Homogeneity Test based Voice Activity Detection

Ouahbi Rekik, Mustapha Djeddou

الملخّـص

في هذا العمل نقترح طريقة جديدة لكشف النشاط الصوتي، ترتكز على اختبار تجانس نموذجين ذاتيي الارتداد، يمثّلان قطعتي إشارة كلاميّة، وذلك بعد حساب مسافة معيّنة. يصاغ هذا الاختبار باعتباره يمثّل اختبار فرضيّات، يتمّ فيه تحديد عتبة، وفقا لاحتمال إنذار كاذب معين. أعطت الاختبارات التي أجريت على قاعدة البيانات «Aurora» نتائج مُرضية، مقارنة مع طرق أخرى.

الكلمات المفاتيح : كشف النّشاط الصّوتيّ، نموذجان ذوا ارتداد ذاتي، اختبار فرضيات.

Homogeneity Test based Voice Activity Detection

Ouahbi Rekik, Mustapha Djeddou

Communication Systems Laboratory
Ecole Militaire polytechnique, Algiers, Algeria
rekikouahbi@gmail.com, djeddou.mustapha@gmail.com

Abstract

In this paper a new approach for voice activity detection (VAD) is proposed. This technique is based on homogeneity test of two autoregressive (AR) processes; each one models a speech window and involves the measure of a defined distance. The homogeneity test is formulated as a hypothesis test with a threshold derived analytically according to a user-defined false-alarm probability. Results using Aurora database shows the effectiveness of the proposed technique compared to other methods and standards.

Keywords: voice activity detection, homogeneity test, autoregressive process.

1. Introduction

The Voice activity detection (VAD), or commonly named speech activity detection, refers to a set of signal processing methods used to detect speech (or non-speech) segments in an audio stream [1].

When speech signals are transmitted or processed, noise is brought in inevitably. Thus, a variety of audio and speech processing applications, including speech enhancement; voice coding, speaker recognition, speech segmentation and labeling, need a VAD operation.

Systems using VAD avoid processing (coding, transmitting …) frames with non speech segments in audio stream. This is very important for systems using autonomous power supply.

In a speech segmentation process, a well-designed VAD highly improves the performance by enhancing the accuracy and by reducing the computational cost. Speakers' modeling needs only speech segments to discriminate different speakers. Furthermore, eliminating non-speech segments (noise, music, and silence zones) reduces the processing time.

The required characteristics for an efficient VAD are [2]: robustness, accuracy, adaptation, simplicity, real-time processing and no prior knowledge of noise. Among these characteristics, robustness in noisy environment has been the most difficult objective to attain.

A review of the state of the art in VAD techniques shows that a variety of algorithms has been proposed. These algorithms may be grouped into two categories [3]: algorithms of the first category use time-domain features, such as techniques based on short-time energy [4], and zero-crossing rate algorithm [5]. Algorithms of the second category are based on frequency-domain analysis of speech signal, such as the frequency band variance [6] and wavelet analysis [7].

Many other works have been carried out by focusing on signal features' fusion [8]. Looking for new VAD approaches is still of interest to many researchers [3].

The main drawback of the majority of VAD algorithms is the decrease of their performance in low signal-to-noise ratio (SNR) levels and with change of the noise source. Hence, finding robust VAD algorithms in such conditions would be of great interest.

In the present work, we propose a new VAD approach. It is based on a homogeneity test of two autoregressive (AR) processes, and involves the use of a distance as a test statistic.

This paper is organized as follows: Section 2 introduces the proposed approach and defines the procedure of the VAD. Experimental results are presented in section 3. Finally, a conclusion is given in section 4.

2. Proposed approach

In this section, we describe the procedure to detect the voice activity by using a homogeneity test of two AR processes.

Throughout this paper, *X(n)* and *Y(n)* will refer to two distinct AR processes; whereas, P_x and P_y will refer to their orders:

$$X(n) + \sum_{k=1}^{P_x} a_x(k)X(n-k) = e_x(n) \quad (1)$$

$$Y(n) + \sum_{k=1}^{P_y} a_y(k)Y(n-k) = e_y(n) \quad (2)$$

where

$[a_x(1), a_x(2), \dots, a_x(P_x)]$ and $[a_y(1), a_y(2), \dots, a_y(P_y)]$

are the AR models coefficients.

