A non-linear VAD for noisy environments

Jordi Solé-Casals, Vladimir Zaiats

Digital Technologies Group, University of Vic, Sagrada Família 7,
08500 Vic, Spain
{jordi.sole, vladimir.zaiats}@uvic.cat

Abstract. This paper deals with non-linear transformations for improving the performance of an entropy-based voice activity detector (VAD). The idea to use a non-linear transformation has already been applied in the field of speech linear prediction, or linear predictive coding (LPC), based on source separation techniques, where a score function is added to classical equations in order to take into account the true 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 the signal is clean, the estimated entropy is essentially the same; if the signal is noisy, however, the frames transformed using the score function may give entropy that is different in voiced frames as compared to nonvoiced ones. Experimental evidence is given to show that this fact enables voice activity detection under high noise, where the simple entropy method fails.

Keywords: VAD, score function, entropy, speech

1 Introduction: cognitive modeling of speech

[1] gives a detailed account that the process of speech comprehension can be divided into a number of subprocesses ranging from those responsible for matching acoustic input against some internal representation of the words in the language to those involved in extracting meaning from the string(s) of lexical items hypothesized by the lexical-access process and constructing a message-level interpretation of the sentence. The modeling of speech comprehension, similar to that in other domains of cognitive modeling, can be guided by the following three considerations. First, to what extent can a subprocess involved in the comprehension process be assigned to separate informationally encapsulated modules? Second, to what extent do these modules interact? And third, what kinds of computational architectures are adequate for modeling these processes?

There are a number of properties of the raw acoustic input that makes the task of spoken-word recognition, and speech comprehension in general, particularly difficult. One of these problems is that the input is not conveniently divided into acoustic segments that correspond neatly to the individual words uttered by the speaker [1].

One of the steps in order to obtain this necessary segmentation is the so-called Voice Activity Detection (VAD). A voice activity detector is used to detect the presence of speech in an audio signal. VAD plays an important role as a pre-

processing stage in different audio processing applications. For example, the performance of speech recognition, speaker recognition, and source localization can be improved by applying these algorithms only to parts of the audio that are identified as speech. Furthermore, in voice over IP (VoIP) and mobile telephony applications, VAD can reduce bandwidth usage and network traffic by transmitting audio packets only if speech is detected. Video conferencing is another challenging application of source localization where VAD is useful. In this application, source localization is performed and the video camera is steered in the direction of the audio source when speech is detected using VAD. VAD is also used to identify noise; therefore it can be waived in systems such as hearing aids and audio conferencing devices [2].

2 Fundamentals of VAD and applications

In this section, we give a short account of the fundamentals of VAD and some important applications. For more details, see [3] and the references therein.

An important drawback affecting most of speech processing systems is the environmental noise and its harmful effect on the system performance. Examples of such systems are new wireless communication voice services or digital hearing aid devices. In speech recognition, there are still technical barriers inhibiting such systems from meeting the demands of modern applications. Speech/non-speech detection is an unsolved problem in speech processing which has influence over different applications including robust speech recognition, discontinuous transmission, real-time speech transmission in Internet, or combined noise reduction and echo cancellation schemes in the context of telephony. The speech/non-speech classification task is not as trivial as it may seem to be, and most of the VAD algorithms fail when the level of the background noise increases.

VAD is employed in many areas of speech processing. Recently, VAD methods have been described in the literature for several applications including mobile communication services. The most important VAD applications in speech processing are in the fields of coding, enhancement and recognition.

In coding, VAD is widely used within the field of speech communication for achieving high speech coding efficiency and low-bit rate transmission. The concept of silence detection and that of comfort noise generation lead to dual-mode speech coding techniques. Different modes of operation of a speech codec are: (i) active speech codec, and (ii) silence suppression and comfort noise generation modes. The International Telecommunication Union (ITU) adopted a toll-quality speech coding algorithm known as G.729 to work in combination with a VAD module in DTX mode. A full rate speech coder is operational during active voice speech, but a different coding scheme is employed for an unvoiced signal, using fewer bits and resulting in a higher overall average compression ratio.

