VOICE ACTIVITY DETECTION USING A CONTEXTUAL INFORMATION AND MULTIPLE HYPOTHESIS TESTING

J. Ramírez, J. C. Segura, J. M. Górriz, A. de la Torre, L. García and C. Benítez

Dept. of Signal Theory, Networking and Communications
University of Granada (SPAIN)

ABSTRACT

This paper shows a revised statistical test for voice activity detection in noise adverse environments. The method is based on a revised contextual likelihood ratio test (LRT) defined over a multiple observation window. The new approach not only evaluates the two hypothesis consisting on all the observations to be speech or non-speech but all the possible hypothesis defined over the individual observations. The implicit hangover mechanism artificially added by the original method was not found in the revised method so its design can be further improved. With these and other innovations the proposed method showed a high speech/non-speech discrimination over a wide range of SNR conditions. The experimental framework showed that the revised method yields significant improvements over standardized VADs for discontinous voice transmission and distributed speech recognition, as well as over recently reported methods.

1. INTRODUCTION

Emerging applications in the field of speech processing are demanding increasing levels of performance in noise adverse environments. Examples of such systems are the new voice services including discontinuous speech transmission [1, 2, 3] or distributed speech recognition (DSR) over wireless and IP networks [4]. These systems often require a noise reduction scheme working in combination with a precise voice activity detector (VAD) in order to compensate its harmful effect on the speech signal.

During the last decade numerous researchers have studied different strategies for detecting speech in noise and the influence of the VAD on the performance of speech processing systems. Sohn *et al*. [5] proposed a robust VAD algorithm based on a statistical likelihood ratio test (LRT) involving a single observation vector. Later, Cho *et al* [6] suggested an improvement based on a smoothed LRT. Most VADs in use today normally consider hangover algorithms based on empirical models to smooth the VAD decision. It has been shown recently that incorporating contextual information in a multiple observation LRT (MO-LRT) [7] reports benefits for speech/pause discrimination in high noise environments. This paper analyzes this method and shows a new LRT VAD that extends the number of hypothesis on the individual multiple observation that are tested.

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

2. MULTIPLE OBSERVATION LIKELIHOOD RATIO TEST

In a two-hypothesis test, the optimal decision rule minimizing the error probability is the Bayes classifier. Given an observation vector $\tilde{\mathbf{y}}$ to be classified, the problem is reduced to selecting the class (G_0 or G_1) with the largest posterior probability $P(G_i|\tilde{\mathbf{y}})$. From the Bayes rule, a likelihood ratio test (LRT) can be defined as:

$$L(\tilde{\mathbf{y}}) = \frac{p(\tilde{\mathbf{y}}|G_1)}{p(\tilde{\mathbf{y}}|G_0)} \underset{G_0}{\overset{G_1}{\gtrless}} \frac{P(G_0)}{P(G_1)} \quad (1)$$

where the observation vector is classified as G_1 if the likelihood ratio $L(\tilde{\mathbf{y}})$ is greater than the ratio $P(G_0)/P(G_1)$ between the *a priori* class probabilities, otherwise it is classified as G_0. Frequently, there is a need to shift the operating point of the classifier in favor of one of the two classes so that $L(\tilde{\mathbf{y}})$ is compared to a threshold η representing the separation between the classes.

A LRT for detecting the presence of speech in a noisy signal based on a Gaussian model was proposed by Sohn *et al.* [5] and several improvements [6, 8] have been considered to improve its performance. Among them, the multiple observation LRT (MO-LRT) [7] considers not just a single observation vector $\tilde{\mathbf{y}}_t$ measured at a frame t, but also an N-frame neighborhood $\{\tilde{\mathbf{y}}_{t-N}, ..., \tilde{\mathbf{y}}_t, ..., \tilde{\mathbf{y}}_{t+N}\}$:

$$\ell(\tilde{\mathbf{y}}_{t-N}, ...\tilde{\mathbf{y}}_{t+N}) = \frac{p_{\mathbf{y}_{t-N},...,\mathbf{y}_{t+N}|G_1}(\tilde{\mathbf{y}}_{t-N}, ..., \tilde{\mathbf{y}}_{t+N}|G_1)}{p_{\mathbf{y}_{t-N},...,\mathbf{y}_{t+N}|G_0}(\tilde{\mathbf{y}}_{t-N}, ..., \tilde{\mathbf{y}}_{t+N}|G_0)} \quad (2)$$