$e_x(n)$, $e_y(n)$ are two independent and identically distributed random variables, with zero means and with respective variances σ_x^2 and σ_y^2.

The main purpose for adopting AR modeling is its ability to provide the same resolution as that provided by the FFT method, but with smaller sample sizes. This makes the AR approach more advantageous, especially for real-time implementation.

An interesting characteristic of an AR process is the fact that estimating the first p + 1 autocorrelation functions, leads to entirely defining the process, i.e.: estimating the coefficients a(i) and the noise variance σ^2 . This is done by using the Yule-Walker equations:

$$\Gamma_x(l) + \sum_{i=1}^{P_x} a_x(i)\,\Gamma_x(l-i) = \sigma_x^2\delta(l,0) \quad (3)$$

where

$l = 0,1, \dots P_x$; $\Gamma_x(l)$ is the autocorrelation function of *X(n)* and $\delta(l,0)$ is the Kronecker function. A similar equation is used for *Y(n)*.

The coefficients a_x (or a_y) are estimated using the Levinson-Durbin algorithm [10]. It is a recursive method that takes advantage of the Hermitian Toeplitz structure of the autocorrelation matrix.

The AR model is determined by the model order (P_x. or P_y.) and the cients a_x(or a_y). The most used criterion for selecting the model's order is the "Minimum Description Length" (MDL) [9] defined as follows:

$$MDL(P_x) = Nlog(\sigma_x^2) + P_x\log(N) \quad (4)$$

where N is the data length and σ_x^2 is the prediction variance error associated with P_x.

After estimating the necessary parameters, they are used to calculate the power-spectrum densities $S_x(f)$ and $S_y(f)$, of *X(n)* and *Y(n)*, respectively, as follows :

$$S_x(f) = \frac{\sigma_x^2}{|P_x(f)|^2} = \frac{\sigma_x^2}{\left|1+\sum_{k=1}^{P_x} a_x(k)\exp(-j2\pi fk)\right|^2} \quad (5)$$

$$S_y(f) = \frac{\sigma_y^2}{|P_y(f)|^2} = \frac{\sigma_y^2}{\left|1+\sum_{k=1}^{P_y} a_y(k)\exp(-j2\pi fk)\right|^2} \quad (6)$$

Let $r(f)$ denote the following non-negative ratio:

$$r(f) = \frac{S_x(f)}{S_y(f)} = \frac{\sigma_x^2}{\sigma_y^2} * \frac{\left|1+\sum_{k=1}^{P_y} a_y(k)\exp(-j2\pi fk)\right|^2}{\left|1+\sum_{k=1}^{P_x} a_x(k)\exp(-j2\pi fk)\right|^2} \quad (7)$$

In order to test the homogeneity of the two AR processes *X(n)* and *Y(n)*, let us introduce the following distance, denoted $D(r)$ [11], that involves the ratio $r(f)$:

$$D(r) = log \int_{-\frac{1}{2}}^{+\frac{1}{2}} r(f)df - \int_{-\frac{1}{2}}^{+\frac{1}{2}} \log\big(r(f)\big)\, df \quad (8)$$

An interesting property of this distance is that $D(r) = 0$ *iff* $r(f) = c$ where c is an arbitrary positive constant. The following test will be based on this property.

Voice activity detection based on homogeneity test of two AR processes will be seen as a hypothesis test involving the distances $D(r)$ as its statistic:

$$\begin{cases} H_0, & D(r) = 0 \quad \text{if} \quad (r(f) = c) \\ H_1, & D(r) \neq 0 \quad \text{if} \quad (r(f) \neq 0) \end{cases} \quad (9)$$

Given two frames (segments) of the observed signal (see Figure.1), with *N* samples, and separated by *M* samples to guarantee the independence assumption of the two frames, we identify: in the first frame, an AR process, denoted by *Y(n)* and, in the second frame, another AR process, denoted by *X(n)*.

