Proc. IEEE Canadian Conference on Electrical and Computer Engineering (St. John's, Nfld.), pp. 470-473, May 1997

COMPARISON OF VOICE ACTIVITY DETECTION ALGORITHMS FOR WIRELESS PERSONAL COMMUNICATIONS SYSTEMS†

Khaled El-Maleh and Peter Kabal

TSP Laboratory, Dept. of Electrical Eng.
McGill University
Montreal, Quebec, Canada H3A 2A7
email: {khaled,kabal}@tsp.ee.mcgill.ca

ABSTRACT

Voice activity detection (VAD) algorithms have become an integral part of many of the recently standardized wireless cellular and Personal Communications Systems (PCS). In this paper, we present a comparative study of the performance of three recently proposed VAD algorithms under various acoustical background noise conditions. We also propose new ideas to enhance the performance of a VAD algorithm in wireless PCS speech applications.

1. INTRODUCTION

Conversational speech is a sequence of consecutive segments of silence and speech. In wireless telephony, the user is often roaming and thus encountering different types and levels of background acoustical noises. This background noise which contaminates the signal results in either noise only or speech plus noise segments. In many speech processing applications, it is desirable to detect speech in the noise. This process is called voice activity detection (VAD) [1]. The VAD operation can be viewed as a decision problem in which the detector decides between noise only, or of speech plus noise. This is a challenging problem in noisy acoustical environments.

Voice activity detection is used in a variety of speech communication systems such as speech coding, speech recognition, hands-free telephony, audio-conferencing, and echo cancellation. The recently proposed multiple-access schemes, such as CDMA, and enhanced TDMA for cellular and PCS systems use some form of VAD [1]. Moreover, in GSM-based wireless systems a VAD module is used for discontinuous transmission to save the battery life of portable units [2]. Variable bit rate (VBR) coding has been recently adopted for CDMA-based cellular and PCS systems to enhance capacity by reducing interference. A VAD device is an indispensable part of any VBR codec as it controls both the average bit rate, and the overall quality of the coder [3].

This paper is organized in five parts. Section 2 starts with a general review of the basics of a VAD design. It then describes, in some detail, three selected VAD designs for the this study. Section 3 reports the results of a simulation study that has been performed to study the performance of the three VADs under various background noise conditions. We then discuss and analyze the simulation results in Section 4. Finally, conclusions are presented in Section 5.

2. VOICE ACTIVITY DETECTION ALGORITHMS

The basic principle of a VAD device is that it extracts some measured features or quantities from the input signal and then compare these values with thresholds, usually extracted from noise only periods. Voice activity (VAD=1) is declared if the measured values exceed the thresholds. Otherwise, no speech activity or noise (VAD=0) is present. What generally characterizes a VAD design is the way it selects its features, and the way it defines and updates the thresholds. In general, a VAD algorithm outputs a binary decision in a frame-by-frame basis where a "frame" of the input signal is a short unit of time such as 20–40 ms. Accuracy, robustness to noise conditions, simplicity, adaptation, and real-time processing are some of the required features of a good VAD.

In the early VAD algorithms, short-time energy, zero crossing rate, and LPC coefficients were among the common features used in the detection process [4]. Cepstral features [5], formant shape [6], a least-square periodicity measure [7] are some of the recent ideas in

†This work was supported by a grant from the Canadian Institute for Telecommunications Research under the NCE program of the Government of Canada

VAD designs.

In this work, we consider three recently proposed VAD algorithms. These include the VAD used in the GSM cellular system [1,2], the VAD used in the enhanced variable rate codec (EVRC) of the North American CDMA-based PCS and cellular systems [3], and a third-order statistics (TOS)-based VAD [8].

2.1. The GSM VAD

In the GSM VAD, an adaptive noise-suppressor filter is used to filter the input signal frame. The coefficients of the filter are computed during noise-only periods. The energy of the filtered signal is compared to a noise-dependent threshold. As both the filter coefficients and the threshold are computed during noise-only frames, special measures are taken to identify noise frames. These include both signal stationarity and periodicity tests. The major weakness of this VAD lies on the stationarity assumption of background noise. This is not always the case for many of the commonly encountered noises in wireless telephony.

