An Efficient Bispectrum Phase Entropy-based Algorithm for VAD

J.M. Górriz, J. Ramírez, C.G. Puntonet, J.C. Segura

Department of Signal Theory
University of Granada, Granada, Spain
gorriz@ugr.es

Abstract

In this paper we propose a novel Voice Activity Detection (VAD) algorithm, based on the integrated bispectrum function (IBI), for improving Automated Speech Recognition (ASR) systems that work in noisy environments. In particular we use the combination of two features, IBI magnitude and IBI phase to formulate a robust and smoothed decision rule for speech/pause discrimination. The analysis performed on the new combined feature highlighted: i) the advantages of each individual feature, while compensating the drawback of each other, and ii) the higher ability for endpoint detection given by a lower variance of the decision function in pause/speech frames. The experiments conducted on the Spanish SpeechDat-Car database showed that the proposed algorithm outperforms ITU G.729, ETSI AMR1 and AMR2 and ETSI AFE standards as well as other recently reported VAD methods in speech/non-speech detection performance.

Index Terms: voice activity detection, clustering analysis, bispectrum function, entropy.

1. Introduction

Voice Activity Detectors (VAD) have been applied successfully to numerous applications of speech technologies (particularly in mobile communications, robust speech recognition or digital hearing aid devices), in combination with a noise reduction scheme [1]. During the last decade numerous researchers have studied different strategies for detecting speech in noise and the influence of the VAD decision on speech processing systems [2, 3, 4, 5].

Most of the algorithms for detecting the presence of speech in a noisy signal only exploit the power spectral content of the signals and require knowledge of the noise power spectral density [3, 5, 6, 7]. One of the most important disadvantages of these approaches is that no *a priori* information about the statistical properties of the signals is used. Higher order statistics methods rely on an *a priori* knowledge of the input processes and has been considered for VAD since they can distinguish between Gaussian signals (which has a vanishing bispectrum) from non-Gaussian signals. However, the main limitations of bispectrum-based techniques are that they are computationally expensive and the variance of the bispectrum estimators is much higher than that of power spectral estimators for identical data record size [8]. These problems were addressed in [9], where a computationally efficient and reduced variance statistical test, based on the *magnitude* of the IBI, for detecting speech periods was shown. On the other hand, the complexity of the proposed VADs can be additionally improved using clustering analysis on the noise subspace of subband energies [10], achieving efficient VADs for real time applications. This paper advances in the field and shows an effective VAD based on the combination of two features (bispectrum magnitude and phase) and the clustering techniques for voice activity detection. The proposed approach also incorporates contextual information to the decision rule, a strategy first proposed in [11] that has reported significant benefits, particularly, in robust speech recognition applications [12, 13]. The paper includes a carefully derivation of the decision rule based on the two features: subband bispectrum magnitude clustering, and bispectrum phase entropy.

2. Noise subspace clustering applied to bispectrum magnitude

Clustering analysis is useful tool in VAD to model the noise subspace in terms of a set of prototypes of energy subbands [10]. The noise prototypes, which are obtained from the minimizing process of the cost function (mean squared error) over a set of initial pause feature vectors, give a smoothed and low dimensional representation of the noise subspace. The presence of the cluster "speech" is detected by means of a Euclidean distance between the mean of the prototype subband energies and the current feature. In this section we define the set of prototypes in terms of the IBI magnitude since they have shown an special ability in VAD [13, 9]. Let $x(n)$ be a discrete time signal and $s(n) = x^2(n) - E[x^2(n)]$. Denote by $\mathbf{y}_{n'}$ the ensemble average of the product samples:

$$y_{n'}(\tau) = \{E[x(n+\tau)\cdot s(n)]\} = \{E[x(i+n'\cdot D+\tau)\cdot s(i+n'\cdot D)]\}; \quad (1)$$

where $i = 0 \dots L-1$, τ is the correlation lag, D is the window shift, L is the number of samples in each frame and n' selects a certain data window. Consider the set of $2 \cdot m + 1$ averages $\{y_{l-m}, \dots y_l \dots, y_{l+m}\}$ centered on average y_l, and denote by $Y(s, n')$, $n' = l-m, \dots l \dots, l+m$ its Discrete Fourier Transform (DFT) resp., that is, the IBI of $x(n)$ [14]:

$$Y_{n'}(\omega_s) \equiv Y(s, n') = \sum_{\tau=0}^{N_{FFT}-1} y_{n'}(\tau) \cdot exp\left(-\mathbf{j} \cdot \tau \cdot \omega_s\right). \quad (2)$$