Processing the entire audio stream will be accomplished using two sliding windows. Depending on the position of the slider, the two windows can be described by one of the two cases of the binary hypothesis testing:

- Under the null hypothesis H_0, there is only noise in the first frame ;
- Under the alternative hypothesisH_1, the first frame contains speech signal with noise.

The sliding offset can be one sample or a block of samples. It is chosen so as to make a trade-off between the accuracy and the computational cost. Figure 1 illustrates such a process.

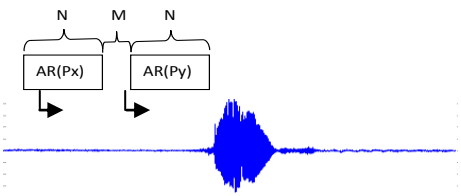


Figure 1: VAD procedure using two sliding windows

According to [11], the distribution of the statistic $D(r)$ is given by:

Under the alternative hypothesis H_1 :

$$\sqrt{N}\big(\widehat{D} - D\big) \rightarrow \mathcal{N}\ (0, v) in\ distribution \quad (10)$$

where $v =$ the variance of the normal distribution ;D = the biais of the distance; *N*= the size of the frame; $\widehat{D}$ = the estimated value of $D(r)$.

Under the null hypothesis H_0 :

$\frac{N}{2}\widehat{D} \rightarrow \mathcal{X}_p^2$ in distribution (11)

where: $\mathcal{X}_p^2$ is the *Chi-2* distribution with *p* degrees of freedom.

With these two distributions, we may use the Neyman-Pearson formalism to go through hypothesis testing. Such a test involves the choice of a significance level α (the false-alarm probability) and the estimation of the decision threshold T_α in order to detect the change points between speech and non-speech segments.

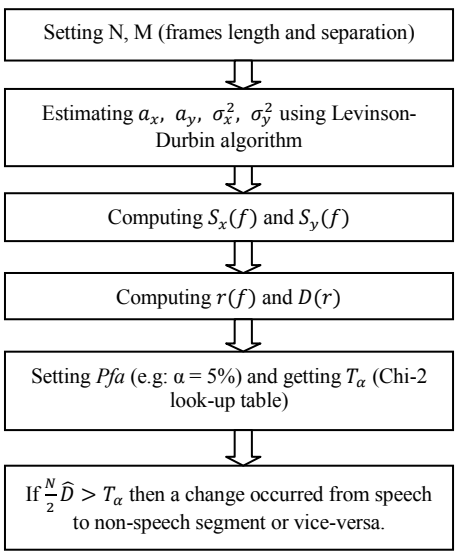
Setting the false-alarm probability to a fixed value (typically 5% or 1%), the probability of false alarm can be expressed as follows [4]:

$$P_{H0}\left(\frac{N}{2}\widehat{D} > T_\alpha\right) = \alpha \quad (12)$$

The threshold can be get using Chi-2 look-up table. Accordingly, the detection probability is given by the following expression:

$$P_d = P_{H1}\left(\frac{N}{2}\widehat{D} > T_\alpha\right) \quad (13)$$

where P_{H1} and P_{H0} denote the probability under the alternative hypothesis and the null hypothesis of the statistical test, respectively. The following steps resume the proposed approach:



It is important to note that with long frames, the estimation of AR parameters is more accurate. On the other hand, due to the non-stationary nature of the speech signal, the processing is performed on short frames. Therefore, the length of frames is chosen so that to make a trade-off between the accuracy of parameters estimation and the non-stationary nature of speech signal.