To improve the performance of the GSM VAD for both stationary and non-stationary noises, Srinivasan and Gersho [1] proposed several new features to the basic VAD design. These include a multi-band (4 bands) energy comparison, spectral flatness measurement, and using the fraction of the energy of the low frequency band. This improved GSM VAD is more powerful as it relies on multiple-thresholds to make the final decision. Some of these thresholds are determined empirically and the others are dynamically updated based on signal measurements.

2.2. The EVRC VAD

The EVRC coder [3] uses a 3-rate determination algorithm (RDA) to select the appropriate rate and coding strategy for each input frame. The lowest rate signifies a noise-only frame. For our comparative study, we have changed this RDA to output a binary VAD flag. The basic idea of this VAD is similar to the GSM VAD or its improved version. However, the novel part of this VAD is its dynamic updating of the thresholds in a way that copes with different background noise environments and conditions. The spectrum of the input signal is divided into two bands and the energy in each band is compared against two thresholds. Speech is detected if the energy in each band is greater than the corresponding lowest threshold. The thresholds are scaled versions of estimated sub-band noise energies from previous frames. For more details about the VAD implementation, see [3].

2.3. The TOS VAD

Symmetrically distributed (non-skewed) processes are characterized to have a third-order cumulant (TOC) that is identically zero at all lags. However, speech signals have been observed experimentally to be skewed enough to produce significantly non-zero TOC at all lags. Under the assumption that many noises can be modeled as Gaussian or symmetrically distributed processes, it is possible to discriminate speech from noise. In [8], a novel time domain Gaussianity test is used in the speech detection process. The test statistic of this VAD, $\hat{d}$ is defined as

$$\hat{d} = \hat{c}_{3y}^{t} \hat{C}_0^{-1} \hat{c}_{3y}. \quad (1)$$

In this equation, $\hat{c}_{3y}$ is the third-order cumulant of a given frame, and $\hat{C}_0$ is the covariance matrix of the TOC estimated from R initial noise-only frames. Speech is detected if $\hat{d}$ exceeds a selected threshold, otherwise the frame contains noise. One major feature of this VAD is that it has a fixed noise-independent threshold, $\mathcal{T}$ given as $\chi_Q^2(\alpha)$, where α is a pre-selected probability of false alarm (P_F) and Q is the number of lags used in the TOC computation. The value of the threshold is obtained from the chi-square (χ_Q^2) table [8].

3. SIMULATION RESULTS

We have implemented the aforementioned VAD algorithms and tested their performance for different noise environments and at various noise levels. For the purpose of this study, we have recorded several acoustical environmental noises (bus, street, restaurant) and used some noise signals from the NOISEX-92 database (car noise, babble) [9]. Background noise was digitally added to clean speech with SNR values of 20, 10, and 0 dB. The performance of a given VAD algorithm is a function of both the noise level (SNR) and the structure of the background noise (stationary, non-stationary, white, or periodic). In Figures 1–8, we show on each figure the binary output of each VAD superimposed on a noisy speech signal. Due to the space limitations, we show the VAD results only for a high SNR (20 dB) and for a very noisy environment (0 dB).

It is common in modern VAD algorithms to use a 'hangover' period of few frames to delay any pre-mature transition from speech to noise [1,2,3]. This is to minimize the probability of missing speech especially for low-energy unvoiced speech. These hangover mechanisms are generally not effective in correcting isolated VAD errors (i.e 'one' among a sequence of zeros or vice versa). For many VAD applications (especially speech coding), it is desirable to clean up such errors. We

have developed an isolated error correction mechanism (IECM) that significantly corrects the VAD decision in away that makes it more useful to speech applications. The basic idea of the IECM is that we delay the decision by 2 to 3 frames to monitor the VAD decisions in neighboring frames. If the current frame VAD decision is different from its close neighbors, then its VAD flag is changed to be similar to the other frames. This is repeated for each frame to remove any isolated errors. In Figure 7, we show the effectiveness of this algorithm in enhancing the VAD results.

