Independent Component Analysis Applied to Voice Activity Detection

J.M. Górriz[1], J. Ramírez[1], C.G. Puntonet[3],
E.W. Lang[3], and K. Stadlthanner[3]

[1] Dpt. Signal Theory, Networking and communications, University of Granada, Spain
gorriz@ugr.es
http://www.ugr.es/~gorriz
[2] Dpt. Computer Architecture and Technology, University of Granada, Spain
[3] AG Neuro- und Bioinformatik, Universität Regensburg, Deutschland

Abstract. In this paper we present the first application of Independent Component Analysis (ICA) to Voice Activity Detection (VAD). The accuracy of a multiple observation-likelihood ratio test (MO-LRT) VAD is improved by transforming the set of observations to a new set of independent components. Clear improvements in speech/non-speech discrimination accuracy for low false alarm rate demonstrate the effectiveness of the proposed VAD. It is shown that the use of this new set leads to a better separation of the speech and noise distributions, thus allowing a more effective discrimination and a tradeoff between complexity and performance. The algorithm is optimum in those scenarios where the loss of speech frames could be unacceptable, causing a system failure. The experimental analysis carried out on the AURORA 3 databases and tasks provides an extensive performance evaluation together with an exhaustive comparison to the standard VADs such as ITU G.729, GSM AMR and ETSI AFE for distributed speech recognition (DSR), and other recently reported VADs.

1 Introduction

The demands of modern applications of speech communication are related to the need for increasing levels of performance in noise adverse environments. The new voice services including discontinuous speech transmission [1] or distributed speech recognition (DSR) over wireless and IP networks [2] are examples of such applications. These systems often require a noise reduction scheme working in combination with a precise VAD in order to palliate the harmful effect of the acoustic environment on the speech signal. Thus, numerous researchers have studied different strategies for detecting speech in noise and the influence of the VAD on the performance of speech processing systems [3, 4, 5, 6, 7]. Recently, an improved VAD using a long-term LRT test defined on the Bispectra coefficients [8] has shown significant improvements of the decision rule but with high computational cost. The latter VAD is based on a MO-LRT over a set of averaged bispectrum coefficients $\mathbf{y}_k$ which are assumed to be independent. This approach

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part I, LNCS 3991, pp. 234–241, 2006.

is also assumed in the Fourier domain [7]. We use the latter algorithm in conjunction with the recursive PCA algorithm presented in [9] to build an efficient "on line" VAD, assessing its performance in an HMM-based speech recognition system.

The rest of the paper is organized as follows. Section 2 presents the Blind Source Separation (BSS) problem when dealing with Gaussian distributions. In section 3 we review the theoretical background related to MO-LRT applied to VAD showing the proposed signal model and analyzing the motivations for the proposed algorithm by showing the speech/non-speech correlation plots. Section 4 introduce a variation on the recursive PCA for the evaluation of the decision rule. Section 5 describes the experimental framework considered for the evaluation of the proposed statistical decision algorithm. Finally, in section we state some conclusions.

2 BSS in Gaussian Scenario

Let us denote by $\mathbf{x} = (x_1, \ldots, x_m)$ a zero m-dimensional random variable that can be observed and $\mathbf{s} = (s_1, \ldots, s_m)$ its m-dimensional statistically independent transform satisfying that:

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{1}$$

where $\mathbf{W}$ is a constant (weight) square matrix. The BSS problem [10, 11] is to determine the previous matrix extracting the independent features s_i, $i = 1, \ldots, m$ and assuming $\mathbf{W}$ is constant. There are some ambiguities in the determination of the linear model (i.e. variances and order of independent components) but fortunately they are insignificant in most applications.

The multivariate Gaussian probability density function (pdf) of an $m \times 1$ random variable vector $\mathbf{x}$ is defined as:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} \det^{1/2}(\mathbf{R}(\mathbf{x}))} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{R}(\mathbf{x})^{-1}(\mathbf{x}-\mu)\right] \tag{2}$$

where μ is the mean vector and $\mathbf{R}(\mathbf{x})$ is the covariance matrix. It is assumed that $\mathbf{R}(\mathbf{x})$ is positive definite and hence $\mathbf{R}(\mathbf{x})^{-1}$ exists. The mean vector is defined as $[\mu]_i = E(x_i), \quad i = 1, \ldots m$ and the covariance matrix as $\mathbf{R}(\mathbf{x}) = E((\mathbf{x} - E(\mathbf{x}))(x - E(\mathbf{x}))^T))$. Uncorrelated jointly Gaussian variables (i.e. components of vector $\mathbf{x}$ in equation 2) satisfy that:

$$p(\mathbf{x}) = \prod_{i=1}^{m} p(x_i) \tag{3}$$

since its covariance matrix is diagonal. That is, they are statistically independent, then decorrelation and statistical independence are equivalent for jointly Gaussian variables. In VAD applications, the DFT coefficients of the incoming signal are usually assumed to be Gaussian independent somehow, using the model of overlapped MO-window (see section 3), they are not.