This test involves the evaluation of an *N*-th order LRT incorporating contextual information to the decision rule and exhibits significant improvements in speech/pause discrimination over the original LRT proposed by Sohn [5].

This smoothed test introduces a non-controllable hangover mechanism that needs to be studied and discussed. This paper reformulates the MO-LRT previously proposed in [7] and shows a new and effective LRT yielding high speech/pause discrimination accuracy. The hangover is then eliminated and not affected by the selection of the number of frames involved in the LRT.

3. REVISED MO-LRT

It is interesting to analyze the hypothesis that are being tested in the evaluation of the previous MO-LRT VAD. Note that the decision is made in favor of one of the two hypothesis:

$$\begin{array}{lll} G_1 & : & \hat{\mathbf{y}}_l = \hat{\mathbf{s}}_l + \hat{\mathbf{n}}_l \\ G_0 & : & \hat{\mathbf{y}}_l = \hat{\mathbf{n}}_l \end{array} \tag{3}$$

for $l = t - N, ..., t, ..., t + N$. The VAD operates on a frame by frame basis and assigns a class to the central frame at time t. In this way, the test evaluates the probability that "all" the observations in the N-frame neighborhood of the central frame to be non-speech or speech. This is the reason to revise the method in order to evaluate not just the two previous hypothesis G_0 and G_1 but also other hypothesis that could be equally possible.

Let $\aleph = \{H_m, m = 1, 2, ..., 2^{2N+1}\}$ be the set of all the possible hypothesis considering all the individual observations to be speech or non-speech in the multiple observation vector $\{\tilde{\mathbf{y}}_{t-N}, ..., \tilde{\mathbf{y}}_t, ..., \tilde{\mathbf{y}}_{t+N}\}$ that is reindexed as $\hat{\mathbf{Y}}$= $\{\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_{N+1}, ..., \hat{\mathbf{y}}_{2N+1}\}$ for convenience of the presentation. Each hypothesis H_m can be defined in terms of a binary integer representation:

$$m = \sum_{k=1}^{2N+1} 2^{b_k} \tag{4}$$

where b_k define if the observation k is non-speech ($b_k = 0$) or speech ($b_k = 1$):

$$\begin{array}{lll} b_k = 1 & : & \hat{\mathbf{y}}_k = \hat{\mathbf{s}}_k + \hat{\mathbf{n}}_k \\ b_k = 0 & : & \hat{\mathbf{y}}_k = \hat{\mathbf{n}}_k \end{array} \qquad k = 1, 2, ..., N + 1 \tag{5}$$

Thus, each hypothesis H_m consists of $2N$+1 individual hypothesis involving the $2N$+1 observations. The classification problem is then reformulated as selecting the class i with the higher posterior probability $P(H_i|\hat{\mathbf{Y}})$ and assigning speech (G_1) or non-speech (G_0) to the current frame depending on the bit b_{N+1} associated to H_i.

