Improved MO-LRT VAD based on bispectra Gaussian model

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

A robust algorithm for voice activity detection (VAD) is presented. It defines a likelihood ratio test (LRT) involving multiple and independent observations of the bispectra. The proposed VAD provides significant improvements in speech/pause discrimination when compared to standardised and recently reported VADs.

Introduction: Voice activity detection (VAD) remains a challenging problem in speech processing and affects a number of applications including noise reduction for digital hearing aid devices, speech recognition systems and speech coding for discontinuous speech transmission (DTX) in mobile and IP networks. During the last decade, researchers have paid attention to the study of discriminative features for classification and noise robust decision rules. A highly cited work is the VAD proposed by Sohn *et al.* [1], which is based on the evaluation of a single feature vector likelihood ratio test (LRT) and assumes a Gaussian model for the noisy signal DFT coefficients. The proposed algorithm considers a generalised LRT involving multiple and independent observations of the bispectra. The experimental analysis shows significant improvements over standardised and recently reported VAD methods.

Background: Let $\{x(t),\ t=1,\ 2,\ \ldots,\ N\}$ be a sequence of random variables with $E\{x(t)\}=0$. The third-order cumulant is defined as $C_{x_k x_l}=E\{\mathbf{x}_0\mathbf{x}_k\mathbf{x}_l\}$ where $\mathbf{x}_k$ denotes the observation vector at lag k while its 2-D discrete Fourier transform (DFT) is the bispectrum function:

$$C_{x_k x_L}(n,m)=\sum_{k=-M}^{+M}\sum_{l=-M}^{+M} C_{x_k x_l} e^{-j(\omega_n k+\omega_m l)} \quad n,m=0,1,\ldots,M-1 \tag{1}$$

where $\omega_n=2\pi n/M$. Let us define P uniformly distributed points in this grid (m, n), called a coarse grid, and form the L-point fine grid as the L nearest frequency pairs to the coarse grid points [2]. If we reorder the set of bispectrum estimates $C(n_l, n_m)$ where $l=1, 2, \ldots, L$, in the fine grid around the bifrequency pair, into a vector $\boldsymbol{\beta}_{m,l}$ where $m=1, 2, \ldots, P$ indexes the coarse grid and define P vectors $\phi_i=\{\boldsymbol{\beta}_{1,i}, \boldsymbol{\beta}_{2,i}, \ldots, \boldsymbol{\beta}_{Pi}\}$, $i=1, 2, \ldots, L$, we obtain after averaging over i, the set of P bispectrum estimates $\{Y_1, Y_2, \ldots, Y_P\}$ which are an approximately uncorrelated and unbiased complex Gausssian vector on the coarse grid [2]. Fig. 1 illustrates the differences that appear in the third-order cumulants and bispectra of speech and noise and how they are used for speech/pause discrimination.

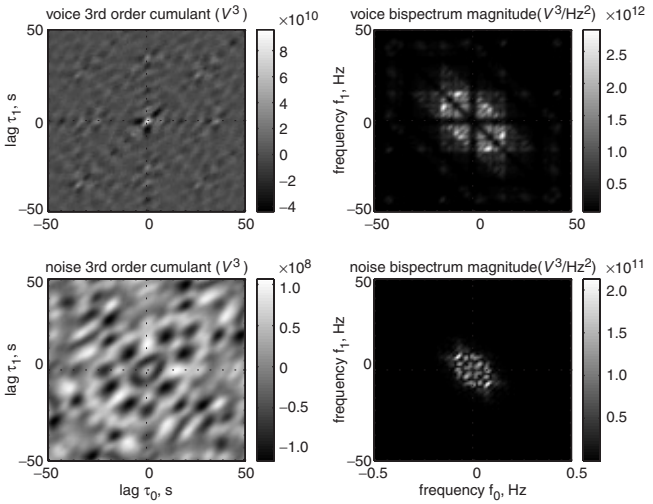

Fig. 1 *Third-order cumulants and bispectra of speech and noise*

Voice activity detection: Recently, a generalisation of the well known VAD proposed by Sohn *et al.*, called MO-LRT [3], has been proposed. This method formulates the decision rule over a sliding window of multiple and independent observation vectors. The benefits of this approach are: (i) the optimal behaviour of the decision rule, and (ii) a multiple observation vector for classification defines a reduced variance LRT reporting clear improvements in robustness against the acoustic noise present in the environment. In this Letter we perform the test in the bispectrum domain where the Gaussian model better represents the observation vectors. The so-called BLRT (bispectrum LRT) is defined by means of an LRT defined over $2m+1$ consecutive observation vectors $\{\hat{\mathbf{y}}_{l-m},\ \ldots,\ \hat{\mathbf{y}}_{l-1},\ \hat{\mathbf{y}}_l,\ \hat{\mathbf{y}}_{l+1},\ \ldots,\ \hat{\mathbf{y}}_{l-m}\}$:

