Exploring Non-linear Transformations for an Entropy-based Voice Activity Detector

Jordi Solé-Casals, Pere Martí-Puig, Ramon Reig-Bolaño

Digital Technologies Group, University of Vic, Sagrada Família 7,
08500 Vic, Spain
{jordi.sole, pere.marti, ramon.reig}@uvic.cat

Abstract. In this paper we explore the use of non-linear transformations in order to improve the performance of an entropy based voice activity detector (VAD). The idea of using a non-linear transformation comes from some previous work done in speech linear prediction (LPC) field based in source separation techniques, where the score function was added into the classical equations in order to take into account the real distribution of the signal. We explore the possibility of estimating the entropy of frames after calculating its score function, instead of using original frames. We observe that if signal is clean, estimated entropy is essentially the same; but if signal is noisy transformed frames (with score function) are able to give different entropy if the frame is voiced against unvoiced ones. Experimental results show that this fact permits to detect voice activity under high noise, where simple entropy method fails.

Keywords: VAD, score function, entropy, speech

1 Introduction

In speech and speaker recognition a fast and accurate detection of the speech signal in noise environment is important because the presence of non-voice segments or the omission of voice segments can degrade the recognition performance [1-3]. On the other hand in a noise environment there are a set of phonemes that are easily masked, and the problem of detecting the presence of voice cannot be solved easily. The problem is further complicated by the fact that the noise in the environment can be time variant and can have different spectral properties and energy variations. Also there are limitations on the admissible delay between the input signal, and the decision of the presence or absence of voice.

In the last several decades, a number of endpoint detection methods have been developed. According to [4] we can categorize approximately these methods into two classes. One is based on thresholds [4-7]. Generally, this kind of method first extracts the acoustic features for each frame of signals and then compares these values of features with preset thresholds to classify each frame. The other is pattern-matching method [8-9] that needs estimate the model parameters of speech and noise signal. The detection process is similar to a recognition process. Compared with pattern-

matching method, thresholds-based method does not need keep much training data and train models and is simpler and faster.

Endpoint detection by thresholds-based method is a typical classification problem. In order to achieve satisfied classification results, it is the most important to select appropriate features. Many experiments have proved that shortterm energy and zero-crossing rate fail under low SNR conditions. It is desirable to find other robust features superior to short-term energy and zero-crossing rate. J. L. Shen [10] first used the entropy that is broadly used in the field of coding theory on endpoint detection. Entropy is a metric of uncertainty for random variables, thus it is definite that the entropy of speeches is different from that of noise signals because of the inherent characteristics of speech spectrums.

However, it is found that the basic spectral entropy of speech varies to different degrees when the spectrum of speech is contaminated by different noise signals especially high noise signals. The varieties make it difficult to determine the thresholds. Moreover, the basic spectral entropy of various noises disturbs the detection process. It is expected that there exists a way by which it is possible that (1) the entropy of various noise signals approaches to one another under the same SNR condition, (2) the curve of noise entropy is flat, and (3) the entropy of speech signals differs from that of noise signals obviously.

This paper investigates different non-linear transforamtions on the input signal to improve voice activity detection based on spectral entropy. Preliminary experimental results shown that it is possible to improve basic spectral entropy, specially in the presence of non-gaussian noise or colored noise.

2 Entropy

Originally, the entropy was defined for information sources by Shannon [11] and is a measure of the uncertainty associated with a random variable. Is defined as:

$$H(S) = -\sum_{i=1}^{N} p(s_i) \log p(s_i) \tag{1}$$

where S = [s_1, ..., s_i, ..., s_N] are all the possible values of the discrete random variable S, and $p(s_i)$ is the probability mass function of S.

In case of speech, for certain phonemes, the energy is concentrated in a few frequency bands, and therefore will have low entropy as the signal spectrum is more organized during speech segments; while in the case of noise with flat spectrum or low pass noise, the entropy will be higher. The measure of entropy is defined in the spectral energy domain as:

$$p_j(k) = \frac{\left|S_j(k)\right|}{\sum_{m=1}^{N} \left|S_j(m)\right|} \tag{2}$$

where $S_j(k)$ is the kth DFT coefficient in the jth frame. Then the measure of entropy is defined in the spectral energy domain as:

$$H(j) = -\sum_{i=1}^{N} p_j(k) \log p_j(k) \quad (3)$$

As $H(j)$ is maximum when S_j is a white noise and minimum (null) when it is a pure tone, 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 or non-Gaussian noises. We will see that applying some nonlinear function on the signal the entropy based method can deal with these cases.

3 Exploring score function as non-linear transformation

Inspired in BSS/ICA algorithms [see 12 and references therein] or blind linear/non-linear deconvolution [13-14], we propose to use score function to non-linearly modify the signal before calculating entropy for VAD process. What we expect is that as score function is related to pdf of the signal, we will enhance the difference between voice and non voice frames, even in very noisy environments.