Principal Component Analysis (PCA) is a very efficient and popular tool for extracting uncorrelated components from a set of signals. PCA tries to find a linear transformation (Karhunen-Loéve) $\widetilde{\mathbf{s}} = \mathbf{W}^T\mathbf{x}$ into a new orthogonal basis (columns of $\mathbf{W}$) such that the covariance matrix in the new system is diagonal. Since $\mathbf{R}(\mathbf{x})$ is symmetric we can find $\mathbf{W}$ such that:

$$\mathbf{W}^T\mathbf{R}(\mathbf{x})\mathbf{W} = \mathbf{\Lambda} \tag{4}$$

Hence the PCA decorrelates the vector $\mathbf{x}$ also achieving statistical independence when it is multivariate Gaussian distributed[1].

3 PCA-MO-Likelihood Ratio Test

In a two hypothesis test (ω_0 =noise & ω_1 =speech in noise), the optimal decision rule that minimizes the error probability is the Bayes classifier. Given an J-dimensional observation vector $\hat{\mathbf{x}}$ to be classified, the problem is reduced to selecting the class (ω_0 or ω_1) with the largest posterior probability $\mathrm{P}(\omega_i|\hat{\mathbf{x}})$. From the Bayes rule:

$$L(\hat{\mathbf{x}}) = \frac{p_{\mathbf{x}|\omega_1}(\hat{\mathbf{x}}|\omega_1)}{p_{\mathbf{x}|\omega_0}(\hat{\mathbf{x}}|\omega_0)} \begin{array}{c} > \\ < \end{array} \frac{P[\omega_0]}{P[\omega_1]} \Rightarrow \begin{array}{l} \hat{\mathbf{x}} \leftrightarrow \omega_1 \\ \hat{\mathbf{x}} \leftrightarrow \omega_0 \end{array} \tag{5}$$

In the LRT, it is assumed that the number of observations is fixed and represented by a vector $\hat{\mathbf{x}}$. The performance of the decision procedure can be improved by incorporating more observations to the statistical test. When $M = 2m + 1$ measurements $\hat{\mathbf{x}}_{-m}$, $\hat{\mathbf{x}}_{-m+1}$, ..., $\hat{\mathbf{x}}_m$ are available in a two-class classification problem, a MO-LRT can be defined by:

$$L_M(\hat{\mathbf{x}}_{-m}, \hat{\mathbf{x}}_{-m+1}, ..., \hat{\mathbf{x}}_m) = \frac{p_{\mathbf{x}_{-m},\mathbf{x}_{-m+1},...,\mathbf{x}_m|\omega_1}(\hat{\mathbf{x}}_{-m}, \hat{\mathbf{x}}_{-m+1}, ..., \hat{\mathbf{x}}_m|\omega_1)}{p_{\mathbf{x}_{-m},\mathbf{x}_{-m+1},...,\mathbf{x}_m|\omega_0}(\hat{\mathbf{x}}_{-m}, \hat{\mathbf{x}}_{-m+1}, ..., \hat{\mathbf{x}}_m|\omega_0)} \tag{6}$$

In order to evaluate the proposed MO-LRT VAD on an incoming signal, an adequate statistical model for the feature vectors in presence and absence of speech needs to be selected. The model selected is similar to that used by Sohn *et al.* [3] 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 case, the decision rule is formulated over a sliding window consisting of $2m + 1$ observation vectors around the frame for which the decision is being made (see figure 1), then we have to assume they are at least jointly Gaussian distributed as in equation 2.

The PCA algorithm presented in the next section decorrelates the observed signals $\hat{\mathbf{x}}_k$ into a set of independent signal $\hat{\mathbf{s}}_\mathbf{k}$ hence the MO-LRT in equation 6 can be expressed as:

$$\ell_{l,m} = \sum_{k=l-m}^{l+m} \ln \frac{p_{\mathbf{s}_k|\omega_1}(\hat{\mathbf{s}}_k|\omega_1)}{p_{\mathbf{s}_k|\omega_0}(\hat{\mathbf{s}}_k|\omega_0)} \tag{7}$$

[1] Observe how any transformation of the kind $\widetilde{\mathbf{x}} = (\mathbf{VW})^\mathbf{T}\mathbf{x}$ where $\mathbf{V}$ is a orthogonal matrix ($\mathbf{VV}^T = \mathbf{V}^T\mathbf{V} = I$) yields the same result.