In speech enhancement, the aim is to improve the performance of speech communication systems in noisy environments. It mainly deals with suppressing background noise from a noisy signal. A difficulty in designing efficient speech enhancement systems is the lack of explicit statistical models for speech signals and noise processes. In addition, speech signals, and possibly noise processes, are not

strictly stationary. Speech enhancement normally assumes that the noise source is additive and uncorrelated with the clean speech signal. One of the most popular methods for reducing the effect of a background (additive) noise is spectral subtraction. The popularity of spectral subtraction is largely due to its relative simplicity and easy implementation. The spectrum of the noise N(f) is estimated during unvoiced periods detected by a VAD, and subtracted from the spectrum of the current frame X(*f*). This gives an estimate of the spectrum S(*f*) of the clean speech.

In recognition, the performance of speech recognition systems is strongly influenced by the quality of the speech signal. VAD is a very useful technique for improving the performance of these systems working in noisy scenarios. A VAD module is used in most of the speech recognition systems within the feature extraction process for speech enhancement. The noise statistics such as spectrum are estimated during unvoiced periods in order to apply a speech enhancement algorithm (spectral subtraction or a Wiener filter). On the other hand, non-speech frame-dropping (FD) is also a frequently used technique in speech recognition to reduce the number of insertion errors caused by the noise. It consists in dropping unvoiced periods (based on the VAD decision) from the input of the speech recognizer. This reduces the number of insertion errors due to the noise that can be a serious error source under high mismatch training/testing conditions.

In the last several decades, a number of endpoint detection methods have been developed. According to [4] we can approximately classify these methods into two classes. One class is based on thresholds [4-7]. Generally, this kind of method first extracts acoustic features for each frame of signals and then compares these values of features with preset thresholds, in order to classify each frame. Another class is pattern-matching [8-9] that requires estimation of model parameters for speech and noise signals. The detection process is similar to a recognition process. Compared with the pattern-matching method, the thresholds-based method does not need keep much training data and training models, is simpler and faster.

Endpoint detection by the thresholds-based method is a typical classification problem. In order to achieve reasonable classification results, it is the most important to select appropriate features. Many experiments have proved that short-term 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. The entropy is a measure of uncertainty for random variables; therefore one can guess that the entropy of speech is different from that of noise due to the intrinsic characteristics of speech spectrums.

However, it has been discovered that the basic spectral entropy of speech varies in a different manner when the speech spectrum is contaminated by different noise signals, especially by high noise signals. High variability makes it very difficult to determine thresholds. Moreover, the basic spectral entropy of different noises disturbs the detection process. It is expected that there exists a way such that: (i) the entropy of different noise signals approaches one another under the same SNR condition, (ii) the noise entropy curve is flat, and (iii) the entropy of speech signals is clearly different from that of the noise.

This paper focuses on non-linear transformations of input signals enabling us to improve voice activity detection based on the spectral entropy. Earlier experimental

results shown that the basic spectral entropy can be improved, especially in the presence of non-gaussian noise or colored noise.

3 Entropy

Originally, entropy was defined for information purposes by C. Shannon [11] in 1948. Entropy is a measure of uncertainty associated with a random variable. For a discrete random variable S, it is defined as follows:

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

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

In the case of speech, the energy of certain phonemes is concentrated in a few frequency bands. Therefore the entropy will be low when the signal spectrum is more organized during speech segments. 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 in the following way:

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

where $S_j(k)$ is the kth discrete Fourier transform (DFT) coefficient in the jth frame. Then the measure of entropy is defined in the spectral energy domain by the following formula:

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

Since $H(j)$ attains its maximum when S_j is a white noise and becomes minimum (null) when it is a pure tone, the entropy of the noise frame does not depend on the noise level, and the threshold can be estimated a priori. With this observation, the entropy-based method fits well for speech detection in white or quasi-white noises, but performs poorly for colored or non-Gaussian noises. We will see that application of a nonlinear function to the signal will enable the entropy-based method to deal with these cases.