If the set $\aleph$ of all the possible hypothesis is splitted depending on the value of the central frame bit b_{N+1} as:

$$\begin{array}{l} M_1 = \{H_m \in \aleph : b_{N+1} = 1\} \\ M_0 = \{H_m \in \aleph : b_{N+1} = 0\} \end{array} \tag{6}$$

the posterior probabilities are defined to be:

$$\begin{array}{l} p(G_1|\hat{\mathbf{Y}}) = \sum_{m \in M_1} p(H_m|\hat{\mathbf{Y}}) \\ p(G_0|\hat{\mathbf{Y}}) = \sum_{m \in M_0} p(H_m|\hat{\mathbf{Y}}) \end{array} \tag{7}$$

and

$$\begin{array}{l} P(G_1) = \sum_{m \in M_1} P(H_m) \\ P(G_0) = \sum_{m \in M_0} P(H_m) \end{array} \tag{8}$$

Using the Bayes rule:

$$\begin{array}{l} p(G_1|\hat{\mathbf{Y}}) = \frac{1}{P(\hat{\mathbf{Y}})} \sum_{m \in M_1} P(H_m) p(\hat{\mathbf{Y}}|H_m) \\ p(G_0|\hat{\mathbf{Y}}) = \frac{1}{P(\hat{\mathbf{Y}})} \sum_{m \in M_0} P(H_m) p(\hat{\mathbf{Y}}|H_m) \end{array} \tag{9}$$

and a revised LRT can be defined as:

$$\Lambda = \frac{p(G_1|\hat{\mathbf{Y}})}{p(G_0|\hat{\mathbf{Y}})} = \frac{\sum_{m \in M_1} P(H_m) p(\hat{\mathbf{Y}}|H_m)}{\sum_{m \in M_0} P(H_m) p(\hat{\mathbf{Y}}|H_m)} \tag{10}$$

An effective approximation to the statistical test described above is to replace the summation by the maximum value of the probability of the hypothesis in M_1 and M_0:

$$\Lambda^* = \frac{\text{máx}_{m \in M_1}\, p(\hat{\mathbf{Y}}|H_m)}{\text{máx}_{m \in M_0}\, p(\hat{\mathbf{Y}}|H_m)} \tag{11}$$

By taking logarithms this test is expressed in a more compact form:

$$\log \Lambda^* = \max_{m \in M_1} l_m - \max_{m \in M_0} l_m \tag{12}$$

where:

$$l_m = \sum_{k=1}^{2N+1} \log p(\hat{\mathbf{y}}_k|b_k) \tag{13}$$

If we restrict the number of possible hypothesis by removing those corresponding to more than one speech/non-speech or non-speech/speech transition in the N-frame neighborhood, the test can be rewritten in matrix form:

$$\mathbf{L} = \mathbf{K}\mathbf{B}_1 + (\mathbf{I} - \mathbf{K})\mathbf{B}_0 \tag{14}$$

where:

$$\begin{array}{l} \mathbf{L} = [l_1, l_2, ..., l_{2N+1}]^T \\ \mathbf{B}_0 = [\log p(\hat{\mathbf{y}}_1|0), ..., \log p(\hat{\mathbf{y}}_{2N+1}|0)]^T \\ \mathbf{B}_1 = [\log p(\hat{\mathbf{y}}_1|1), ..., \log p(\hat{\mathbf{y}}_{2N+1}|1)]^T \end{array} \tag{15}$$

and $\mathbf{K}$ is the Hankel matrix:

$$\mathbf{K} = \begin{bmatrix} 0 & 0 & ... & 0 & ... & 0 & 0 \\ 0 & 0 & ... & 0 & ... & 0 & 1 \\ 0 & 0 & ... & 0 & ... & 1 & 1 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & 1 & ... & 1 & ... & 1 & 1 \\ 1 & 1 & ... & 1 & ... & 1 & 1 \\ 1 & 1 & ... & 1 & ... & 1 & 0 \\ & & & & & & \\ 1 & 1 & ... & 0 & ... & 0 & 0 \\ 1 & 0 & ... & 0 & ... & 0 & 0 \end{bmatrix} \tag{16}$$