where $\omega_s = \frac{2\pi \cdot s}{N_{FFT}}$, $0 \le s \le N_{FFT}-1$, N_{FFT} is DFT resolution, $\mathbf{j}$ denotes the imaginary unit and $y_{n'}(\tau)$ is the n'-th component of the vector $\mathbf{y}^{\tau}_{n'}$. The averaged IBI "energies" for each n'-th frame, $E(k, n')$, in K subbands $(k = 1, 2, ..., K)$, are computed by means of:

$$E(k, n') = \left[\frac{2K}{N_{FFT}} \sum_{s=s_k}^{s_{k+1}-1} |Y(s, n')|^2\right] \quad (3)$$
$$s_k = \left\lfloor \frac{N_{FFT}}{2K}(k-1) \right\rfloor \; k = 1, 2, ..., K$$

where an equally spaced subband assignment is used and $\lfloor \cdot \rfloor$ denotes the "floor" function. Hence, the IBI magnitude is averaged over K subbands obtaining a suitable representation of the input signal for VAD [12], the observation vector at each frame n', defined as:

$$\mathbf{E}(n') = (E(1,n'),\ldots,E(K,n'))^T \in \mathbb{R}^K \quad (4)$$

Once the feature vector is defined, the noise model is obtained by minimizing the dissimilarity measure, based on the squared Euclidean distance:

$$d(\mathbf{E}_j,\mathbf{E}_{j'}) = \sum_{k=1}^{K}(E(k,j)-E(k,j'))^2 = ||\mathbf{E}_j-\mathbf{E}_{j'}||^2 \quad (5)$$

over a set of N initial pause feature vectors, which can be defined as:

$$\begin{array}{l} J(\mathcal{C}) = \frac{1}{2}\sum_{i=1}^{C}\sum_{\mathcal{C}(j)=i}\sum_{\mathcal{C}(j')=i}||\mathbf{E}_j-\mathbf{E}_{j'}||^2 \\ = \frac{1}{2}\sum_{i=1}^{C}\sum_{\mathcal{C}(j)=i}||\mathbf{E}_j-\bar{\mathbf{E}}_i||^2 \end{array} \quad (6)$$

where $\mathcal{C}(j)=i$ denotes a many-to-one mapping, that assigns the j-th observation to the i-th prototype and

$$\begin{array}{l} \bar{\mathbf{E}}_i = (\bar{E}(1,i),\ldots,\bar{E}(K,i))^T = mean(\mathbf{E}_j); \\ \forall j, \quad \mathcal{C}(j)=i, \quad i=1,\ldots,C \end{array} \quad (7)$$

is the mean vector associated with the i-th prototype. Thus, the loss function is minimized by assigning N observations to C prototypes in such a way that within each prototype the average dissimilarity of the observations is minimized. Once convergence is reached, N K-dimensional pause frames are efficiently modeled by C K-dimensional noise prototype vectors denoted by $\bar{\mathbf{E}}_i^{opt}$, $i=1,\ldots,C$. The presence of the second "cluster" (speech frame) is detected if the following ratio holds:

$$\eta(l) = log\left[1/K\sum_{k=1}^{K}\frac{\hat{E}(k,l)}{<\bar{\mathbf{E}}_i>(k)}\right] > \gamma \quad (8)$$

where $<\bar{\mathbf{E}}_i>= 1/C\sum_{i=1}^{C}\bar{\mathbf{E}}_i = 1/C\sum_{i=1}^{C}1/N_i\sum_{j=1}^{N}\gamma_{ij}\mathbf{E}_j$ is the averaged noise prototype center, $\hat{E}(k,l) = \max\{\mathbf{E}(j)\}$, $j = l-m,\ldots,l+m$ and γ is the decision threshold. As it is shown in equation 8, the VAD decision rule is formulated over a sliding window consisting of $2m$+1 observation (feature) vectors around the frame for which the decision is being made (l). This strategy, known as "long term information"[8], provides very good results using several approaches for VAD, however it imposes an m-frame delay on the algorithm that, for several applications including robust speech recognition, is not a serious implementation obstacle.

3. A new discriminative feature: bispectrum phase entropy

The entropy, a measure of amount of expected information, is broadly used in the field of coding theory. In [15] it is used in combination with energy for endpoint detection showing that voiced spectral entropy is quite different from non-voiced one.

