# Robust Voice Activity Detection Based on the Entropy of Noise-Suppressed Spectrum

*Zoltán Tüske, Péter Mihajlik, Zoltán Tobler+ and Tibor Fegyó+.*

Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Hungary
mihajlik@tmit.bme.hu

+ AITIA International Inc.
tfegyo@aitia.ai

## Abstract

A novel noise robust voice activity detection approach is introduced. The novelty of the method that it uses noise suppressed spectrum of the input signal for spectral entropy calculation. As a result excellent end-pointing performance is observed based on predefined global entropy threshold and time constraints. The effect of frame dropping controlled by the proposed algorithm was investigated on the accuracy of automatic speech recognition. The experiments were performed on Hungarian publicly available noisy and normal telephony speech databases. The relative improvement due to dropping of non-speech frames was positive in all test configurations with a maximum of 29,5%. Besides, in average more than 50% of the frames were dropped.

## 1. Introduction

The need for actually noise robust speech endpoint detection is increasingly growing. Perhaps, the most demanding application is Noise Robust Automatic Speech Recognition (NRASR). A reliable endpoint detector can straightforwardly reduce the number of recognition errors by discarding non-speech – but possibly high-energy – sound events of the noisy environment. A good voice activity detector, furthermore, is not only able to improve the accuracy and speed of a speech recognition system, but also can boost the efficiency of Wiener-filter based noise suppression. I.e. many of the NRASR systems use Wiener-filter technology, like the ETSI ADSR (Advanced Distributed Speech Recognition) standard [1] that requires distinguishing speech and non-speech segments.

In the last few decades several methods have been developed aiming at noise robust speech end pointing or voice activity detection in adverse (highly noisy) conditions. According to [2], the methods can be categorized into two classes. One of them is based on thresholds [2-3]. Generally, this kind of method first extracts certain acoustic features for each frame of the input signal and then compares these values of features with preset thresholds to classify each frame. The other one is a pattern-matching method [4] that needs the estimation of the model parameters both for speech and noise signals. The detection process is similar to a recognition process. Compared with the pattern matching method, the thresholds-based method needs neither much training data nor model training therefore it is generally simpler and faster. When using voice activity detection in a front-end, there is, however, no option for the pattern matching method.

A favourable approach amongst the threshold-based methods is where short-time spectral entropy is used as acoustic feature – rather than energy [2-3], [5]. Since the entropy is a metric of uncertainty for random variables, the entropy of speech segments is obviously different from that of the noise signals because of the inherent characteristics of speech spectrum. Though entropy based methods usually outperforms energy-based approaches especially under noisy conditions and though many papers present improvements of entropy based endpoint detection methods they all suffer from degradation of accuracy if SNR (Signal-to-Noise Ratio) falls below a limit. This is because of the “flattening” of the entropy curve as the energy of the noise - even if it is white - is rising. Therefore [6] tries to combine the energy and entropy which is a desirable approach, but cannot solve the problem for various kinds of noises.

So, short-time entropy is a preferable acoustic feature when constructing a noise robust speech endpoint detection, but it is not effective enough in the presence of high-complexity noises (narrow band, harmonic noises, background music, etc.).

Our idea was based on the behavior of the human auditory system to suppress at least the relatively slow varying noises. Therefore, after [7] we applied a minimum spectral sub band energy based noise estimation and suppression stage before the entropy calculation. Essentially the estimated noise spectrum is used to whiten the signal spectrum. As a result we observed an outstanding voice activity detection performance.

The efficiency of the developed VAD (Voice Activity Detector) algorithm was measured indirectly by speech recognition test on noisy and normal telephony speech databases in various front-end configurations.

## 2. Entropy of the magnitude spectrum

Even at low SNR the magnitude spectrum shows the speech regions more organized than the noisy ones. The assumption is that the signal spectrum is more organized during speech segments than during noise segments. The “measure of organization” of a discrete magnitude spectrum can be described similarly as the entropy of an information source by Shannon [8].