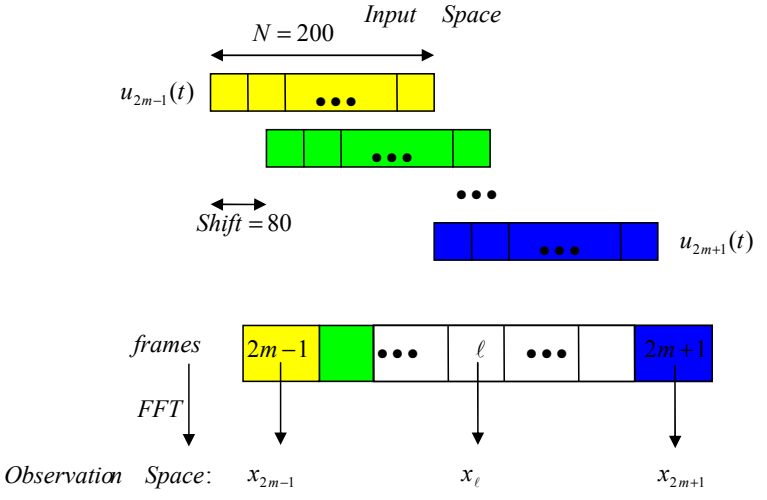


Fig. 1. Signal Model used in the PCA-MO-LRT. Observe how the use of overlapped windows introduces some correlation in the Observation Space then the assumption of statistical independence is not appropriate.

where l denotes the frame being classified as speech (ω_1) or non-speech (ω_0) and the pdf of the observations can be computed using:

$$\begin{aligned} p(\hat{\mathbf{s}}|\omega_0) &= \prod_{j=0}^{J-1} \frac{1}{\pi\lambda_N(j)} \exp\left\{-\frac{|s_j|^2}{\lambda_N(j)}\right\} \\ p(\hat{\mathbf{s}}|\omega_1) &= \prod_{j=0}^{J-1} \frac{1}{\pi[\lambda_N(j)+\lambda_S(j)]} \exp\left\{-\frac{|s_j|^2}{\lambda_N(j)+\lambda_S(j)}\right\} \end{aligned} \tag{8}$$

where s_j represents the uncorrelated noisy speech DFT coefficients, J is the DFT resolution and $\lambda_N(j)$ and $\lambda_S(j)$ denote the variances of N_j and S_j, respectively.

By defining: $\Phi(k) = \ln \frac{p_{\mathbf{s}_k|\omega_1}(\hat{\mathbf{s}}_k|\omega_1)}{p_{\mathbf{s}_k|\omega_0}(\hat{\mathbf{s}}_k|\omega_0)}$, the LRT can be recursively computed:

$$\ell_{l+1,m} = \ell_{l,m} - \Phi(l-m) + \Phi(l+m+1) \tag{9}$$

and the decision rule is defined by:

$$\ell_{l,m} \begin{array}{ll} \geq \eta & \text{frame } l \text{ is classified as speech} \\ < \eta & \text{frame } l \text{ is classified as non - speech} \end{array} \tag{10}$$

where η is the decision threshold which is experimentally tuned for the best trade-off between speech and non-speech classification errors.

4 Recursive PCA Applied to VAD

In order to recursively evaluate the LRT over the set of uncorrelated signals in the current frame l we use a result in [9]. In the frame $l+1$ the PCA components

for the MO-window $l-m,\ldots,l+m$ are computed as a function of the previous MO-window centered at frame l. Since:

$$\mathbf{R}_M = \frac{1}{M}\sum_{i=1}^{M}\mathbf{x}_i\mathbf{x}_i^T = \frac{M-1}{M}\mathbf{R}_{M-1} + \frac{1}{M}\mathbf{x}_M\mathbf{x}_M^T \tag{11}$$

where M denotes the number of observation ($M = 2m+1$), we obtain the following recursive formula for the eigenvectors and eigenvalues:

$$\mathbf{Q}_M M\mathbf{\Lambda}_M\mathbf{Q}_M^T = \mathbf{Q}_{M-1}[(M-1)\mathbf{\Lambda}_{M-1} + \alpha_M\alpha_M^T]\mathbf{Q}_M^T \tag{12}$$

where $\mathbf{R}_M = \mathbf{Q}_M\mathbf{\Lambda}_M\mathbf{Q}_M^T$ and $\alpha_M = \mathbf{Q}_{M-1}^T\mathbf{x}_M$. Using a matrix perturbation analysis approach of a matrix in the form $(\mathbf{\Lambda}+\alpha\alpha^T)$, we can obtain a recursion for the eigenvalues and eigenvectors as [9]:

$$\begin{aligned}\mathbf{Q}_M &= [\mathbf{Q}_{M-1}(\mathbf{I}+\mathbf{P}_\mathbf{V})]\mathbf{T}_M \\ \mathbf{\Lambda}_M &= [(1-\lambda_M)\mathbf{\Lambda}_{M-1} + \mathbf{P}_\mathbf{\Lambda}]\mathbf{T}_M^{-2}\end{aligned} \tag{13}$$