Moreover, if the matrix $\mathbf{K}$ is splitted into two submatrices $\mathbf{K}_0$ and $\mathbf{K}_1$ by extracting from $\mathbf{K}$ the rows with a central 0 or 1, respectively, the test is easily reduced to:

$$\log \Lambda^* = \text{máx}\, \mathbf{L}_1 - \text{máx}\, \mathbf{L}_0 \tag{17}$$

where:

$$\begin{array}{l} \mathbf{L}_1 = \mathbf{K}_1\mathbf{B}_1 + (\mathbf{I} - \mathbf{K}_1)\mathbf{B}_0 \\ \mathbf{L}_0 = \mathbf{K}_0\mathbf{B}_1 + (\mathbf{I} - \mathbf{K}_0)\mathbf{B}_0 \end{array} \tag{18}$$

Note that $\mathbf{K}_0 = \mathbf{I} - \mathbf{K}_1$ and equation 18 is reduced to:

$$\begin{array}{l} \mathbf{L}_1 = \mathbf{K}_1\mathbf{B}_1 + (\mathbf{I} - \mathbf{K}_1)\mathbf{B}_0 \\ \mathbf{L}_0 = (\mathbf{I} - \mathbf{K}_1)\mathbf{B}_1 + \mathbf{K}_1\mathbf{B}_0 \end{array} \tag{19}$$

As an example, for $N = 1$, the matrices $\mathbf{K}$, $\mathbf{K}_0$ and $\mathbf{K}_1$ are defined to be:

$$\mathbf{K} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \mathbf{K}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad \mathbf{K}_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \tag{20}$$

and $\mathbf{L}_1$ and $\mathbf{L}_0$ are computed by:

$$\begin{array}{l} \mathbf{L}_1 = \mathbf{K}_1\mathbf{B}_1 + (\mathbf{I} - \mathbf{K}_1)\mathbf{B}_0 = \\ = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \log p(\hat{\mathbf{y}}_1|1) \\ \log p(\hat{\mathbf{y}}_2|1) \\ \log p(\hat{\mathbf{y}}_3|1) \end{bmatrix} + \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \log p(\hat{\mathbf{y}}_1|0) \\ \log p(\hat{\mathbf{y}}_2|0) \\ \log p(\hat{\mathbf{y}}_3|0) \end{bmatrix} \\ \mathbf{L}_0 = (\mathbf{I} - \mathbf{K}_1)\mathbf{B}_1 + \mathbf{K}_1\mathbf{B}_0 = \\ = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \log p(\hat{\mathbf{y}}_1|1) \\ \log p(\hat{\mathbf{y}}_2|1) \\ \log p(\hat{\mathbf{y}}_3|1) \end{bmatrix} + \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \log p(\hat{\mathbf{y}}_1|0) \\ \log p(\hat{\mathbf{y}}_2|0) \\ \log p(\hat{\mathbf{y}}_3|0) \end{bmatrix} \end{array} \tag{21}$$

The algorithm for voice activity detection is based on a comparison of a likelihood ratio to a given threshold η:

$$\log \Lambda^* \underset{G_0}{\overset{G_1}{\gtrless}} \eta \tag{22}$$

For the computation of the logarithmic probability vectors $\mathbf{B}_0$ and $\mathbf{B}_1$, an adequate statistical model needs to be selected. In this work, the discrete Fourier transform (DFT) coefficients of the clean speech (S_j) and the noise (N_j) are assumed to be asymptotically independent Gaussian random variables:

$$\begin{array}{l} p(\hat{\mathbf{y}}|G_0) = \prod_{j=0}^{J-1} \frac{1}{\pi\lambda_N(j)} \exp\{-\frac{|Y_j|^2}{\lambda_N(j)}\} \\ p(\hat{\mathbf{y}}|G_1) = \prod_{j=0}^{J-1} \frac{1}{\pi[\lambda_N(j)+\lambda_S(j)]} \exp\{-\frac{|Y_j|^2}{\lambda_N(j)+\lambda_S(j)}\} \end{array} \tag{23}$$

where Y_j represents the noisy speech DFT coefficients and $\lambda_N(j)$ and $\lambda_S(j)$ denote the variances of N_j and S_j, respectively. Thus, the logarithmic probabilities found in $\mathbf{B}_0$ and $\mathbf{B}_1$ can be computed as in [7] through the "a priori" and "a posteriori" SNRs defined to be:

$$\gamma_j = \frac{|Y_j|^2}{\lambda_N(j)} \qquad \xi_j = \frac{\lambda_S(j)}{\lambda_N(j)} \tag{24}$$

that are estimated using the Ephraim and Malah minimum mean-square error (MMSE) estimator [9].