3.1 Score function

Given a vector Y, the so-called score is defined as:

$$\psi_Y(u) = \frac{\partial \log p_Y(u)}{\partial u} = \frac{p'_Y(u)}{p_Y(u)} \quad (4)$$

Since we are concerned by nonparametric estimation, we will use a kernel density estimator [15]. This estimator is easy to implement and has a very flexible form, but suffers from the difficulty of the choice of the kernel bandwidths. Formally we estimate $p_Y(u)$ by:

$$\hat{p}_Y(u) = \frac{1}{hT} \sum_{t=1}^{T} K\left(\frac{u - y(t)}{h}\right) \quad (5)$$

from which we get an estimate of $\psi_Y(u)$ by $\psi_Y(u) = \frac{\hat{p}'_Y(u)}{\hat{p}_Y(u)}$. Many kernel shapes can be good candidates, for our experiments we used the Gaussian kernel. A "*quick and dirty*" method for the choice of the bandwidth consists in using the rule of thumb $h = 1.06\hat{\sigma}T^{-1/5}$. Other estimators may be found, and used, but experimentally we noticed that the proposed estimator works fine.

3.2 Other functions

In many BSS/ICA algorithms, score function is approximated by a fixed function, depending on the sub-Gaussian or super-Gaussian character of the signals. In this case, functions like $\tanh(u)$, $u\exp(-u^2/2)$ or u^3 are used instead of calculating the true score function $\psi_Y(u)$. As a preliminary analysis only results with the true score function will be presented in this paper, but these more simple functions can maybe be a good candidates, especially for real-time applications, as they avoid the task of estimate the true score function.

4 Proposed method

The proposed method in order to explore these non-linear functions for VAD is shown in figure 1. The signal is framed and score function is estimated for each frame, using this output as the input to the next block (entropy calculation) instead of the original frame. What we are interested is looking at this entropy of the scored frame compared with the original frame (without score function).

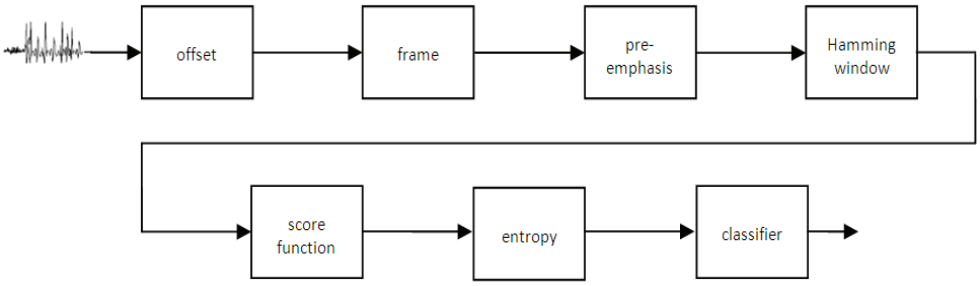

Fig. 1. Block diagram of the proposed method.

Pre-processing stage will be done according to ETSI standard [16]. According to that, a notch filtering operation is applied to the digital samples of the input speech signal s_{in} to remove their DC offset, producing the offset-free input signal s_{of} :

$$s_{of}(n) = s_{in}(n) - s_{in}(n-1) + 0.999 s_{of}(n-1) \quad (6)$$

The signal is framed in a 25 ms frame length, that corresponds to 200 samples for a sampling rate $f_s = 8\,kHz$, with frame shift interval of 10 ms, that corresponds to 80 samples for a sampling rate $f_s = 8\,kHz$. A pre-emphasis filter is applied to the framed offset-free input signal,

$$s_{pe}(n) = s_{of}(n) - 0.97 s_{of}(n-1) \quad (7)$$

and finally a Hamming window is applied to the output of the pre-emphasis block. Once obtaining windowed frame of *N* samples, score function is estimated according to eq. 4 and 5, and then spectral entropy is computed by means of eq. 3. The final decision (voiced frame – unvoiced frame) is taken by means of a threshold, even if more complex and better rules can be considered, but here we are only exploring the differences between estimated entropy by using score function or not, in order to facilitate the classifier block.

5 Experiments

Several experiments are done in order to investigate the performance of the system. First of all, we are interested in looking how looks like a scored frame compared to a simple frame. In figure 2 we can see a voiced signal, its estimated entropy and its estimated entropy over the score function. In this case, when the voice signal is clean (good SNR), we can observe a similar shape of the entropy for the original frames and scored frames.