A second variant of the proposed technique may be used. In this case, only the first frame will slide along the audio stream. We assume that the second frame is fixed and contains only noise. As soon as the first frame reaches a speech segment, the quantity $\frac{N}{2}\widehat{D}$ exceeds the threshold T_α. Figure.2 illustrates this procedure:

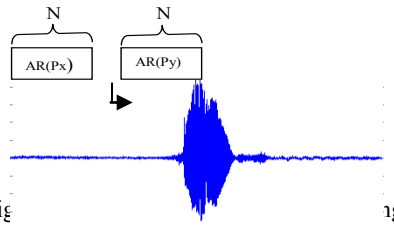


Fig ıg

window

In this case, the AR process *X(n)* is assumed to be white noise with variance σ_x^2. The ratio $r(f)$ will be expressed as:

$$(f) = \frac{S_y(f)}{S_x(f)} = \frac{\sigma_y^2}{\left|1+\sum_{k=1}^{Py} a_y(k)\exp(-j2\pi fk)\right|^2} * \frac{1}{\sigma_x^2} = \beta S_y(f) \quad (14)$$

where β replaces the quantity $\frac{1}{S_x(f)}$

An important property of the distance $D(r)$ is given by [11]:

$$D(\lambda r) = D(r) \quad (15)$$

where λ stands for a positive constant.

We substitute in (8), $r(f)$ previously defined in (14). The statistic now becomes:

$$D(r) = log \int_{-\frac{1}{2}}^{+\frac{1}{2}} S_y(f)df - \int_{-\frac{1}{2}}^{+\frac{1}{2}} \log\left(S_y(f)\right) df \quad (16)$$

In this case, the distribution of the statistic $D(r)$ will be as follows:

1) Under the alternative hypothesis H_1 :

$$\sqrt{N}\left(\hat{D} - D\right) \rightarrow \mathcal{N}(0, v_w) \quad \text{in distribution} \quad (17)$$

Where v_w is defined in [11].

2) under the null hypothesis H_0 :

$$N\hat{D} \rightarrow \mathcal{X}_p^2 \quad \text{in distribution} \quad (18)$$

The VAD process is faster in the second variant than in the first one.

The statistic values obtained will be filtered in order to decrease the false-alarm probability and the probability of miss. This is mainly a smoothing procedure. First we set a minimum period of time for both speech and silence segments, then, we eliminate segments misjudged. The final results will be binary values; where "0" stands for non-speech segments, and "1" for speech segments.

3. Experimental results

In this section we assess the performance of the proposed approach throughout several tests. We perform our experiments using Aurora database [12]. Several SNR levels are used; ranging from 25 dB to –5 dB. Test signals are divided into three noisy conditions: « Quiet », « Low » and « High ». Many sources of noise are used (train, babble, restaurant, street, airport and exhibition hall).

The obtained results are compared to that of SOHN algorithm [13], from « voicebox » toolbox, and different VAD standards such as G.729 [14], AMR1/2 [15] and AFE (FD/WF)[16].

Several VAD performance criteria are used in the literature. The most used ones are [1]:

- Non-speech Hit-Rate, denoted by *HR0*, and expressed by :

$$HR0 = \frac{N_{0,0}}{N_0^{ref}} \quad (19)$$

- Speech Hit-Rate, denoted by *HR1*, and expressed by :

$$HR1 = \frac{N_{1,1}}{N_1^{ref}} \quad (20)$$

where $N_{0,0}$ and $N_{1,1}$ are the number of detected non-speech and speech segments (frames). N_0^{ref} and N_1^{ref} are the true number, taken as reference, of non-speech and speech segments (frames). The reference labels are set manually or by using a VAD algorithm not already used in the tests. Clean signals (without noise) of the database are used to label each signal segments as speech or no-speech for reference.