$$\ell_{l,m}=\sum_{k=l-m}^{l+m}\ln\frac{P_{\mathrm{y}_k|H_1}(\hat{\mathbf{y}}_k|H_1)}{P_{\mathrm{y}_k|H_0}(\hat{\mathbf{y}}_k|H_0)} \tag{2}$$

where l denotes the frame being classified as speech (H_1) or non-speech (H_0). Note that $\ell_{l,m}$ can be computed recursively: $\ell_{l+1,m}=\ell_{l,m}-\Phi(l-m)+\Phi(l+m+1)$, where:

$$\Phi(k)=\ln\frac{p_{y_k|H_1}(\hat{\mathbf{y}}_k|H_1)}{p_{y_k|H_0}(\hat{\mathbf{y}}_k|H_0)} \tag{3}$$

Thus, if $\ell_{l,m}$ is greater than a fixed threshold η, the current frame l is classified as speech, otherwise it is classified as non-speech. In order to evaluate the proposed BLRT $\ell_{l,m}$ on an incoming signal, an adequate statistical model for the feature vectors in the presence and absence of speech needs to be selected. The model selected is similar to that used by Sohn *et al.* [1] that assumes the DFT coefficients of the clean speech (S_j) and the noise (N_j) to be asymptotically independent Gaussian random variables. In our algorithm, we work with the bispectrum coefficients instead:

$$\begin{aligned} p_{\mathrm{y}|\mathrm{H}_0}(\hat{\mathbf{y}}|\mathrm{H}_0)&=\prod_{j=0}^{p-1}\frac{1}{\pi\lambda_N(j)}\exp\left\{-\frac{|Y_j|^2}{\lambda_N(j)}\right\}\\ p_{\mathrm{y}|\mathrm{H}_1}(\hat{\mathbf{y}}|\mathrm{H}_1)&=\prod_{j=0}^{p-1}\frac{1}{\pi[\lambda_N(j)+\lambda_S(j)]}\exp\left\{-\frac{|Y_j|^2}{\lambda_N(j)+\lambda_S(j)}\right\}\end{aligned} \tag{4}$$

where Y_j represent the noisy speech bispectrum coefficients, and $\lambda_N(j)$ and $\lambda_S(j)$ denote the variances of the bispectrum function of N_j and S_j, respectively. Thus, evaluating the log-LRT and averaging leads to:

$$\Phi(k)=\frac{1}{P}\sum_{j=0}^{p-1}\left[\frac{\gamma_{k,j}\xi_{k,j}}{1+\xi_{k,j}}-\log(1+\xi_{k,j})\right] \tag{5}$$

where $\gamma_{k,j}$ and $\xi_{k,j}$ are the *a priori* and *a posteriori* bispectrum SNRs, which are estimated using the Ephraim and Malah minimum mean-square error (MMSE) estimator. Note that, $\gamma_{k,j}$ and $\xi_{k,j}$ have to be computed m frames in advance. This fact imposes an m-frame delay on the algorithm that, for several applications including robust speech recognition, is not a serious implementation obstacle. It is worthwhile clarifying that, unless the frames do not overlap and the signal and noise are white, the successive observations may not be independent. However, the independence assumption enables modelling the joint probability distribution of the observations more easily [3] and it is guaranteed by the bispectrum properties as shown in [4].

Results: The ROC curves are used in this Section for the evaluation of the proposed VAD. These plots describe completely the VAD error rate and show the trade-off between the speech and non-speech error probabilities as the threshold η varies. The Spanish SpeechDat-Car database was used in the analysis. This database contains recordings in a car environment from close-talking and hands-free microphones. Utterances from the close-talking device with an average SNR of about 25 dB were labelled as speech or non-speech for reference while the VAD was evaluated on the hands-free microphone. Thus, the speech and non-speech hit rates (HR_1, HR_0) were determined as a function of the decision threshold η for each of the VAD tested. Fig. 2 shows the ROC curves in the most unfavourable conditions (high-speed, good road) with a 5 dB average SNR. It is shown that increasing the number of observation vectors m improves the performance of the proposed BLRT VAD. This is motivated by a shift up and to the left of the ROC curve which enables working with improved speech and non-speech hit rates. The best results are obtained for $m=8$ while increasing the number of observations over this value reports no additional improvements. The proposed

VAD outperforms the Sohn's VAD [1], which assumes a single observation in the decision rule together with an HMM-based hangover mechanism, as well as standardised VADs such as G.729 and AMR [5, 6]. Fig. 3 compares the proposed BLRT VAD to our previous work MO-LRT [3], which applies an LRT assuming a Gaussian model for the noisy speech DFT coefficients, and recently reported methods [1, 7–9]. Thus, the proposed VAD works with improved speech/non-speech hit rates when compared to the most relevant algorithms to date.