The algorithm is adaptive and suitable for non-stationary noise environments since the statistical properties are updated when the frame is classified as a non-speech frame. In this way, the variance of the noise λ_N is updated as:

$$\lambda_N(j) = \alpha\lambda_N(j) + (1 - \alpha)|Y_j|^2 \tag{25}$$

Figure 1 shows the operation of the original MO-LRT VAD and the revised one over an utterance of the Spanish SpeechDatCar database [10] in clean conditions (25 dB SNR). Note that the new algorithm removes the saving period at the word beginnings and endings being more accurate in such a low noise conditions. It is interesting to point out that the hangover of the original MO-LRT was a result of extending the decision over a neighborhood of the current frame. However, the new statistical test exhibits the same smoothing process and reduced variance of the decision variable with the benefit of being suitable for a more effective hangover mechanism development. Under the noisiest conditions (5 dB SNR), the new algorithm has a similar behavior to the previous VAD as shown in figure 2.

4. EXPERIMENTAL RESULTS

The ROC (receiving operating characteristic) curves have shown to be very effective for the evaluation of voice activity detectors [11, 12]. These plots, which show the trade-off between the error probabilities of speech and non-speech detection as the threshold η varies, completely describe the VAD error rate. In this analysis, the Spanish SpeechDat-Car (SDC)[10] database was used. This database consists of recordings from distant and close-talking microphones in car environments at different driving conditions. For the computation of the speech and non-speech distributions, a semiautomatic "speech/non-speech" labeling process was conducted on the close talking microphone.

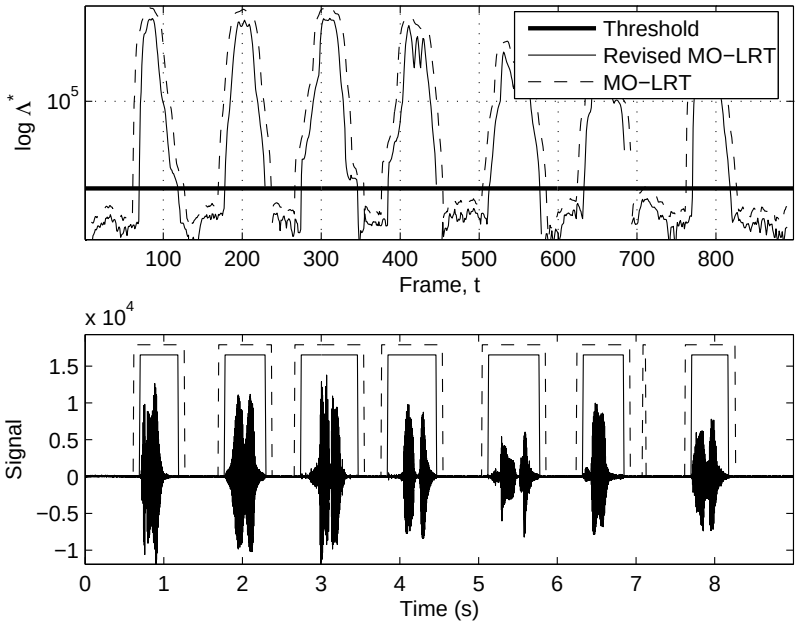

Figura 1. *Comparison between the original MO-LRT and the revised MO-LRT for VAD in clean conditions.*