4. DISCUSSION

In this paper, the simulation results shown in Figures 1–6 are the VAD decisions after isolated error correction. The results show a consistent superiority of the EVRC VAD in detecting speech for almost all types of noise and even for very low SNR. However, it occasionally detects noise as speech (false alarm) especially for babble (simultaneous background conversations) noise as SNR gets low. The TOS VAD is ranked overall second in performance and shows almost-perfect detection results for babble noise at 0 dB.

The GSM VAD exhibits good performance under stationary noise environments while it has difficulty distinguishing speech from noise in non-stationary noises such as buses, babble, and street. Also, its performance deteriorates for low SNR (below 20 dB). The GSM VAD results clearly get improved using the modifications suggested in [1] but still the robustness of the VAD is not guaranteed at low SNRs. We have observed that high-energy voiced speech segments are always detected in all VADs under very noisy conditions. However, low-energy unvoiced speech is commonly missed.

Conventionally, the input to the VAD is the conversational speech signal. The linear prediction (LP) residual has been used before as a tool in voicing decision algorithms to classify speech as voiced or unvoiced. In this work, we have also evaluated using the LP residual as the input signal to the VAD algorithm. Thus a standard linear prediction analysis is done first and then the output of the LP analysis filter (residual signal) is fed to the VAD to make the binary decision. The results show that the accuracy of the VAD decisions has been improved in almost all cases when we used the LP residual instead of directly using the input signal (see Figure 8).

5. CONCLUSIONS

We have presented in this paper an experimental comparative study of three VAD algorithms under different background noise conditions. The results show a consistent superiority of both the EVRC and the TOS VADs when compared with the GSM-based VADs. We have also shown that VAD decisions were improved by using the proposed isolated error correction mechanism and the LP residual as the input signal to the VAD.

6. ACKNOWLEDGMENTS

We would like to thank B. V. Nguyen, P. Lim and C. M. Leung for participating in an early stage of this study.

7. REFERENCES

[1] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Proc. of the IEEE Speech Coding Workshop.*, pp. 85–86, October 1993.

[2] D. K. Freeman, G. Cosier, C.B. Southcott, and I. Boyd, "The voice activity detector for the Pan-European digital cellular mobile telephone service," in *Proc. Intl. Conf. Acoust., Sp., & Sig. Proc.*, pp. 369–372, Glasgow, May 1989.

[3] TIA Document, PN-3292, Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems, January, 1996.

[4] L. R. Rabiner, and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," in *Proc. Intl. Conf. Acoust., Sp., & Sig. Proc.*, pp. 323–326, May 1977.

[5] J. A. Haigh, and J. S. Mason, "Robust voice activity detection using cepstral features," in *IEEE TENCON*, pp. 321–324, China, 1993.

[6] J. D. Hoyt, and H. Wechsler, "Detection of human speech in structured noise," in *Proc. Intl. Conf. Acoust., Sp., & Sig. Proc.*, pp. II-237–II-240, Australia, May 1994.

[7] R. Tucker, "Voice activity detection using a periodicity measure," in *IEE Proceedings-I*, Vol. 139, No. 4, pp. 377–380, August 1992.

[8] M. Rangoussi, and G. Carayannis, "Higher order statistics based Gaussianity test applied to on-line speech processing ," in *Proc. of the IEEE Asilomar Conf.,* pp. 303–307, 1995.

[9] H.J. M. Steeneken, and F. W. M. Geurtsen, " Description of the RSG.10 noise database," Report IZF 1988-3, TNO Institute for Perception, Soesterberg, The Netherlands.