The entropy of an information source is defined as:

$$H(S) = -\sum_{i=1}^{N} P(\,s(i)\,)\cdot\log_2(\,P(s(i))\,) \tag{1}$$

Where N is the number of the symbols, s(i) is the symbol i., and P(i) is the a-posteriori probability of the symbol i. The entropy can be defined in the spectral energy domain as:

$$H(|Y(\omega,t_0)|^2) = -\sum_{\omega=1}^{\Omega} \{P(\,|Y(\omega,t_0)|^2)\cdot\log_2(\,P(|Y(\omega,t_0)|^2))\} \tag{2}$$

Where “spectral probabilities” are calculated as:

$$P(\,|Y(\omega_0,t_0)|^2) \;=\; \frac{|Y(\omega_0,t_0)|^2}{\sum_{\omega=1}^{\Omega} |Y(\omega,t_0)|^2} \qquad (3)$$

$|Y(\omega,t_0)|^2$ is the spectral energy of the frame $t_0$, and $P(|Y(\omega_0,t_0)|^2)$ is the probability of the frequency band $\omega_0$ of frame $t_0$. [3]

The entropy will be maximal, if the signal is white noise, $H_{max}=\log(\Omega)$ and minimal if it is a sinusoid $H_{min}=0$.

So the range of this measure is bounded and spectral entropy does not depend on the energy level of the frame. This makes entropy particularly suitable for threshold-based voice activity detection.

In noise, however, there are difficulties that prevent the direct application of spectral entropy in a VAD. First of all, when the level of the noise is higher than the level of the speech the entropy curve becomes hardly suitable for threshold-based decisions (see Figure 1.). The situation is even worse if the noise is not white and consequently it shows some kind of organization. In this cases [3] proposes to divide the spectrum of each frame by the average spectrum computed over T frames (4) and so whiten the spectrum, while [2] suggests a modified entropy calculation formula not detailed here.

$$\hat{Y}(\omega,t_0) = \frac{|Y(\omega,t_0)|}{\frac{1}{T}\sum_{i=1}^{T} |Y(\omega,t)|} \qquad (4)$$

We experienced that the application of (4) whitens the speech spectrum, as well, and so it does not help significantly.

Nevertheless, we found that none of these methods was able to compensate the effect of complex noises (noises having slow varying spectral harmonics). This may severely influence the accuracy of an entropy based VAD under real-life noisy conditions. (Note that the spectral entropy of a simple sine wave is maximal!). So we looked for a solution that is insensitive for relatively constant narrow band noises.

Our guess was that dividing by the spectrum of an *estimated noise* might result in a better VAD performance and fulfill the requirements mentioned previously.

## 3. Noise Estimation and Suppression

After [7] we implemented an algorithm for the estimation of noise. The spectral minimum is computed over a set of prior T smoothed frames. The assumption is that the speech has a non-stationary nature and so the energy of speech sometimes falls to zero in all frequencies, though not simultaneously. Only the energy of noise is relatively constant in a given frequency. Therefore the minima in each spectral bin over a sufficiently long interval can give an estimation of the noise, if the spectral changes of the noise are slower than that of the speech.

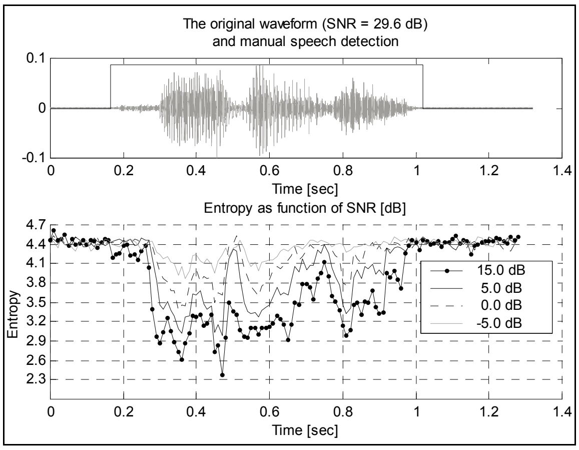

*Figure 1.* Entropy under various SNR

So the spectrum of the noise is estimated by finding the minimum of each frequency band ω (or FFT bin, in practice) as follows:

$$Y_{noise}(\omega,t_0) = \min_{t=t_0-T \ldots t_0} \{\, Y_{signal}(\omega,t) \,\} \qquad (5)$$

where $Y_{signal}(\omega,t)$ is the magnitude spectrum of the signal.

