A New Kullback–Leibler VAD for Speech Recognition in Noise

Javier Ramírez, *Student Member, IEEE*, José C. Segura, *Senior Member, IEEE*, Carmen Benítez, *Member, IEEE*, Ángel de la Torre, and Antonio J. Rubio, *Member, IEEE*

***Abstract*—This letter shows an innovative voice activity detector (VAD) based on the Kullback–Leibler (KL) divergence measure. The algorithm is evaluated in the context of the recently approved ETSI standard for distributed speech recognition (DSR). The VAD uses long-term information of the noisy speech signal in order to define a more robust decision rule yielding high accuracy. The Mel-scaled filter bank log-energies (FBE) are modeled by means of Gaussian distributions, and a symmetric KL divergence is used for the estimation of the distance between speech and noise distributions. The decision rule is formulated in terms of the average subband KL divergence that is compared to a noise-adaptable threshold. An exhaustive analysis using the AURORA databases is conducted in order to assess the performance of the proposed method and to compare it to existing standard VAD methods.**

***Index Terms*—Kullback–Leibler (KL) divergence, noise reduction, robust speech recognition, voice activity detection (VAD).**

I. Introduction

THE EMERGING applications of speech technologies (especially in mobile communications, robust speech recognition or digital hearing aids devices) often require a noise reduction scheme working in combination with a precise voice activity detector (VAD) [1]. There exist well-known noise suppression algorithms that are widely used in these applications and for which the VAD is critical for the demanded levels of performance. A typical VAD decomposes the input speech signal into frames and decision is made on a basis of the actual frame [2], [3]. These algorithms are effective in numerous applications but often cause detection errors mainly due to the loss of discrimination at low SNR levels. Several algorithms trying to palliate these drawbacks by means of the definition of more robust decision rules [4] have been proposed. These alternative VAD procedures use long-term information about the speech signal and usually yield better discrimination with sustained improvements in speech/nonspeech hit rates.

This letter shows a new VAD based on the Kullback–Leibler (KL) divergence measure that takes advantage of this design strategy. The algorithm is evaluated in the context of the AURORA project [5], [6] and the recently approved Advanced Front-end (AFE) standard [7] for distributed speech recognition (DSR). The quantifiable benefits of this approach are studied by means of an exhaustive performance analysis conducted on the AURORA databases, with standard VADs such as the ITU G.729 [8], ETSI AMR [9] and AFE [7], and the Sohn's algorithm [2] used as a reference.

II. Background

The proposed VAD is based on the Kullback–Leibler divergence measure or relative entropy between two probability distributions $p_1(x)$ and $p_2(x)$, which is defined by

$$H(p_1\|p_2) = \int p_1(x)\log\left(\frac{p_1(x)}{p_2(x)}\right)dx. \tag{1}$$

It can be shown [10] that the relative entropy is nonnegative and it is null only if the two probability distributions are identical. Thus, it discriminates statistical processes by indicating how distinguishable $p_1(x)$ is from $p_2(x)$ by maximum-likelihood hypothesis testing when the actual data obeys $p_1(x)$.

The KL divergence can be easily computed in the case of Gaussian distributions. Note that $H(p_1(x)\|p_2(x))$ is the expected value of the function $\log(p_1(x)/p_2(x))$ over $p_1(x)$, i.e., $E_1[\log(p_1(x)/p_2(x))]$. Thus, the KL divergence computation is reduced to the estimation of the means μ_1 and μ_2 and standard deviations σ_1 and σ_2 of the distributions $p_1(x)$ and $p_2(x)$, respectively

$$H(p_1\|p_2) = \frac{1}{2}\left[\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) - 1 + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2}\right]. \tag{2}$$