Shannon's entropy H_S, measures the average length of binary word per symbol under optimal coding for some information source S and it is defined:

$$H_S = -\sum_{k=1}^{M}P\left[s_k\right]\log_2\left(P\left[s_k\right]\right), \quad (9)$$

where $S = [s_1,\ldots,s_k,\ldots,s_M]$ represents the information source with M symbols and $P[s_k]$ is the probability of emission of symbols i. This causes minimum entropy to occur when one symbol has an emission probability 1 and other symbols have emission probability 0. Respectively, maximum entropy occurs when all the symbols have same probability, i.e., $s_k = 1/M$ for all i. Considering the normalized IBI phase $\widehat{Y}(s,n')$ of the frame n' as a probability distribution, the entropy in the phase domain can be computed by substituting the symbol probabilities $P[s_k]$ with probability of the sth frequency band given by:

$$P\left[\ \widehat{Y}(s,n')\ \right] = \frac{|\widehat{Y}(s,n')|}{\sum_s|\widehat{Y}(s,n')|}, \quad (10)$$

resulting the IBI phase entropy at frame n':

$$H_{\widehat{Y}}(n') = -\sum_k P\left[|\widehat{Y}(s,n')|\right]\log_2 P\left[|\widehat{Y}(s,n')|\right]. \quad (11)$$

Assuming that phase vectors are independent over the sliding window, then the overall IBI phase entropy is given by:

$$H_{\widehat{Y}} = \sum_{n'}H_{\widehat{Y}}(n') \quad (12)$$

This new feature is used in combination with equation 8 by multiplication as in [15]. It works because both "energy" and entropy has their limitations. The blind spots in either "energy" or entropy, or both, can be canceled by the multiplication (in our case by division). In other words, energy covers the case that was failed in entropy: babble and background music in speaker utterance; whereas the entropy covers the case that was failed in energy: non-stationary noise which belongs to mechanical sounds [15].

4. Remarks

Fig.1 shows the operation of the proposed algorithm the IBI-Phase Entropy (IBI-P Entropy) VAD, on an utterance of the Spanish SpeechDat-Car (SDC) database [16]. The phonetic transcription is: ["sjete", "θinko", "dos", "uno", "otSo", "sejs"]. We also include de decision functions using the same clustering scheme without the phase entropy feature (IBI-Magnitud Clustering (IBI-Mag Cl), green line) and the IBI magnitude VAD without clustering (IBI-Mag, red line). As it is shown in the bottom of this figure the accuracy in detection of word endings of the proposed VAD is higher than the other approaches. The decrease of the decision function variance in pause periods leads to a better classification of the silencie/speech frames as we will see in the following section.

In figures 2 and 3 we show the different features used for pause/speech discrimination. In particular we show the bi-frequency components of magnitude and phase before averaging. It can be clearly seen how the phase components in pause frames (see botton-right in figure 2) mean a situation of maximum entropy, as it is shown in figure 4, unlike the correlated phase components of speech frames (idem in figure 3). Thus, this new feature could be used in combination with the decision function in terms of the IBI magnitude in equation 8 as:

$$\eta_{new}(l) = \eta(l)/\bar{H}_{\widehat{Y}} \quad (13)$$

where $\bar{H}_{\widehat{Y}}$ denotes the normalized IBI-P entropy (with respect to the IBI-P entropy of a set of initial pause frames) over the sliding window of observations (equation 11).