4 Exploring score function as non-linear transformation

Inspired in BSS/ICA algorithms, see [12] and the references therein, or in blind linear/non-linear deconvolution [13-14], we propose to use a score function for a non-

linear modification of the signal prior to calculation of the entropy for VAD process. What we expect is that since the score function is related to pdf of the signal, we will enhance the difference between voiced and unvoiced frames, even in very noisy environments.

4.1 Score function

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

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

Since we deal with nonparametric estimation, we will use a kernel density estimator [15]. This estimator is easy to implement, its form is very flexible, but the choice of the kernel bandwidth is a certain drawback. Formally, we estimate $p_Y(u)$ in the following way:

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

Then an estimate of $\psi_Y(u)$ is obtained by $\psi_Y(u) = \frac{\hat{p}'_Y(u)}{\hat{p}_Y(u)}$. Many kernel shapes can be good candidates; we used the Gaussian kernel for our purposes. A "*quick and dirty*" method for choosing the bandwidth consists in using the rule of the thumb giving $h = 1.06\hat{\sigma}T^{-1/5}$. Of course, other estimators may be found and used; what we claim is that our estimator behaves fine in simulations.

4.2 Other functions

In many BSS/ICA algorithms, the score function is approximated by a fixed function, depending on sub-Gaussian or super-Gaussian character of the signal. In this case, functions like $\tanh(u)$, $u\exp\left(-u^2/2\right)$ or u^3 are used instead of calculation of the true score function $\psi_Y(u)$. As a preliminary analysis, we present here some results related to the true score function only. In real-time applications, these simple functions can be good candidates, since they avoid the need to estimate the true score function.

5 Proposed method

The method proposed for exploration of non-linear functions for VAD is shown in Figure 1. The signal is framed and the 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 in is looking at this entropy of the scored frame compared with the original frame (without using the score function).

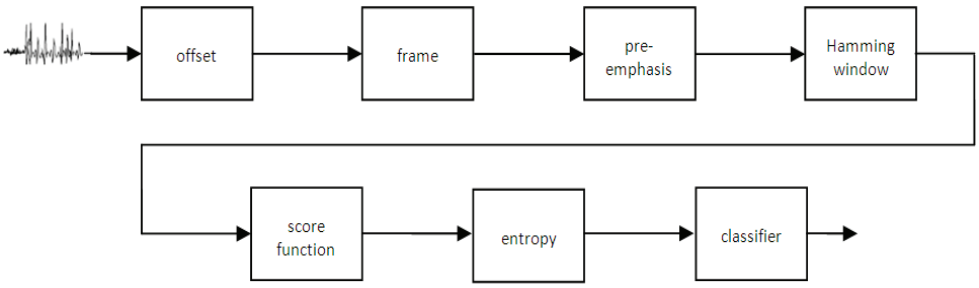

Fig. 1. Block diagram of the proposed method.

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

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

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

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

Finally a Hamming window is applied to the output of the pre-emphasis block. Once the windowed frame of N samples is obtained, the score function is estimated according to Eqns. 4 and 5, and then the spectral entropy is calculated from Eqn. 3. The final decision (voiced/unvoiced frame) is taken by means of a threshold. Even if more complex and better rules can be used, our purpose is only to explore the differences between the estimated entropy with the score function and without it in order to simplify the classifier block.

6 Experiments

Several experiments have been carried out in order to investigate the performance of the system. First of all, we are interested in looking at a scored frame as compared to a simple frame. Figure 2 shows a voiced signal, its estimated entropy and its estimated entropy through the score function. In this case, when the voice signal is clean (having a good SNR), we can observe a similar shape of the entropy for the original frames and the scored frames.