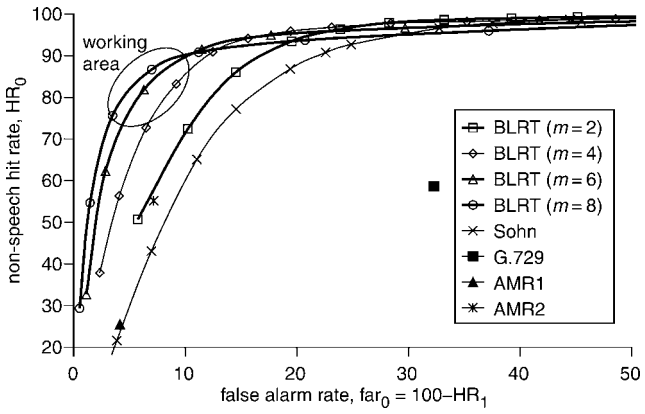

Fig. 2 *Influence of number of feature vectors m on ROC curves*

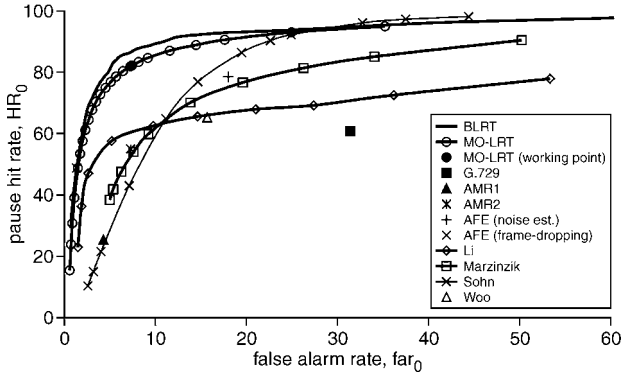

Fig. 3 *ROC curves of proposed BLRT VAD and comparison to standard and recently reported VADs*

Conclusion: A new VAD for improving speech detection robustness in noisy environments is proposed. The proposed BLRT VAD is defined as a generalisation of an LRT that considers multiple and independent observations of the bispectra. The proposed BLRT VAD outperformed MO-LRT VAD, that uses a similar LRT model of the DFT spectrum, Sohn's VAD, that defines the LRT on single observation, and other methods including the standardised G.729, AMR and AFE VADs, in addition to recently reported VADs.

Acknowledgments: This work has been funded by the European Commission (HIWIRE, IST No. 507943) and the Spanish MEC project TEC2004-03829/FEDER.

13 May 2005
Electronics Letters online no: 20051761
doi: 10.1049/el:20051761

J.M. Górriz, J. Ramírez and J.C. Segura (*Dpto. de Teoría de la Señal, Telemática y Communicaciones, Periodista Daniel Saucedo Aranda, 1871 Granada, Spain*)

C.G. Puntonet (*Dpto. Arquitectura y Tecnología de Computadores, Periodista Daniel Saucedo Aranda, 18071 Granada, Spain*)

E-mail: javierrp@ugr.es

References

1 Sohn, J., Kim, N.S., and Sung, W.: 'A statistical model-based voice activity detection', *IEEE Signal Process. Lett.*, 1999, **16**, (1), pp. 1–3
2 Subba-Rao, T.: 'A test for linearity of stationary time series', *J. Time Ser. Anal.*, 1982, **1**, pp. 145–158
3 Ramírez, J., *et al.*: 'Statistical voice activity detection using a multiple observation likelihood ratio test', to appear in *IEEE Signal Process. Lett.*, 2005
4 Subba-Rao, T., and Gabr, M.: 'An introduction to bispectral analysis and bilinear time series models' (Springer-Verlag, 1984)
5 ITU-T Recommendation G.729-Annex B, A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70, 1996
6 ETSI EN 301 708 Recommendation, Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels, 1999
7 Woo, K., *et al.*: 'Robust voice activity detection algorithm for estimating noise spectrum', *Electron. Lett.*, 2000, **36**, (2), pp. 180–181
8 Li, Q., *et al.*: 'Robust endpoint detection and energy normalization for real-time speech and speaker recognition', *IEEE Trans. Speech Audio Process.*, 2002, **10**, (3), pp. 146–157
9 Marzinzik, M., and Kollmeier, B.: 'Speech pause detection for noise spectrum estimation by tracking power envelope dynamics', *IEEE Trans. Speech Audio Process.*, 2002, **10**, (2), pp. 109–118