It was found that if the estimated noise spectrum (5) was substituted into (4) instead of the average spectrum the resulting noise-suppressed spectrum showed considerably better basis for entropy calculation than the original spectrum. In this way the spectrum of a color noise *during speech* can be whitened without whitening the speech spectrum itself. Therefore the entropy curve of a signal is much more emphasized during noisy speech segments than during various noise segments. In addition, any relatively constant harmonic noises are suppressed because the described modification of (4) means spectral subtraction of noise in the logarithmic domain.

However, if the spectrum of the noise changes suddenly e.g., the energy of some frequency band rises abruptly the estimator will follow this changing with considerable latency.

As a result a part of the noise to be suppressed resides in the spectrum potentially resulting in a voice activity detection failure.

Therefore we extended the noise estimation as follows. Patterns from the future and the past are also used for reducing the effects of sudden changes. Two estimations of noise spectrum are computed from the past and from the future, respectively. The actual noise spectrum is obtained bin-by-bin through choosing the *larger* value among future and past estimated noise spectrum bin.

$$\hat{Y}_{noise}(\omega,t_0) = \mathrm{MAX}\,\Big[\; \min_{t=t_0-T_1 \ldots t_0} \{\, Y_{signal}(\omega,t) \,\},\; \min_{t=t_0 \ldots t_0+T_2} \{\, Y_{signal}(\omega,t) \,\} \;\Big] \qquad (6)$$

This modification causes a latency of $T_2$ (250 ms in our experiments), which is allowable in real-time ASR.

## 4. Overview of the NSSE-VAD algorithm

In the following we briefly summarize the developed voice activity detection algorithm. For convenience it will be called NSSE-VAD (Noise-Suppressed Spectral Entropy-based Voice Activity Detection) in the rest of this paper.

**Step 1: STFT**

Before the 128-bin FFT calculation a 32 ms long 31.2% overlapped Hanning-window was applied. To obtain the smoothed spectrum a two-dimensional FIR filter (S) was used on the FFT output. The algorithm gave better result, if smoothing in both frequency and time domain was used.

$$Y_{\text{smoothed}}(\omega_0,t_0) = \sum_{\omega=-2}^{2} \sum_{t=-2}^{2} Y_{\text{original}}(\omega_0+\omega,t_0+t)\cdot S(\omega+3,t+3) \qquad (7)$$

Where $Y_{\text{original}}(\omega,t)$ is the magnitude spectrum as the output of the FFT of the t-$^{\text{th}}$ frame of the signal, and S(i,j) is the element of i-$^{\text{th}}$ row and j-$^{\text{th}}$ column of the smoothing matrix S.

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 3 & 2 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \cdot \frac{1}{35} \qquad (8)$$

**Step 2: Noise Estimation and Subtraction**

To improve the robustness of spectral entropy against various noises we use the previously described noise estimation algorithm: (6) is applied on $Y_{\text{smoothed}}$ (with $T_1$ = 750ms). Then the smoothed spectrum is divided by the estimated noise spectrum.

$$Y_{\text{noise-suppressed}} = Y_{\text{smoothed}} / \hat{Y}_{\text{noise}} \qquad (9)$$

**Step 3: Entropy calculation**

The spectral entropy is computed on the noise-suppressed spectrum applying (2).

**Step 4: Classification**

We used a global threshold (91% of the maximum entropy in our experiments). If the computed entropy of the actual frame was less than the threshold, the frame was considered as speech, else as noise.

**Step 5: Final speech/non-speech decisions**

In order to extend voice activity detection to endpoint detection – if required – a final layer is applied to make voice activity detection regions coherent. This operation is based on simple time constraints. For example, as speech contains silence, too, the algorithm is insensitive to short silences (~0.1 sec) between two speech regions. So, in these cases the speech/non-speech output of the classification is judged over.

## 5. Evaluation

Various speech recognition tests were carried out in order to measure the effect of the NSSE-VAD in term of recognition error rate improvement.

### 5.1. Databases

We used two test corpuses, a normal and a noisy one. The first one is called BESZTEL or Normal DB and the test set consists of about 6000 utterances – typically command words – which were not marked as noisy at the annotation process. The other database was designed especially for NRASR purposes therefore the recordings contained various environmental noises during the speech. This database is called TESZTEL [9] or Noisy DB and the size of the test set was about 1200 and the utterances were typically 2 to 4 syllable length words. A limitation of this database is that some of the recordings are AGC (Automatic Gain Control) distorted.