III. KL–FBE VAD

The proposed VAD works in the Mel-scaled energy domain and assumes a Gaussian model for the logarithmic filter bank energy (FBE) distributions of speech p_S and noise p_N in each band. The detection algorithm is based on the symmetric KL "distance," $\rho_{S,N} = H(p_S\|p_N) + H(p_N\|p_S)$, or equivalently

$$\rho_{S,N} = \int (p_S(x) - p_N(x))\log\left(\frac{p_S(x)}{p_N(x)}\right)dx \tag{3}$$

which for Gaussian distributions is given by

$$\rho_{S,N} = \frac{1}{2}\left[\frac{\sigma_S^2}{\sigma_N^2} + \frac{\sigma_N^2}{\sigma_S^2} - 2 + (\mu_S-\mu_N)^2\left(\frac{1}{\sigma_S^2} + \frac{1}{\sigma_N^2}\right)\right] \tag{4}$$

where μ_S and μ_N are the means of the signal and noise log-Energy distributions, respectively, and σ_S and σ_N their corresponding standard deviations.

Manuscript received March 17, 2003; revised June 3, 2003. This work was supported in part by the Spanish Government under the CICYT project TIC2001-3323. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alex Acero.

The authors are with the Department of Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, Campus Universitario Fuentenueva S/N, 18071 Granada, Spain (e-mail: javierrp@ugr.es).

Digital Object Identifier 10.1109/LSP.2003.821762

The algorithm can be described as follows. First, the signal is preemphasized and segmented into 25-ms frames with a 10-ms window shift. The Mel-scaled log-Energies $\{E(n,k)\}_{k=0}^{K_{FB}-1}$ are then computed for the kth filter and the nth frame by applying a Mel-scaled triangular filter bank to the signal spectrum magnitude [6], [7]. The VAD models the subband Log-energies by means of Gaussian distributions being each band independently processed by means of a $(2N+1)$-frame sliding window

$$W^{(k)} = \{E(j,k)\}_{j=n-N}^{n+N} \tag{5}$$

which is subdivided as the inferior and superior windows

$$W_1^{(k)} = \{E(j,k)\}_{j=n-N}^{n-1} \quad W_2^{(k)} = \{E(j,k)\}_{j=n+1}^{n+N} \tag{6}$$

respectively. In a second stage, the mean value of the energy windows $W_1^{(k)}$ and $W_2^{(k)}$, $\mu_1(k)$ and $\mu_2(k)$, and their standard deviations, $\sigma_1(k)$ and $\sigma_2(k)$, respectively, are computed and on-line averaged by a first-order IIR smoothing filter

$$\begin{aligned}\hat{\mu}_i(k) &= \lambda\hat{\mu}_i(k) + (1-\lambda)\mu_i(k)\\ \hat{\sigma}_i(k) &= \lambda\hat{\sigma}_i(k) + (1-\lambda)\sigma_i(k), \qquad i=1,2.\end{aligned} \tag{7}$$

The k-band signal mean and standard deviation required by (4) are estimated using the Log-energy window $W_2^{(k)}$ as $\mu_S(k) = \hat{\mu}_2(k)$ and $\sigma_S(k) = \hat{\sigma}_2(k)$, while noise statistics $\mu_N(k)$ and $\sigma_N(k)$ are updated during nonspeech periods to track nonstationary noise environments by

$$\begin{aligned}\mu_N(k) &= \lambda\mu_N(k) + (1-\lambda)\min\{\hat{\mu}_1(k), m(k), \hat{\mu}_2(k)\}\\ \sigma_N(k) &= \lambda\sigma_N(k) + (1-\lambda)\min\{\hat{\sigma}_1, \hat{\sigma}_2(k)\}\end{aligned} \tag{8}$$

where λ is a forgetting factor and $m(k)$ is the median of the whole log-Energy window $W^{(k)}$.