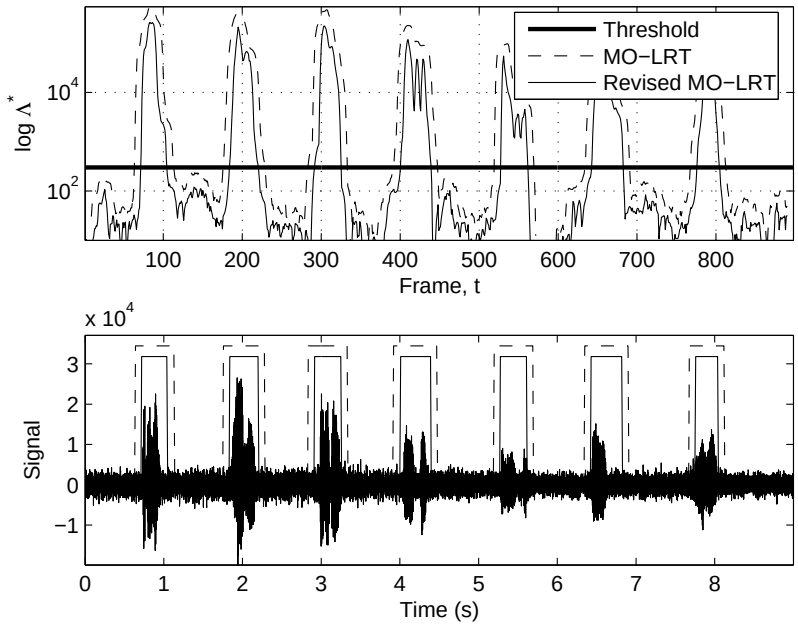

Figura 2. *Comparison between the original MO-LRT and the revised MO-LRT for VAD in high noise car environment.*

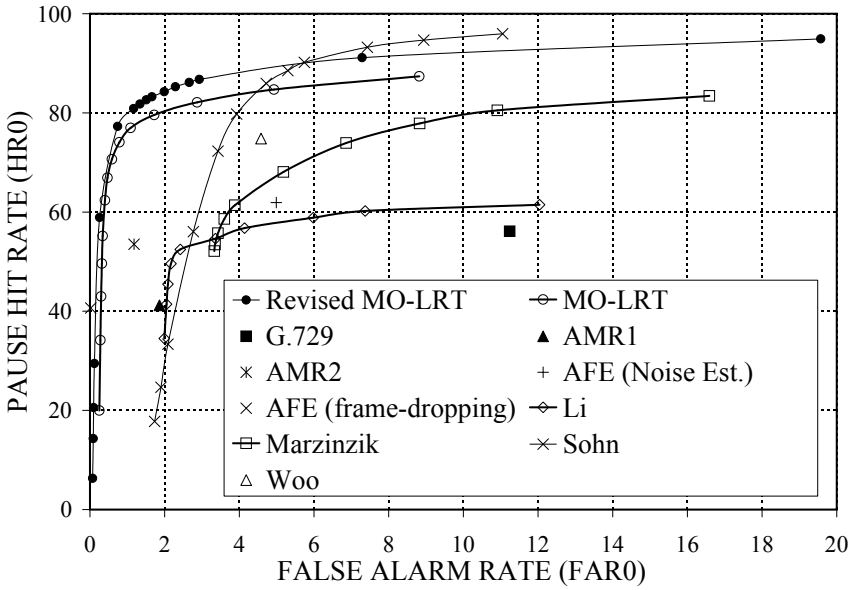

Figura 3. *ROC curves in quiet noise conditions (stopped car and engine running) and close talking microphone.*