The training database (MTBA [10]) was a normal telephone speech database therefore the test condition in case of BESZTEL can be considered as well-matched while testing on TESZTEL as highly mismatched due to noises.

### 5.2. Recognition tasks

Middle vocabulary speech recognition tests were performed on both databases using explicit endpoint-detection and not. The vocabulary size was about 1000 in case of normal test and about 250 at the noisy test. There was implicit endpoint detection at the recognition phase due to a silent model.

### 5.3. Evaluation methodology

Three series of experiments were run with three different front-end configurations augmented with the NSSE-VAD. The relative improvement achieved by our detector through non-speech frame dropping was measured by four recognition tests in each turn. In one half of the experiments standard features were used (39 dimensional feature vectors), in the other half absolute energy was suppressed (38 dimensional features).

The effect of NSSE-VAD was first investigated with a simple MFCC front-end. Then this MFCC front-end was extended with Blind Equalization (MFCC+BEQ). Finally the ETSI standard ADSR front-end was used in the experiments. In the latter case the standard's VAD was evaluated, too.

### 5.4. Results

Table 1 shows the reference recognition error rates of three front-ends as a reference. Table 2 a) presents the absolute error rates when non-speech frame dropping was switched on. The effect of NSSE-VAD and ADSR-VAD is emphasized in Table 2 b) where the relative improvements caused by the explicit end pointing are displayed. Note, that while the ADSR-VAD dropped 3,5% and 24,9% of frames at the Noisy and the Normal DB, respectively, the dropping rate of NSSE-VAD was 60% and 52,6% on the two databases accordingly.

*Table 1,* Reference word error rates (WER, %) of the front-ends without VAD

| | With energy | | Absolute energy suppressed | |
|---|---|---|---|---|
| | Normal | Noisy | Normal | Noisy |
| ADSR | **5,23** | **51,24** | **6,26** | **21,20** |
| CC | 4,78 | 45,61 | 5,26 | 27,33 |
| CC+BEQ | 4,76 | 43,60 | 5,43 | 19,97 |