The algorithm measures the KL "distance" $\rho_{S,N}(k)$ through (4) with the subband probabilities modeled by means of Gaussians distributions. Assuming two hypothesis: H_0 (speech absent) and H_1 (speech present), the decision is formulated by averaging the $K_{\rm FB}$ subband KL distances

$$\hat{\rho}_{S,N} = \frac{1}{K_{\rm FB}} \sum_{k=0}^{K_{\rm FB}-1} \rho_{S,N}(k) \mathop{\lesseqgtr}_{H_0}^{H_1} \eta. \tag{9}$$

The detection threshold η can be fixed or adaptable to the observed noise energy E. Optimal thresholds η_0 and η_1 for clean and high noise conditions, respectively, are defined and the linear threshold tuning shown in Fig. 1 is used. This model ensures a high speech/nonspeech discrimination improving speech pause detection at high and medium SNR levels while maintaining a high accuracy for speech periods.

IV. Experimental Framework

Several experiments using the AURORA databases were carried out to evaluate the performance of the KL-FBE VAD and to compare it to the most representative standard methods [7]–[9]. This section evaluates the speech/nonspeech discrimination as a function of the SNR, provides the Receiver Operating Characteristic (ROC) curves for speech databases recorded under real conditions and compares speech recognition performance.

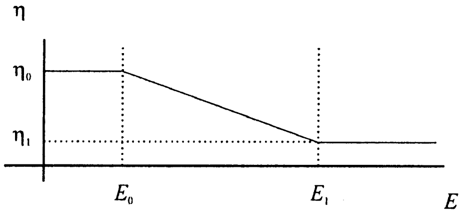


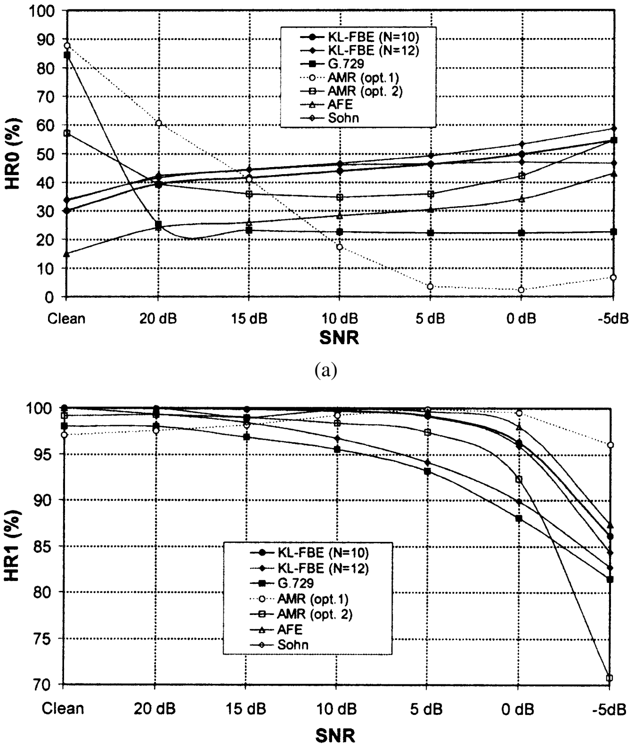
Fig. 1. Adaptive threshold to noise level.