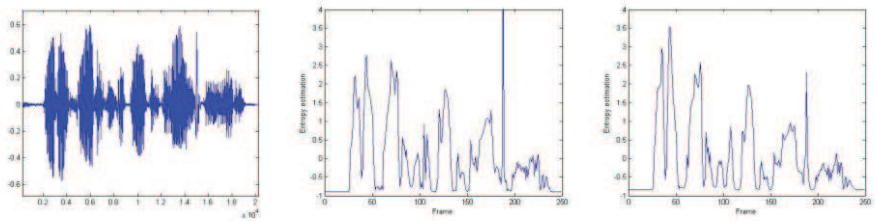

Fig. 2. Signal input (left) and the estimated entropy (without the score function in the middle, and with the score function on the right hand)

If we add a 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 the sake of clarity (bottom left), and the estimated entropy without and with the score function. Even if the noise is very high, we can observe that the entropy is different in those parts of the signal that contain speech. Of course, the difference is not as clear as in Figure 2. We can also observe that the results without and with the score function are not as similar as before. If the noise is much harder, entropy estimation does not enable us to distinguish between the noise and the speech; therefore no voice activity can be detected.

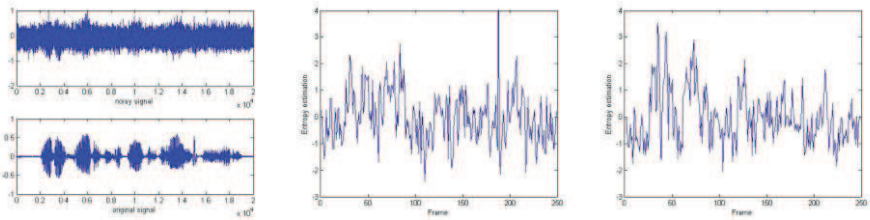

Fig. 3. Signal input (top left) with a Gaussian noise, and the estimated entropy (without the score function in the middle, and with the score function on the right hand).

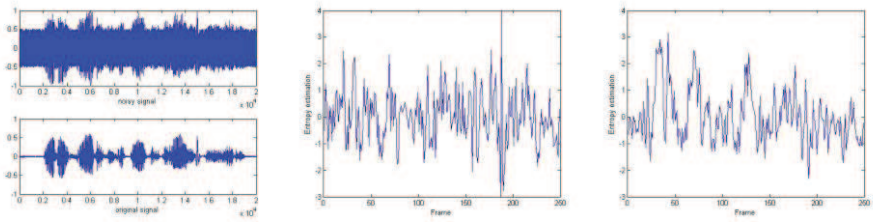

Fig. 4. Signal input (top left) with uniform noise, and estimated entropy (without score function on the middle, and with score function on the right)

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

Using a simple threshold on the estimated entropy, we can make a decision on the signal in order to decide whether or not the frame is voiced. Of course, more sophisticated procedures should be used instead of a simple trigger, as explained in the literature, but we restrict here to results using a threshold for simplicity.

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

On the other hand, if the speech signal is noisy, and since the estimated entropy is no longer equal without or with the score function, speech detection is much more complex, and different results are obtained with or without the score function. The results in this case are given in Figure 6.

One can see that the scored frames give better results and therefore the voice is better detected even if it is hidden in a noise. Of course VAD does not give perfect results, as we can see by comparison of the detection presented in Fig. 6 with the true speech signal, plotted on the bottom of the figure for clarity. As we have already mentioned before, this can be improved by designing a better classifier.