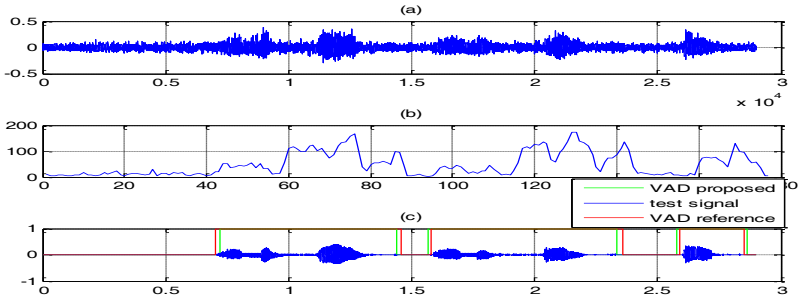


Figure 3: Proposed VAD results:
(a) signal+noise (b) D distance (c) final result

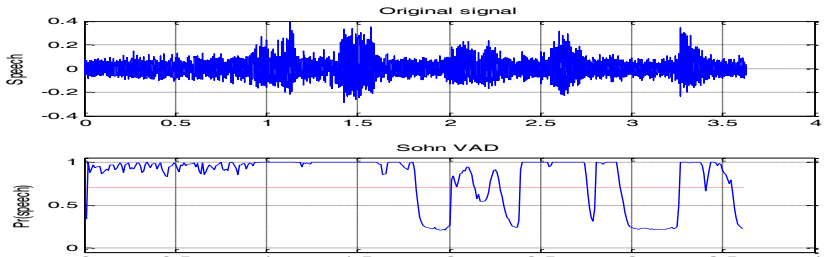


Figure 4: VAD with SOHN (threshold in red line)

Figures 3a, 3b and 3c illustrate the VAD results obtained with the proposed method (green line in Figure.3.c) and compared to a VAD reference (red line in Figure.3.c). The test signal is taken from Aurora database with an SNR level of 0 dB (high noise condition). It is clear that all speech segments are successfully detected, and hence, the proposed technique has a satisfying degree of robustness against noise.

Figure 4 illustrates the result of using SOHN algorithm. We can notice that some non speech segments, at the beginning of the signal, were detected as speech frames (with high probability of presence of speech).

Furthermore, some highly noised speech frames were detected with low probability of presence of speech. The results of evaluating the performance of the proposed method using *HR0* and *HR1* criteria are reported in Table 1.

These results reveal the effectiveness of the proposed technique when compared to other methods. The proposed technique outperforms the G729 and the AFE in all cases. For the comparison against the AMR1 VAD, our proposal gives better performance in terms of HR0 than the AMR1. However, it is slightly less effective in terms of HR1.

Method	SNR Level					
	Quiet		*Low*		*High*	
	HR0	HR1	HR0	HR1	HR0	HR1
Proposed	87	95	75	97	65	93
G729	24	80	21	66	20	70
AMR1	50	97	11	98	5	97
AFE (WF)	52	93	59	90	70	86

Table 1. HR0 AND HR1 (%)

Nonetheless, if we take the mean of the two measures, HR0 and HR1, the proposed algorithm gives better performance.

Other criteria are used to assess the performance of VAD algorithms, such as the detection probability as a function of the false-alarm probability. This is illustrated by ROC (Receiver Operating Characteristics) curves [1]. Samples of these curves, with different noise conditions are reported in Figures 5, 6 and 7.

The results depicted in figures 5, 6 and 7 show the usefulness of the proposed approach, in different noise environment (different noise sources and levels), when compared to the VAD standards: G729b, AMR and AFE. Furthermore, good compromise between false-alarm and detection probabilities is obtained.