A. Speech/Nonspeech Discrimination Analysis

First, the proposed VAD was evaluated in terms of the ability to discriminate between speech and pause periods at different SNR levels. The clean TIdigits database was used to label each utterance as speech or pause frames for reference. Detection performance as a function of the SNR was assessed for the AURORA 2 database in terms of the speech pause hit-rate (HR0) and the speech hit-rate (HR1) (i.e., the fraction of all actual pause or speech frames that are correctly detected as pause or speech frames, respectively). The optimal parameters for the VAD were: $\eta = 0.4$, $N = 12$ and $\lambda = 0.9$, while the filter bank decomposes the signal in $K_{\rm FB} = 23$ Mel-scaled subbands [6], [7]. Fig. 2 shows HR0 and HR1 as a function of the SNR for KL-FBE, G.729, AMR, AFE, and the Sohn's VAD. Table I compares the VADs in terms of the average hit-rates. Thus, KL-FBE obtains the best behavior in detecting speech pauses with a 46.83% HR0 average value, while the G.729, AMR1, AMR2, AFE and the Sohn's VAD yield 31.77%, 31.31%, 42.77%, 28.74%, and 43.66%, respectively. On the other hand, the KL-FBE VAD is the most precise VAD in detecting speech periods exhibiting the slowest decay in performance at unfavorable noise conditions as shown in Fig. 2(b). KL-FBE attains a 96.96% HR1 average value in speech detection while G.729, AMR1, AMR2, AFE, and the Sohn's VAD provide 93.00, 98.18%, 93.76%, 97.70%, and 94.46%, respectively. Although AMR1 and AFE seems to be well suited for maintaining a high-accuracy detecting speech periods at low SNRs, it is only motivated by their extremely conservative behavior that degrades their speech pause detection accuracy being HR0 less than 10% for SNR values below 10 dB. This fact makes them less useful than other VADs in a practical speech processing system where it is typically demanded a 50% speech pause hit-rate for adequately modeling the time-varying noise statistics and the efficient application of the noise compensation algorithms. The KL-FBE VAD yielded better results than the Sohn's algorithm in speech/pause detection with higher speech and nonspeech hit-rates. Thus, considering together speech and pause hit-rates, the proposed VAD yielded the best results when compared to the most representative VADs tested.

B. ROC Curves

An additional test was conducted to compare speech detection performance by means of the VAD ROC curve [11], a frequently used methodology that completely describes the VAD error rate [4]. The Spanish SpeechDat-Car (SDC) database [12] was used in the analysis. This database contains 4914 recordings (files) from more than 160 speakers. Recordings from the



(a)

(b)

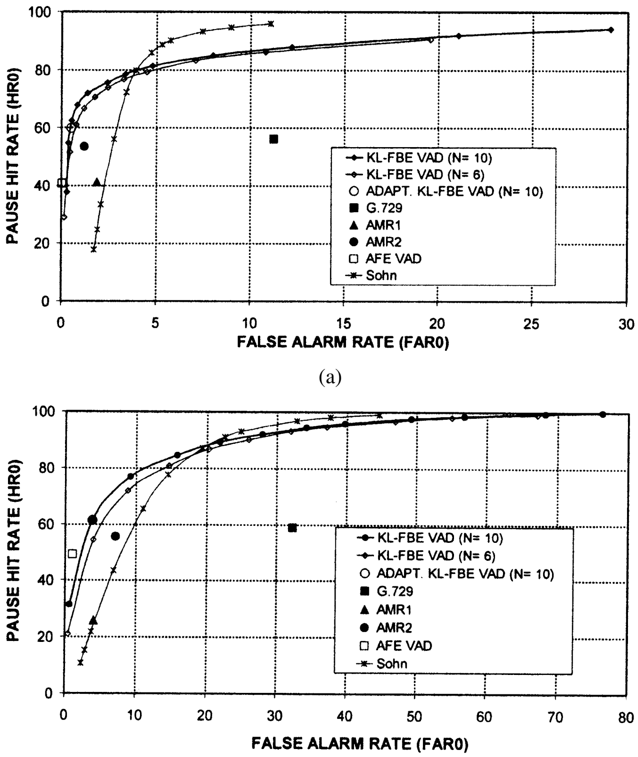
Fig. 2. (a) Nonspeech hit-rate (HR0). (b) Speech hit rate (HR1).

TABLE I
AVERAGE SPEECH/NONSPEECH HIT RATES FOR SNRS FROM "CLEAN" TO −5 dB

	G.729	AMR1	AMR2	AFE	Sohn	KL-FBE
HR0(%)	31.77	31.31	42.77	28.74	43.66	46.83
HR1(%)	93.00	98.18	93.76	97.70	94.46	96.96

close-talking microphone and from one of the distant microphones are included. As in the whole SDC database, the files are categorized into three noisy conditions (quiet, low noisy, and highly noisy) depending on the driving conditions. Thus, recordings from the close-talking microphone are used in the analysis to label speech/pause frames for reference, while recordings from the distant microphone are used to evaluate the different VADs in terms of their ROC curves.