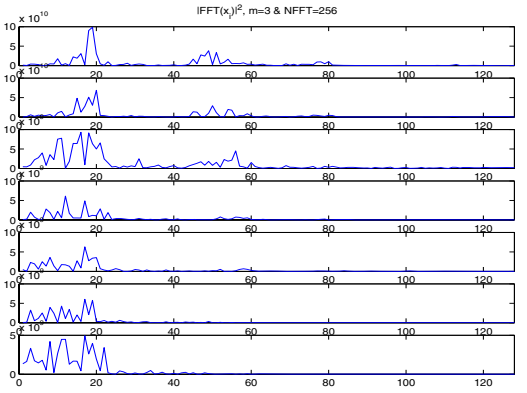
where $\mathbf{T}_M$ is a diagonal matrix containing the inverses of the norms of each column of the matrix in brackets (top); $\mathbf{P}_\mathbf{V}$ is an antisymmetric matrix whose $(i,j)^{th}$ entry is $\alpha_i\alpha_j/(\lambda_j+\alpha_j^2-\lambda_i-\alpha_j^2)$ if $j\neq i$, and 0 if $j=i$; $\mathbf{P}_\mathbf{\Lambda}$ is a diagonal matrix whose i^{th} diagonal entry is α_i^2; and λ_M is a memory depth parameter. Using the set of equations 13 we can obtain the uncorrelated components of the decision frame l from the decision frame $l-1$ in a two step procedure: Let $\mathbf{R}_{M,l}$ denote the covariance matrix on the decision frame l of order $M = 2m+1$.

1. From equation 13 obtain $\mathbf{R}_{M-1,l}$ using $\mathbf{R}_{M,l}$ and $\alpha = \alpha_{l-m}$.
2. From equation 13 obtain $\mathbf{R}_{M,l+1}$ using $\mathbf{R}_{M-1,l}$ and $\alpha = \alpha_{l+m}$.

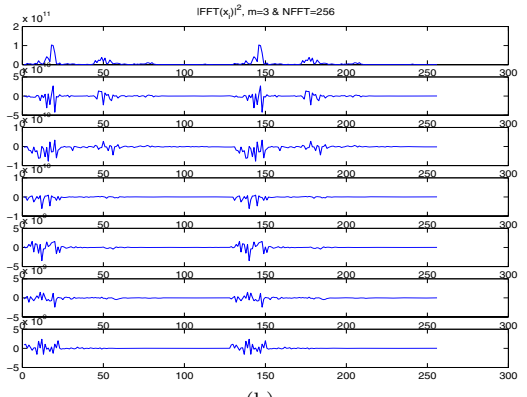
With the aim of this recursion, the proposed VAD is computationally efficient enough to be used on a real time application and eludes the iterative eigenvector decomposition in each MO-window.

5 Experimental Framework

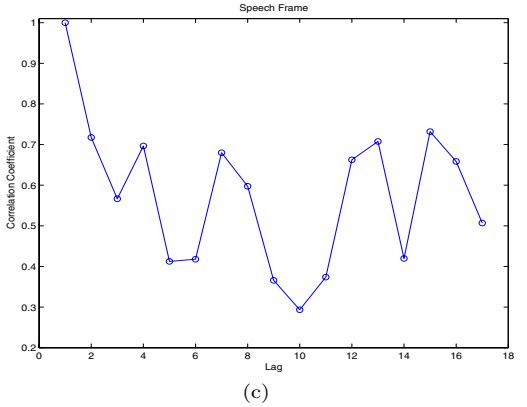
The ROC curves are used in this section for the evaluation of the proposed VAD. These plots completely describe the VAD error rate and show the trade-off between the speech and non-speech error probabilities as the threshold varies. The Aurora 3 Spanish SpeechDat-Car database was used in the analysis. 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 labeled 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 (HR1, HR0) were determined as a function of the decision threshold for each of the VAD tested. In figure 2 we observe an example of the VAD operation showing the components extracted by the recursive PCA and the correlation plot



(a)



(b)



(c)

Fig. 2. Example of VAD operation for $m = 3$ and resolution $J = NFFT = 256$ (a) Set of observed DFT signals ($2m + 1 = 7$ windows) on speech frames. (b) Independent DFT signals on speech frames. (c) The correlation plot of the DFT observation vectors over the set of overlapped windows (m=8) reveals the non suitability of the previously proposed models (they are not independent as they are correlated).