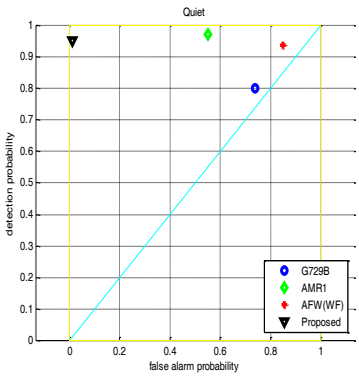


Figure 5: Pd Vs Pfa for Quiet signals

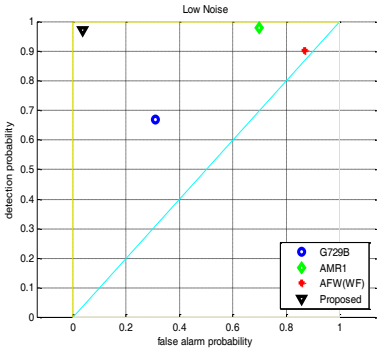


Figure 6. Pd Vs Pfa for Low-Noise signals

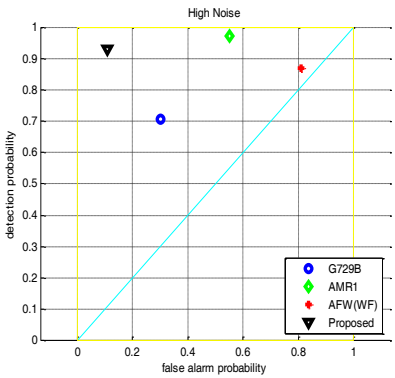


Figure 7: Pd Vs Pfa for High-Noise signals

4. Conclusion

In this paper, a new VAD approach is proposed. This approach is based on a homogeneity test of two AR processes, where each one models a sliding speech window and involves the use of spectral

distance as a test statistic. The detection threshold is set analytically; which is advantageous in such processes. Experimental results have revealed the effectiveness of this method in noisy environments. No prior knowledge of the noise is needed. Two variants are possible with different computation load. Furthermore, the sliding step can be adjusted for a more scalability and trade-off between complexity and precision, which makes the approach suitable for real-time applications.

5. References

[1] Ramirez J., Gorriz J. M. and Sergura J. C.,"Voice activity detction. Fundamentals and speech recognition system robustness in robust Speech recognition and understanding".Intech, June 2007.

[2] Moattar M. H. and Homayounpour M. M., "A simple but efficient real-time voice activity detection algorithm" 17th European Signal Processing Conference (EUSIPCO 2009).. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Wang, Y., Huang, S. and Wei Y., "A voice activity detection algorithm with sub-band detection based on time-frequency characteristics of mandarin" , 2013 6th International Congress on Image and Signal Processing (CISP 2013).

[4] Savoji M. H., "A robust algorithm for accurate endpointing of speech,"in: *Speech Commun.*, 1989,vol. 8, pp. 45-60.

[5] Wang J. F. and Chen S. H., "A voice activity detection algorithm based on perceptual wavelet packet transform and teager energy operator,"International Symposium on Chinese Spoken Language Processing,2002, pp. 177–180.

[6] Hung W.W. and Wang H.C., "On the use of weighted filter bank analysis for the derivation of robust MFCC," IEEE Signal Processing Letters, 2001, pp. 70-73.

[7] Chang J.H., Kim N.S. and Mitra S.K., "Voice activity detection based on multiple statistical models," IEEE Trans Signal Processing, 2006, pp.1965-1976.

[8] Morales-Cordovilla, J.A., Ning Ma, Sanchez, V., Carmona, J.L., Peinado, A.M. and Barker, J., "A pitch based noise estimation technique for robust speech recognition with missing data," Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.IEEE, 2011, pp. 4808–4811.

[9] Rissanen J., "Modeling by shortest data description," Automatica, vol.14, pp. 465–471, 1978.

[10] Levinson N., "The Wiener RMS (Root Mean Square) error Criterion in filter design and prediction," J Math. Phys., vol. 25, 1947, pp. 261-278.

[11] Martinez, R., Gomez, P., and Derouiche, K.,''A test of homogeneity for autoregressive processes '', Int. J. Adapt. Control Signal Process. 2002; 16:213-242.

[12] Hirsch H. and Pearce D., "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions." ISCA ITRW ASR 2000, Paris, France, September 18-20.

[13] Sohn J., Kim N. S. and Sung W., "A statistical model-based voice activity detection," IEEE Signal Process. Lett., vol. 16, no. 1, pp. 1–3, Jan. 1999.

[14] Benyassine A., Shlomot E.,.Su H.-Y. Massaloux, D, Lamblin C. and Petit J-P., "Itu-t recommendation g.729 annex b : A silence compression scheme for use with g.729 optimized for v.70 digital simultaneous voice and data applications. IEEE Communications Magazine, 35(9):64-73, September 1997.

[15] European Standard (Telecommunications series). Voice activity detector (vad) for adaptive multi-rate (amr) speech trafic channels, 1999. ETSI EN 301 708 v7.11 standard description.

[16] European standard (Telecomunicationsnseries). Transmission and quality aspect front-end feature extarction algorithm, 2007. ETSI ES 202 050 standard descriprtion.