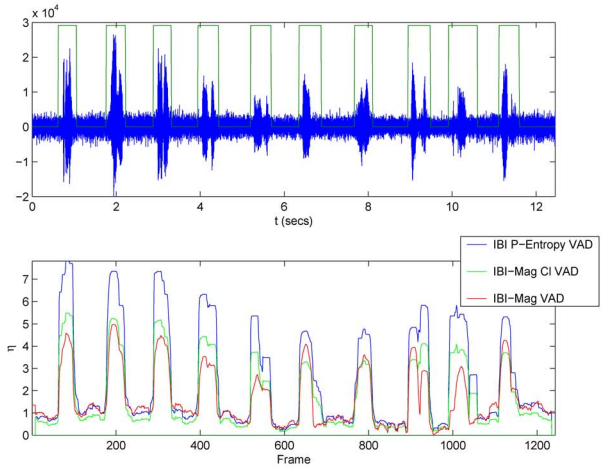

Figure 1: *Operation of the proposed VAD on a utterance of the Aurora3.*

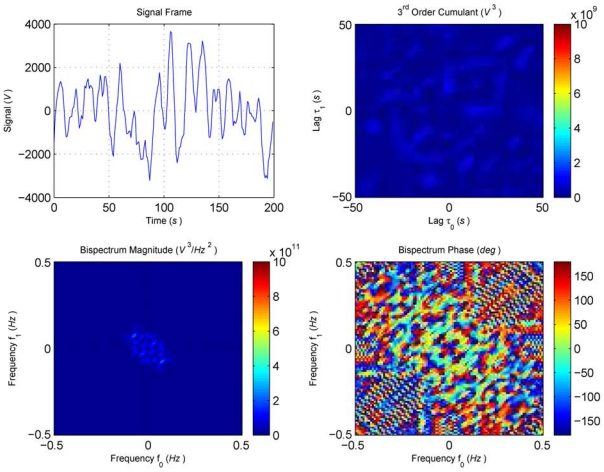

Figure 2: *Pause frame features. Bispectrum magnitude and phase*

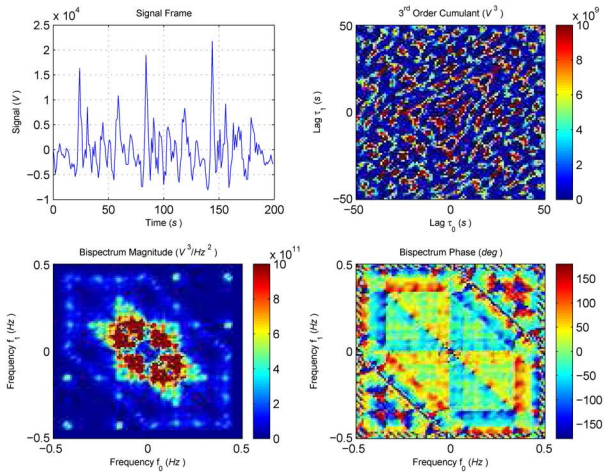

Figure 3: *Speech frame features. Bispectrum magnitude and phase*

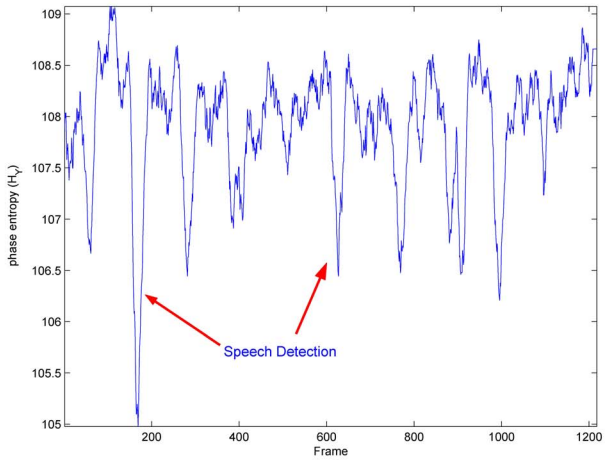

Figure 4: *IBI-phase entropy over the* $2 * m + 1$ *observations on a utterance of the Aurora3.*

5. Experimental framework

The ROC curves are frequently used to completely describe the VAD error rate. The AURORA 3 subset of the original Spanish SpeechDat-Car (SDC) database [16] was used in this analysis. The files are categorized into three noisy conditions: quiet, low noisy and highly noisy conditions, which represent different driving conditions with average SNR values between 25dB, and 5dB. The non-speech hit rate (HR0) and the false alarm rate (FAR0= 100-HR1) were determined in each noise condition. Fig. 5 shows the ROC curve of the proposed IBI-phase entropy VAD when it is defined on multiple observations (m= 8 frame delay) under the worst noise condition (5 dBs). The working points of the ITU-T G.729 [2], ETSI AMR [4] and ETSI AFE VADs [17] are also included as well as other frequently referred algorithms [7, 6, 5, 3] for recordings from the distant microphone in quiet and high noisy conditions.

The proposed VAD yields clear improvements in detection accuracy working closer to the upper left corner than any other algorithm used as a reference. The benefits are especially important over G.729 and over the Li's algorithm [6]. The IBI-P entropy VAD is more effective when multiple observations are considered. It improves Marzinzik's VAD [5], the Sohn's VAD [3], and all recently reported VADs to date for varying significance level. Fig. 5 also assesses the influence of the the combination of the two features (magnitude and phase) on the ROC curves. We observe a significant shift to the left-up corner, specially on the working area in speech recognition as we expected in the previous sections and from [15]. If IBI-phase entropy feature is not used, the clustering algorithm applied to IBI magnitude yields clear improvements over the competing algorithms also. The purpose of this new feature is to achieve a more robust detection algorithm for use in high noise acoustic environments. Thus, IBI-phase entropy leads to an additional shift-up and to the left of the ROC curve in the ROC space without additional computational complex. In addition the ROC curve without clustering is also plotted in order to highlight the improvement achieved using this technique.