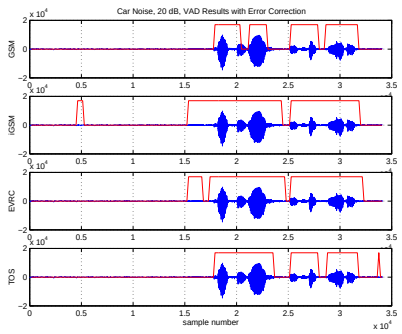

Figure 1: VAD results for car noise at 20 dB SNR.

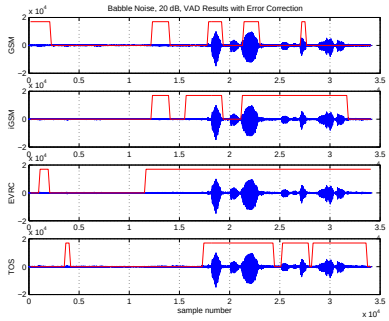

Figure 2: VAD results for babble noise at 20 dB SNR.

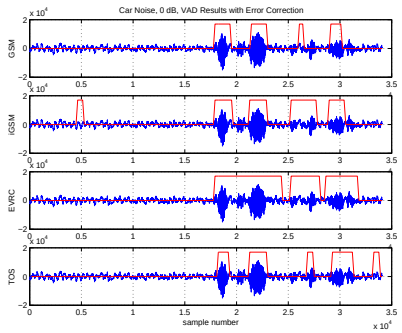

Figure 3: VAD results for car noise at 0 dB SNR.

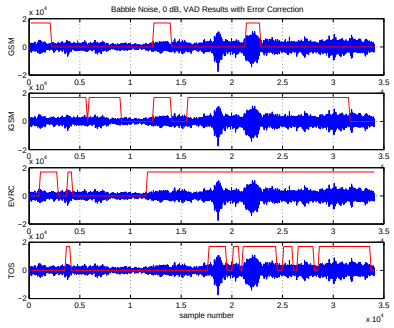

Figure 4: VAD results for babble noise at 0 dB SNR.

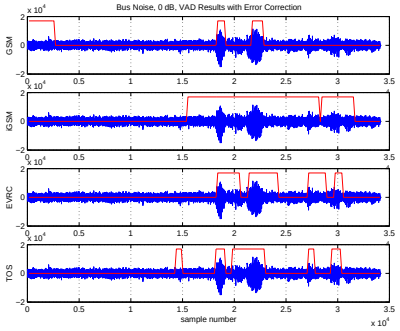

Figure 5: VAD results for bus noise at 0 dB SNR.

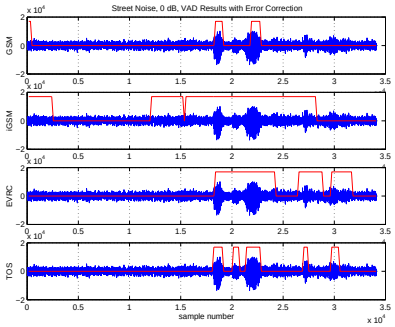

Figure 6: VAD results for street noise at 0 dB SNR.

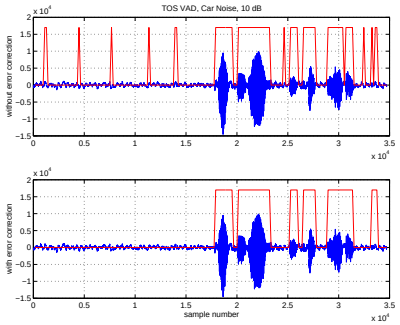

Figure 7: TOS VAD: effect of isolated error correction mechanism.

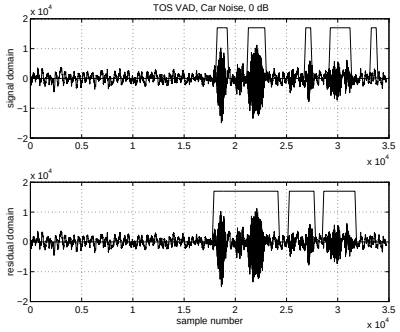

Figure 8: TOS VAD: effect of input signal.