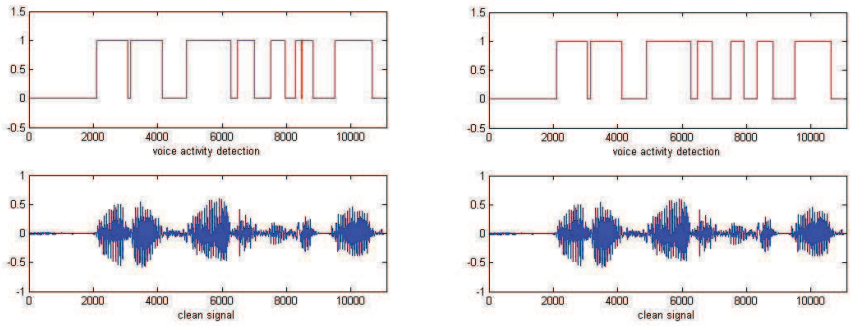

Fig. 5. Voice activity detection obtained using a simple threshold. On the left hand, estimation the entropy without the score function. On the right hand, estimation the entropy with the score function. Since the estimated entropy is essentially the same, the results are quite similar.

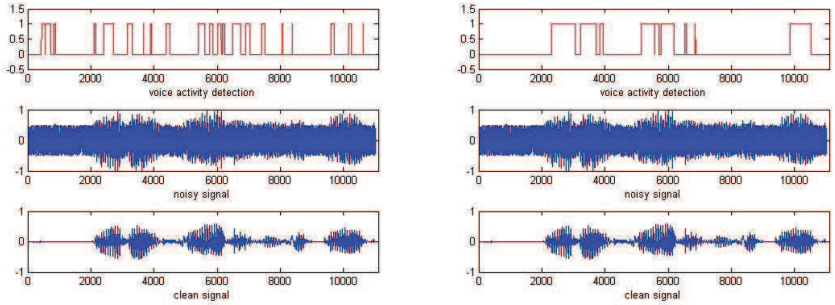

Fig. 6. Voice activity detection obtained using a simple threshold. On the left hand, estimation the entropy without the score function. On the right hand, estimation the entropy with the score function. Since the speech signal is noisy, the estimated entropy is different and therefore the detection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

7 Conclusions

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

The score function is used as a non-linear transformation estimated by means of a Gaussian kernel, and the entropy is used as a criterion to decide whether or not a frame is voiced.

If the speech signal is clean, the results are essentially the same, since the score function does not change the entropy of the signal. However, in the case of a noisy speech signal, the estimated entropy is no longer equivalent, giving therefore a different result. It is in this case where the frames pre-processed with the score function give better results and the voice can be detected in a very noisy signal.

A future work should be done for studying other non-linear transformations, in an attempt to simplify and reduce complexity of the system, as well for being implemented in real-time applications. On the other hand, the classifier should also be improved, for example, by deriving some heuristic rules, or by using more complex systems as neural networks, in order to minimise incorrect activity detections. Here, the results reported in [17-19] should be taken into account.

Acknowledgments. This work has been supported by the University of Vic under grants R0904, R0912, and by the Ministry of Science and Innovation of Spain (MICINN) under grant TEC2008-02717-E/TEC.

References

1. Gerry Altmann, *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*, Publisher: The MIT Press, USA (1995), ISBN-13: 978-0262510844
2. D. Singh and F. Boland, "Voice Activity Detection", ACM Crossroads 13.4: Computer Vision and Speech, 2007.
3. Michael Grimm and Kristian Kroschel (editors), Robust Speech Recognition and Understanding, I-Tech, Vienna, Austria (2007), ISBN 987-3-90213-08-0.
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).
17. Eun-Kyong Kim, Woo-Jin Han, Yung-Hwan Oh, "A score function of splitting band for two-band speech model", *Speech Communication* 41 (2003), 663-674.

18. Kokkinakis, K.; Nandi, A.K. “Flexible score functions for blind separation of speech signals based on generalized Gamma probability density functions”, Acoustics, Speech and Signal Processing, 2006. ICASSP Proceedings, vol. 1, 2006.
19. Tung-Hui Chiang, Yi-Chung Lin, “An integrated scoring function for a spoken dialogue system”, Signal Processing Proceedings, 1998, vol. 1, 617-620, 1998.