The speech pause hit rate (HR0) as a function of the false-alarm rate ($\text{FAR0} = 100 - \text{HR1}$) for $0 < \eta \leq 10$ is shown in Fig. 3 together with the working point of the adaptive KL-FBE VAD, G.729, AMR1, AMR2, and AFE. It is clearly shown that the ability of the adaptive KL-FBE VAD to tune the detection threshold enables working on the optimal point of the ROC curve for different noisy conditions. Optimal detection threshold $\eta_0 = 2$ and $\eta_1 = 0.5$ were determined for clean and noisy conditions, respectively, while the threshold calibration curve was defined between $E_0 = 30$ dB (low noise energy) and $E_1 = 50$ dB (high noise energy). It can be derived from these plots that the KL-FBE VAD, when compared to G.729 and AMR VADs, yields the lowest false-alarm rate for a fixed speech pause hit rate and also, the highest speech pause hit rate for a given false-alarm rate. It must be noted that the AFE VAD is only used in the standard [7] for frame-dropping and it has been planned to be conservative exhibiting poor speech pause detection accuracy thus, working on a low false-alarm rate point of the ROC curve shown in Fig. 3. Thus, the adaptive KL-FBE VAD provides the best results when the speech/nonspeech detection accuracy are considered together being the gains especially important over the G.729 VAD. On the other hand, the proposed VAD has been the most precise one in delimiting speech pauses and, when compared to the AFE VAD, it works in a less conservative point of the ROC curve with the best speech pause detection accuracy suffering only a moderate increase in the false-alarm rate.



(a)

(b)

Fig. 3. KL-FBE ROC curves and working point of the different VAD analyzed. (a) Stopped car, motor running. (b) High speed, good road.

If the proposed VAD is compared to the Sohn's algorithm, it can be concluded that the Sohn's VAD ROC curve is shifted to a higher false-alarm region in the ROC space. Both curves cross over but, in low false-alarm rate space, the proposed algorithm yields reduced false-alarm rate and increased speech pause hit rates. As a result, the proposed VAD can operate on a lower false-alarm rate point of the ROC space with increased speech pause hit rates. On the other hand, reducing the delay of the algorithm to six frames ($N = 6$) only leaded to moderate increase in the false-alarm rate and reduction of the nonspeech hit rate. However, when the SNR conditions get noisier (SNR $\leq$ 0dB), reducing the delay may lead to a more accused increase of the false-alarm rate being this parameter more important for a noise robust VAD decision.

C. Speech Recognition Performance

These improvements were corroborated when the VAD was integrated into a speech recognition system. The reference

TABLE II
RECOGNITION PERFORMANCE RESULTS

	Train/Test Conditions		WAcc.(%)				
			KL-FBE	G.729	AMR1	AMR2	Sohn
WF	AURORA 2	MCT	88.76	75.24	83.64	88.40	88.66
		CT	79.81	57.13	66.30	78.33	79.12
	AURORA 3	HM	70.13	67.31	64.94	65.44	70.22
		MM	81.06	77.66	76.34	76.58	80.15
		WM	93.24	92.21	92.58	92.71	93.21
WF+FD	AURORA 2	MCT	89.81	83.55	83.56	87.29	88.11
		CT	82.27	57.08	65.01	78.48	81.77
	AURORA 3	HM	82.96	63.35	76.47	79.53	80.52
		MM	85.08	67.66	81.29	82.37	84.93
		WM	94.78	88.81	94.93	94.91	94.38

