold for sentence *"three four five"* in white Gaussian noise (SNR=3dB).

Under colored noise the entropy curve of speech regions can be very similar to the entropy of non-speech regions. To allow detection with entropy under colored noise conditions, we propose to divide the spectrum of each frame by the average spectrum computed over all frames, or computed iteratively for each new frame for on-line methods. After division by the average spectrum, the resulting spectrum is similar to the white noise spectrum in non speech region and the entropy VAD can be applied.

$$|\tilde{Y}(\omega,t)| = \frac{|Y(\omega,t)|}{\frac{1}{T}\sum_{i=1}^{T} |Y(\omega,t)|} \quad (7)$$

where $|\tilde{Y}(\omega,t)|$ is the spectrum after the "whitening" filter. Fig. 3 shows the entropy of $|\tilde{Y}(\omega,t)|$. The speech regions can be distinguished in this signal, but there are a lot of variations in the noise segments. This is mainly due to division by the average spectrum in Eq. 7, when the latter is small for certain frequencies. To reduce this effect, before computing the spectrum, a white noise with a small amplitude is added to the signal and the speech regions become clearly detectable with entropy measure (Fig. 4).

Figure 4: Entropy curve for *"three four five"* in band pass white Gaussian noise (750-1250 Hz) after whitening filter and addition of a small magnitude white noise

5. Noise estimation methods

Noise estimation consists of estimating the distribution of the noise in the frequency domain. As the spectrum of the signal is obtained using a Fourier transform and according to the central limit theorem, we make the assumption that the noise magnitude is normally distributed in each frequency band of the spectral domain. If the noise is considered as stationary, in each frequency band its mean and variance can be considered as constant for the whole signal. The estimates of the means and variances of the noise are then calculated from the "noise-only" frames:

$$\hat{\mu}_n(\omega) = \frac{\sum_{t=1}^{T}(1-\nu(t))\cdot|y(\omega,t)|}{\sum_{t=1}^{T}(1-\nu(t))} \tag{8}$$

$$\hat{\sigma}_n^2(\omega) = \frac{\sum_{t=1}^{T}(1-\nu(t))\cdot|y(\omega,t)|^2}{\sum_{t=1}^{T}(1-\nu(t))} - \mu_n^2(\omega). \tag{9}$$

where $\nu(t) = 1$ during speech segments and zero otherwise.

Figure 5: Relative error in the estimation of the means of noise spectra in babble noise

6. Performances of the noise estimation methods

The estimates of the means and variances of noise obtained using Eqs. 8 and 9 are compared to those obtained directly from the noise signal. In order to compare the two different algorithms, we use the relative errors between the estimates of the means and variances and the values calculated directly from the noisy signal. They are expressed as:

$$\Delta\hat{\mu}_n = \frac{1}{\Omega}\sum_{\omega=1}^{\Omega}\frac{|\hat{\mu}_n(\omega)-\mu_n(\omega)|}{\mu_n(\omega)} \tag{10}$$

$$\Delta\hat{\sigma}_n^2 = \frac{1}{\Omega}\sum_{\omega=1}^{\Omega}\frac{|\hat{\sigma}_n^2(\omega)-\sigma_n^2(\omega)|}{\sigma_n^2(\omega)} \tag{11}$$

For this experiment we used the TIDigits database for the utterances and we added noises from the Noisex database. The data are down-sampled to a frequency of 8kHz. The energy-based and entropy based detection methods use window of 256 samples for analysis with a window shift of 80 samples.

Figure 6: Relative error in the estimation of the variances of noise spectra in babble noise

Figs. 5-8 present errors obtained for the estimation of the distribution of the noise magnitude spectra using energy-based detection and entropy-based detection. We observe that the entropy-based approach always outperforms the energy-based method for the estimation of both the means and variances of the noise. We also observe that the estimation error for the parameters of the noise distribution increases as the SNR increases. It is explained by the fact that in this case, an error in the detection of the speech segments leads to include speech segments in the estimation of the noise and therefore increases the relative error of the estimates. We can also observe that the entropy based approach obtains significantly better performance in non-stationary noise like factory noise than energy based method.

Figure 7: Relative error in the estimation of the means of noise spectra in factory noise

Figure 8: Relative error in the estimation of the variances of noise spectra in factory noise

7. Conclusions

In this paper we have presented a new voice activity detector based on the entropy of the magnitude spectrum of the signal. We have shown that this algorithm obtains better performance than energy-based algorithm for the estimation of the noise distribution in the spectral magnitude domain. One of the main advantages of the proposed algorithm is that it is not sensitive to changes in the noise level, but only to the spectral nature of the noise. If the noise changes both in spectral domain and in energy level, the two algorithms can be combined in order to increase the robustness of the voice activity detection algorithm.

8. References

[1] Mak, B., Junqua, J.-C., and Reaves, B., "A robust speech/non speech detection algorithm using time and frequency-based features", *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, pp. 269–272, 1992.

[2] Lamel, L., Rabiner, L., Rosenberg, A., and Wilpon, J., "An improved endpoint detector for isolated word recognition", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, pp. 777–785, August 1981.

[3] Junqua, J.-C., Mak, B., and Reaves, B., "A robust algorithm for word boundary detection in presence of noise", *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 406–412, July 1994.

[4] Abdallah, I., Montrésor, S., and Baudry, M., "Speech signal detection in noisy environment using a local entropic criterion", *in Eurospeech*, Rhodes, Greece, Sep. 1997.

[5] Shannon, C. E., "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, Oct. 1948.

[6] Van Compernolle, D., "Noise adaptation in hidden Markov model speech recognition system", *Computer Speech and Language*, vol. 3, pp. 151–168, 1989.