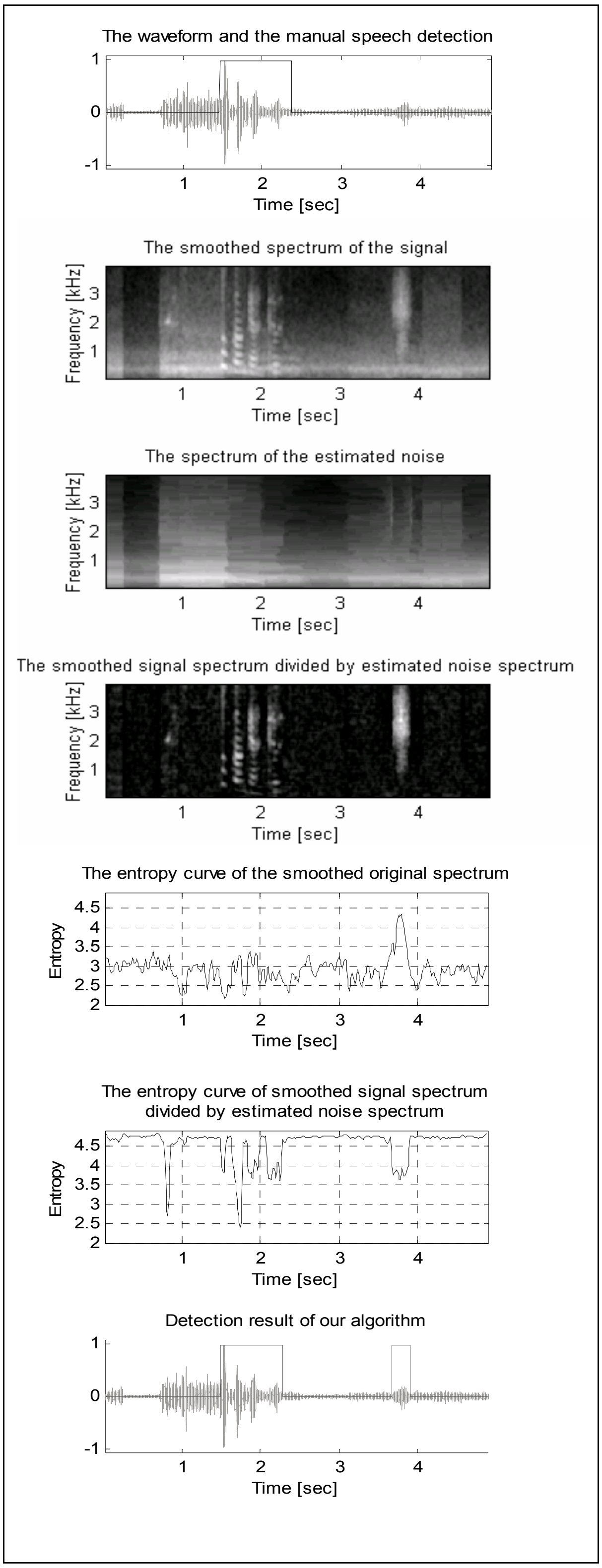

*Figure 3.* Illustration of NSSE-VAD operation

*Table 2, a)* Word error rates (WER, %) of different front-ends with the two end pointing methods

| VAD | Front-end | With energy | | Absolute energy suppressed | |
|---|---|---|---|---|---|
| | | Normal | Noisy | Normal | Noisy |
| ADSR | ADSR | 5,21 | 51,07 | 6,26 | 21,20 |
| NSSE | ADSR | 5,11 | 36,14 | 5,86 | 20,54 |
| NSSE | CC | 4,66 | 35,51 | 5,08 | 22,77 |
| NSSE | CC + BEQ | 4,70 | 33,83 | 5,23 | 18,65 |

*Table 2, b)* Relative improvements (in %) due to end-pointing

| VAD | Front-end | With energy | | | Absolute energy suppressed | | |
|---|---|---|---|---|---|---|---|
| | | Normal | Noisy | Avg. | Normal | Noisy | Avg. |
| ADSR | ADSR | +0,38 | +0,33 | **+0,36** | 0,00 | 0,00 | **0,00** |
| NSSE | ADSR | +2,29 | +29,47 | **+15,88** | +6,39 | +3,11 | **+4,75** |
| NSSE | CC | +2,51 | +22,14 | **+12,33** | +3,42 | +16,68 | **+10,05** |
| NSSE | CC + BEQ | +1,26 | +22,41 | **+11,83** | +3,68 | +6,61 | **+5,15** |

## 6. Conclusions

A novel threshold-based noise robust VAD has been introduced. The novelty of the approach is the noise suppression stage before entropy calculation. This stage has a noise-whitening effect proven extremely useful for this type of speech detection. As the speech recognition results on Hungarian databases show the relative improvements achieved by the proposed VAD algorithm through end pointing are in the range of 1,26% – 29,47%, and the average improvements are 4,75% – 15,88%. Which are much higher than in the case of ADSR-VAD (0% – 0,38%). What is more, the simple MFCC front-end along with the proposed NSSE-VAD performs nearly as well as the ADSR front-end in the highly mismatched, energy-suppressed condition and significantly better in all other test situation. Besides the NSSE-VAD speeds up the recognition process remarkably by extensive (more than 50% in our experiments) frame dropping rate.

## 7. References

[1] *ETSI standard doc., ETSI ES 202 050 v1.1.1.*

[2] Chuan JIA, Bo XU: An Improved Entropy-Based Endpoint Detection Algorithm, ICSLP'02, 2002, Beijing

[3] Philippe Renevey and Andrej Drygajlo: Entropy Based Voice Activity Detection in Very Noisy Conditions, Eurospeech 2001, Aalborgh

[4] E. Kosmides , E. Dermatas, G. Kokkinakis, "Stochastic endpoint detection in noisy speech", SPECOM Workshop, 109-114, 1997.

[5] Jialin Shen, Jeihweih Hung, Linshan Lee, "Robust entropy based endpoint detection for speech recognition in noisy environments", International Conference on Spoken Language Processing, Sydney, 1998.

[6] Liang-sheng Huang, Chung-ho Yang, "A novel approach to robust speech endpoint detection in car environments", International Conference on Acoustic, Speech and Signal Processing, 2000.

[7] Izhak Shafran & Richar Rose: Robust Speech Detection And Segmentation For Real-Time ASR Application

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

[9] http://alpha.ttt.bme.hu/speech/hdbtesztelen.php

[10] http://alpha.ttt.bme.hu/speech/hdbMTBA.php