framework (Base) is the ETSI AURORA project for DSR [6] and performance is assessed in terms of the word accuracy (WAcc.). Two types of experiments were conducted on the AURORA 2 and 3 databases: the effect of the VAD when 1) it is only used for applying Wiener filtering (WF) (as in the first stage of [7] without Mel scale warping) as noise suppression method, and 2) it is applied for both, WF and removing nonspeech periods [WF+frame-dropping (FD)]. The best recognition performance is obtained when the proposed VAD is also used for FD as shown in Table II. In clean training (CT) the relative improvements in the word accuracy were 58.69%, 49.33%, and 17.61% over G.729, AMR1, and AMR2 VADs, respectively, while in multicondition training (MCT) the advantages were of up to 38.05%, 38.02%, and 19.83%. Similar improvements were obtained for the experiments conducted on the AURORA 3 databases [12]–[14] for the three train/test mismatch conditions defined [well-matched (WM), medium-mismatch (MM), and high-mismatch (HM)]. Again, the KL-FBE VAD provided the best results with 53.65%, 21.43%, and 13.96% improvements over G.729, AMR1, and AMR2, respectively, when the VAD is used for both WF and FD.

When compared to the Sohn' algorithm, the adaptive KL-FBE VAD yielded higher recognition performance being the improvements more important when the VAD is used for both WF and FD. This fact is mainly motivated by the robustness of the proposed algorithm against the acoustic environment shown in Sections IV-A and IV-B. As a conclusion, a good VAD for robust speech recognition needs a compromise between speech and pause detection accuracy. When the VAD suffers a rapid performance degradation under severe noise conditions it losses too many speech frames and leads to numerous deletion errors. On the other hand, if the VAD does not correctly identify nonspeech periods it increases the insertion errors and the corresponding FD performance degradation.

V. CONCLUSION

This letter analyzed the performance of an innovative KL-based VAD and its use in a speech recognition system. A comparison with the most representative standard VAD methods was provided. The exhaustive analysis conducted on the AURORA databases showed relevant improvements over G.729 and AMR VADs and the Sohn's algorithm in speech/pause detection accuracy and recognition performance for a representative set of noises and conditions.

REFERENCES

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

[2] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Lett.*, vol. 6, pp. 1–3, Jan. 1999.

[3] Y. D. Cho, K. Al-Naimi, and A. Kondoz, "Mixed decision-based noise adaptation for speech enhancement" and not "A statistical model-based voice activity detection," *Electron. Lett.*, vol. 37, no. 8, pp. 540–542, 2001.

[4] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 109–118, Feb. 2002.

[5] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions," in *Proc. ISCA ITRW ASR2000: Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, Sept. 2000.

[6] ETSI, "Speech processing, transmission and auality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," ETSI, Sophia Antipolis, France, ETSI ES 201 108 Rec., 2000.

[7] ETSI, "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," ETSI, Sophia Antipolis, France, ETSI ES 202 050 Rec., 2002.

[8] ITU, "A silence compression scheme for G.729, optimized for terminals conforming to recommendation V.70," ITU, ITU-T Rec. G.729 (Annex B), 1996.

[9] ETSI, "Voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels," ETSI, Sophia Antipolis, France, ETSI EN 301 708 Rec., Dec. 1999.

[10] R. M. Gray, *Source Coding Theory*. Boston, MA: Kluwer, 1990.

[11] V. Madisetti and D. B. Williams, *Digital Signal Processing Handbook*. Boca Raton, FL: CRC, 1999.

[12] A. Moreno *et al.*, "SpeechDat-Car: A large speech database for automotive environments," in *Proc. II LREC*, June 2000.

[13] Nokia, Baseline results for subset of SpeechDat-Car Finnish database for ETSI STQ WI008 advanced front-end evaluation, Jan. 2000.

[14] Texas Instruments, Description and baseline results for the subset of the Speechdat-Car German database used for ETSI STQ Aurora WI008 advanced DSR front-end evaluation, Dec. 2001.