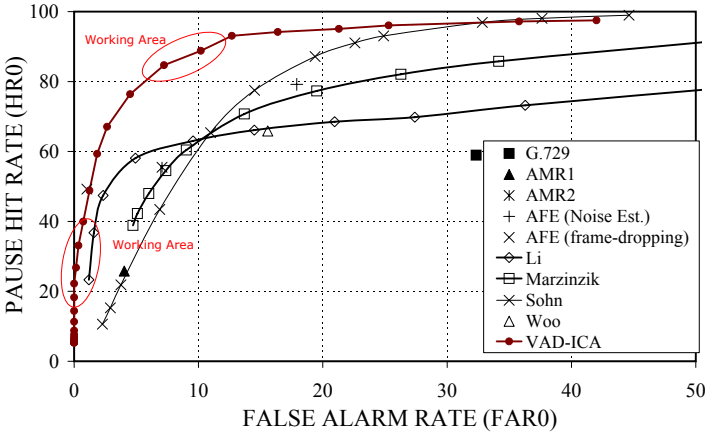


Fig. 3. ROC curves of the proposed BLRT VAD and comparison to standard and recently reported VADs

over the set of overlapped windows which shows that the independent assumption cannot be made.

Fig. 3 shows the ROC curves in the most unfavourable conditions (high-speed, good road) with a 5 dB average SNR. It is shown that the ICA-VAD obtains very good performance at low FAR0 (applications designed for speech detection) and at the common working areas ($FAR0 < 10$ and $HR0 > 80$). 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. In the middle area the VAD reproduces the same accuracy as the previous work MO-LRT [7] using a completely different set of input signals (decorrelated). The improved detection rate over the latter method is major in such applications. The proposed VAD outperforms the Sohn's VAD [3], which assumes a single observation in the decision rule together with an HMM-based hangover mechanism, as well as standardized VADs such as G.729 and AMR and recently reported methods [5, 6, 4]. Thus, the proposed VAD works with improved speech/non-speech hit-rates when compared to the most relevant algorithms to date.

6 Conclusion

In this paper we have introduced the concept of Blind Source Separation in VAD applications. We have shown that PCA can be used as a preprocessing step in this scenario for improving the accuracy of LRT based VADs. The proposed algorithm is optimum in those scenarios ($FAR0 \rightarrow 0$) where the loss of speech frames could be unacceptable, causing a system failure. The VADs have been employed as a part of Speech Recognition Systems (i.e. ASR Systems) in the last years providing significant benefits. Hence the proposed VAD can be used

to obtain relevant improvements over all standardized VADs in speech/pause detection accuracy and recognition performance.

Acknowledgements

This work has received research funding from the EU 6^{th} Framework Programme, under contract number IST-2002-507943 (HIWIRE, Human Input that Works in Real Environments) and SESIBONN project (TEC2004-06096-C03-00) 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.

References

1. ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," *ITU-T Recommendation G.729-Annex B*, 1996.
2. ETSI, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," *ETSI ES 201 108 Recommendation*, 2000.
3. J. Sohn and et al., "A statistical model-based vad," *IEEE Signal Proccessing Letters*, vol. 16, no. 1, pp. 1–3, 1999.
4. M. Marzinzik and et al., "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.
5. K. Woo and et al., "Robust vad algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.
6. Q. Li and et al., "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. on Speech and Audio Proccessing*, vol. 10, no. 3, pp. 146–157, 2002.
7. J. Ramírez and et. al., "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, 2005.
8. J. M. Górriz, J. Ramirez, J. C. Segura, and C. G. Puntonet, "An improved mo-lrt vad based on a bispectra gaussian model," *IEE Electronic Letters*, vol. 41, no. 15, pp. 877–879, 2005.
9. D. Erdogmus, Y. Rao, H. Peddaneni, A. Hegde, and J. Principe, "Recursive principal components analysis using eigenvector matrix perturbation," *EURASIP Journal on Applied Signal Processing*, vol. 13, pp. 2034–2041, 2004.
10. P. Comon, "Independent component analysis—a new concept?" *Signal Processing*, vol. 36, pp. 287–314, 1994.
11. A. J. Bell and T. J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.