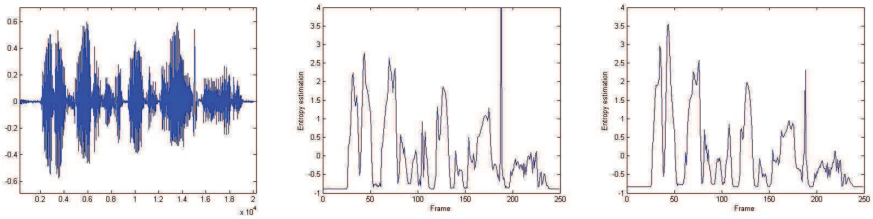

Fig. 2. Signal input (left) and estimated entropy (without score function on the middle, and with score function on the right)

If we add Gaussian noise to the signal, the results begin to be different, as we can see in figure 3, where we show the input signal (top left) and the clean signal for shake of clarity (down left), and the estimated entropy without and with score function. Even if noise is very high, we can observe that entropy is different in the parts of signal containing speech, but of course the difference is not as clear as in figure 2. Also we can observe that results without and with score function are not as similar as before. If noise is much harder, entropy estimation does not permit to distinguish between noise and speech, and then no voice activity can be detected.

On the other hand, if noise is uniform we can obtain better results for the estimation of entropy with score function. Results in figure 4 are obtained with uniform noise, and we observe that without score function we cannot distinguish between noise and speech, while it can be done with score function.

Using a simple threshold on the estimated entropy, we can make a decision on the signal, in order to decide if the frame is a voiced or unvoiced one. Of course, more elaborated procedures must be used instead of a simple trigger, as explained in the literature, but here for the sake of simplicity we will present some results only with a threshold.

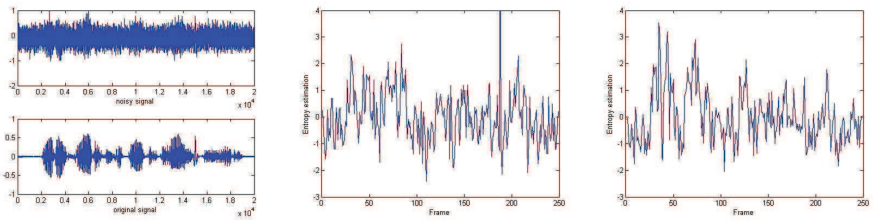

Fig. 3. Signal input (top left) with Gaussian noise, and estimated entropy (without score function on the middle, and with score function on the right)

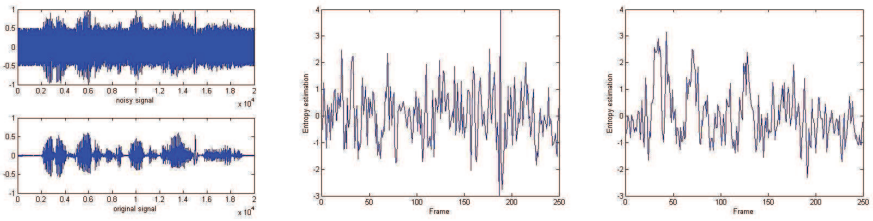

Fig. 4. Signal input (top left) with uniform noise, and estimated entropy (without score function on the middle, and with score function on the right)

Figure 5 shows the results obtained without score function (left) and with score function (right), with a clean speech signal (no noise added). We can see that asimple threshold can give us good results and that they are very similar, as the estimated entropy with and without score function are (approximately) equivalent, as we showed in figure 2.

On the other hand, if speech signal is noisy, and due to the fact that estimated entropy is no more equal without or with score function, the detection of speech is much more hard and different results are obtained using ore don't using score function. Results for this case are presented in figure 6.

In this case, scored frames give better results and hence the voice is better detected even if it is hidden by noise. Of course VAD doesn't gives perfect results, as we can see comparing the detection presented in fig. 6 with the true speech signal, plotted on the bottom of the figure for the sake of clarity, but this can be improved by designing a better classifier, as mentioned before.

6 Conclusions

In this paper, the use of non-linear transformations for improve a voice activity detector is explored.

Score function is used as non-linear transformation, estimated by means of a Gaussian kernel, and entropy is used as a criterion to decide if a frame is voiced or unvoiced.

If speech signal is clean, results are essentially the same, due to the fact that score function doesn't change the entropy of the signal. But in the case of noisy speech signal, the estimated entropy is no more equivalent, hence giving different results. Is in this case where frames pre-processed with score function give better results and voice can be detected into a very noisy signal.

Future work will be done exploring other non-linear transformation, trying to simplify and reduce the complexity of the system in order to be implemented in real-time applications. On the other hand, classifier will also be improved deriving some heuristic rules, for example, or by using more complex systems as neural networks, in order to minimise incorrect activity detections.