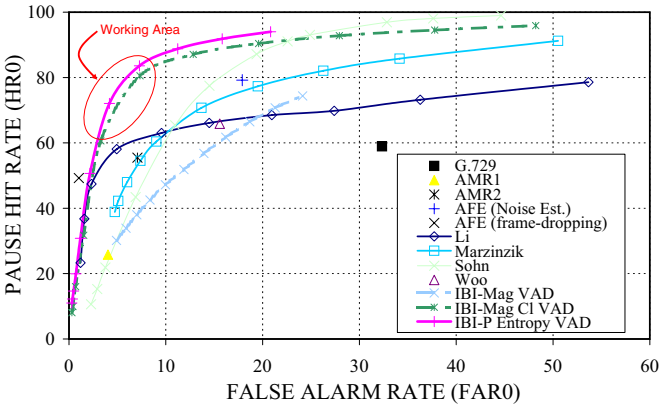

Figure 5: *ROC curves obtained for a subset of the Spanish SDC database at high noisy condition (high speed, good road, 5 dB average SNR)*

6. Conclusions

This paper presented a new technique for improving speech detection robustness in noisy environments. The approach is based on the combination of two features, i.e. the integrated bispectrum function magnitude and phase. The first of them has been used in many some algorithms for VAD [13, 9] unlike the second one. The entropy-based approach is more reliable than pure energy based methods in some cases, particularly when the non-stationary noises are mechanical sounds [15]. In other cases, when entropy becomes very unstable, energy performs well because of its additive property: energy of the sum of speech plus noise is always greater than energy of noise. Thus, the new feature possesses advantages of each individual while compensating the drawback of each other. As a result, it leads to clear improvements in speech/non-speech discrimination especially when the SNR drops. The proposed algorithm outperformed G.729, AMR and AFE standard VADs as well as recently reported approaches for endpoint detection.

7. Acknowledgements

This work has received research funding from the EU 6th Framework Programme, under contract number IST-2002-507943 (HIWIRE, Human Input that Works in Real Environments) and SESIBONN and SR3-VoIP projects (TEC2004-06096-C03-00, TEC2004-03829/TCM) from the Spanish government. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

8. References

[1] R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Communication*, vol. 16, pp. 245–254, 1995.

[2] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *ITU-T Recommendation G.729-Annex B*, 1996.

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

[4] ETSI, "Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels," *ETSI EN 301 708 Recommendation*, 1999.

[5] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.

[6] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, 2002.

[7] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[8] J.M. Górriz, J. Ramírez, J.C. Segura, and C.G. Puntonet, "Improved MO-LRT VAD based on bispectra gaussian model," *Electronics Letters*, vol. 41, no. 15, pp. 877–879, 2005.

[9] J. Ramírez, J. M. Górriz, J. C. Segura, C. G. Puntonet, and A. Rubio, "Speech/non-speech discrimination based on contextual information integrated bispectrum lrt," *IEEE Signal Processing Letters*, 2006.

[10] J. M. Górriz, J. Ramírez, C. G. Puntonet, and J. C. Segura, "An effective cluster-based model for robust speech detection and speech recognition in noisy environments," *In press in the Journal of Acoustical Society of America*, vol. XX, no. X, pp. XXX–XXX, 2006.

[11] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio, "A new adaptive long-term spectral estimation voice activity detector," in *Proc. of EUROSPEECH 2003*, Geneva, Switzerland, September 2003, pp. 3041–3044.

[12] J. Ramírez, José C. Segura, C. Benítez, A. de la Torre, and A. Rubio, "An effective subband osf-based vad with noise reduction for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, 2005.

[13] J. M. Górriz, J. Ramírez, C. G. Puntonet, and J. C. Segura, "Generalized lrt-based voice activity detector," *IEEE Signal Processing Letters*, 2006.

[14] J. K. Tugnait, "Detection of non-gaussian signals using integrated polyspectrum," *IEEE Trans. on Signal Processing*, vol. 42, no. 11, pp. 3137–3149, 1994.

[15] L. S. Huung and C. H. Yung, "A novel approach to robust speech endpoint detection in car environments," in *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, 2000, pp. 1751–1754.

[16] A. Moreno, L. Borge, D. Christoph, R. Gael, C. Khalid, E. Stephan, and A. Jeffrey, "SpeechDat-Car: A Large Speech Database for Automotive Environments," in *Proceedings of the II LREC Conference*, 2000.

[17] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES 201 108 Recommendation*, 2002.