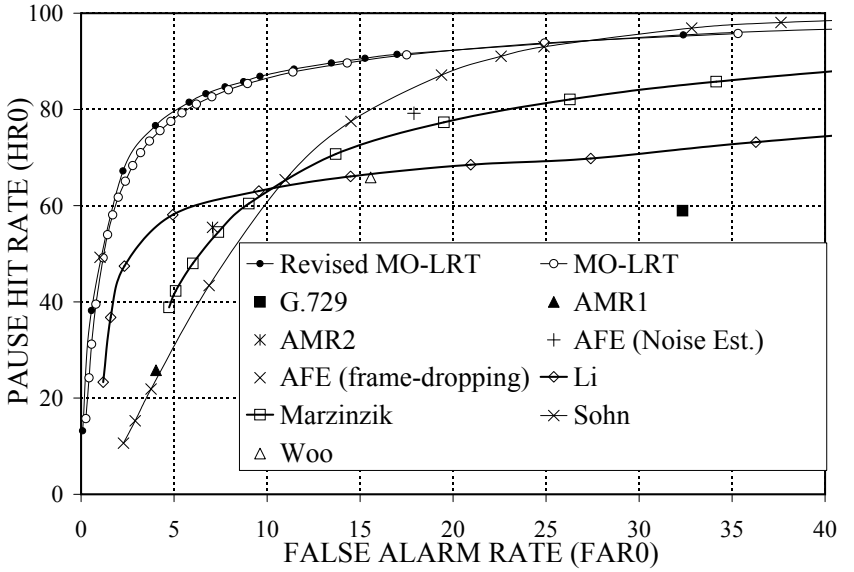

Figura 4. *ROC curves in high noise conditions (high speed over a good road) and distant talking microphone.*

Fig. 3 and 4 shows the non-speech hit rate (HR0) versus the false alarm rate (FAR0=1-HR1, where HR1 denotes the speech hit rate) for recordings from the distant microphone and under quiet and high noise conditions, respectively. The results show that the revised MO-LRT method yields better results than the previous method. These improvements are obtained by the robustness of the decision rule and by removing the implicit hangover found in the previous method and developing a more suitable design. It can be concluded from figures 3 and 4 that the proposed algorithm also outperforms a number of standardized VAD methods including the ITU-T G.729 [2], ETSI AMR (opts. 1 and 2) [3] and the ETSI Advanced Front-End (AFE) [4] for distributed speech recognition (DSR), as well as other recently published VAD methods [5, 13, 14, 12]. The best results are obtained for N= 8 while increasing the number of observations over this value reports no additional improvements. In particular, the proposed VAD outperforms the Sohn's VAD [5], that assumes a single observation in the decision rule and a HMM-based hangover mechanism.

5. CONCLUSIONS

This paper revises a multiple observation likelihood ratio test for voice activity detection in noisy environments. The new approach not only evaluates the two hypothesis consisting on all the observations to be speech or non-speech, but all the possible hypothesis defined over the individual observations. The revised statistical test exhibits the same smoothing process and reduced variance of the decision variable with the benefit of being suitable for a more effective hangover mechanism development. The experimental results showed a high speech/non-speech discrimination accuracy over a wide range of SNR conditions and significant improvements over standardized VADs such as ITU-T G.729, ETSI AMR and ETSI AFE, as well as other publicly available approaches.

6. REFERENCES

[1] A. Benyassine, E. Shlomot, H. Su, D. Massaloux, C. Lamblin, y J. Petit, "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*, vol. 35, no. 9, pp. 64–73, 1997.

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

[4] 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.

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

[6] Y. D. Cho y A. Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 8, no. 10, pp. 276–278, 2001.

[7] J. Ramírez, José C. Segura, C. Benítez, L. García, y A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, 2005.

[8] A. Sangwan, W.P. Zhu, y M.O. Ahmad, "On the competitive neyman-pearson approach for composite hypothesis testing and its application in voice activity detection," in *Proc. of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2006, vol. 3, pp. 301–304.

[9] Y. Ephraim y D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-32, pp. 1109–1121, 1984.

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

[11] J. Ramírez, J. C. Segura, M. C. Benítez, A. de la Torre, y A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[12] M. Marzinzik y 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.

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

[14] Q. Li, J. Zheng, A. Tsai, y 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.