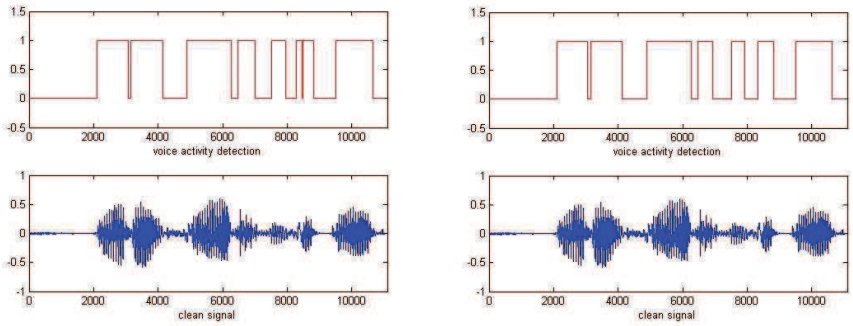

Fig. 5. Voice activity detection obtained with a simple threshold. On the left, estimating the entropy without score function. On the right, estimating the entropy with score function. As the estimated entropy is essentially the same, results are very coincident.

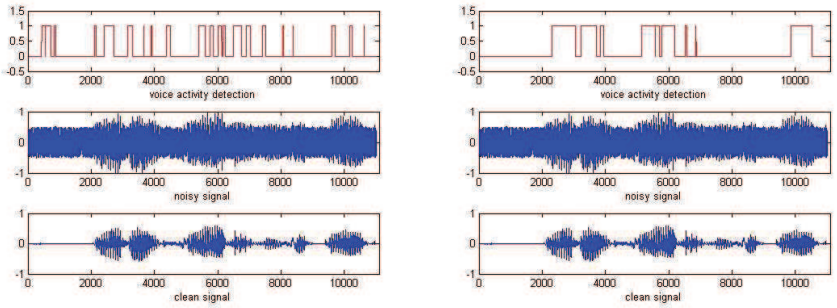

Fig. 6. Voice activity detection obtained with a simple threshold. On the left, estimating the entropy without score function. On the right, estimating the entropy with score function. As now the speech signal is noisy, the estimated entropy is different and hence the detection is also different. We can observe that scored signal gives better results.

Acknowledgments. This work has been supported by the University of Vic under the grant R0904

References

1. G. S. Ying, C. D. Mitchell, and L. H. Jamieson, “Endpoint Detection of Isolated Utterances Based on a Modified Teager Energy Measurement”, *Proc. ICASSP,* II.732-735, 1993.
2. Jia-lin Shen, Jeih-weih Hung, Lin-shan Lee, "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments", *Proc. ICSLP* CD-ROM 1998.
3. W.-H.Shin, B.-S.Lee,Y.-K.Lee,J.-S.Lee, “Speech/Non-Speech Classification Using Multiple Features For Robust Endpoint Detection”, *Proc. ICASSP*, 1399-1402, 2000.
4. Chuan Jia, Bo Xu, "An improved Entropy-based endpoint detection algorithm", *Proc. ICSLP,* 2002.
5. Woo-Ho Shin, Byoung-Soo Lee, Yun-Keun Lee, Jong-Seok Lee, “Speech/non-speech classification using multiple features for robust endpoint detection”, *Proc. ICASSP*, 2000.
6. Stefaan Van Gerven, Fei Xie, “A Comparative study of speech detection methods”, European Conference on Speech, Communication and Techonlogy,1997.
7. Ramalingam Hariharan, Juha Häkkinen, Kari Laurila, “Robust end-of-utterance detection for real-time speech recognition applications”, *Proc. ICASSP*, 2001
8. A. Acero, C. Crespo, C. De la Torre, J. Torrecilla, “Robust HMM-based endpoint detector”, *Proc. ICASSP*, 1994.
9. E. Kosmides, E. Dermatas, G. Kokkinakis, “Stochastic endpoint detection in noisy speech”, *SPECOM Workshop*, 109-114, 1997.
10. Jialin Shen, Jeihweih Hung, Linshan Lee, “Robust entropybased endpoint detection for speech recognition in noisy environments *Proc. ICSLP*, Sydney, 1998.
11. Shannon, C. E., “A mathematical theory of communication”, *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, Oct. 1948.
12. A. Hyvärinen, J. Karhunen, E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001
13. J. Solé-Casals, A. Taleb, C. Jutten “Parametric Approach to Blind Deconvolution of Nonlinear Channels”, *Neurocomputing*, vol.48, pp. 339-355, 2002.
14. J. Solé-Casals, E. Monte, A. Taleb, C. Jutten, "Source separation techniques applied to speech linear prediction", *Proc. ICSLP*, 2000.
15. W Härdle, *Smoothing Techniques with implementation in S*, Springer Verlag, 1990
16. ETSI standard doc., *ETSI ES 201 108 V1.1.3* (2003-09)