International Journal of Computer and Communication Technology

Volume 2 | Issue 4 Article 11

October 2011

A Statistical Approach for Voiced Speech Detection

Mihir Narayan Mohanty
ITER, SOA University, Bhubaneswar, mihirnmohanty@hotmail.com

Aurobinda Routray
I.I.T., Kharagpur, WB, India, aroutray@iitkgp.ac.in

Pruthviraj Kabisatpathy Prof.
CET, Bhubaneswar, pkabisatpathy@gmail.com

A Statistical Approach for Voiced Speech Detection

Mihir Narayan Mohanty
Department of Applied Electronics and Instrumentation
ITER, SOA University
Bhubaneswar
mihirnmohanty@hotmail.com

Aurobinda Routray
Department of Electrical Engineering and Instrumentation
Indian Institute of Technology, Kharagpur
aroutray@iitkgp.ac.in

Prithviraj Kabisatpathy
Department of Instrumentation and Electronics
College of Engineering and Technology, Bhubaneswar
pkabisatpathy@gmail.com

Abstract— **Detection of Voice in speech signal is a challenging problem in developing high-performance systems used in noisy environments. In this paper, we present an efficient algorithm for robust voiced speech detection and for the application to variable-rate speech coding. The key idea of the algorithm is considering speech energy and zero crossings rate (ZCR) information simultaneously when processing speech signals and finding the end point of the signal. Next to it a decision rule and a background noise statistics estimator, by applying a statistical model. A robust decision rule is derived from the generalized likelihood ratio test (LRT) by assuming that the noise statistics are known *a priori*. The algorithm is most efficient for the time-varying noise. According to our simulation results, the proposed algorithm shows significantly better performance in low signal-to-noise ratio and in noisy environments.**

Keywords- Voice Detection; Speech Detection; Zero Crossing Rate (ZCR); Likelihood Ratio Test (LRT).

I. INTRODUCTION

Wireless communication and digital voice storage systems are some of the application of speech coding. The reduction of average bite rate causes by the employment of voice activity detector (VAD). It can be used to detect speech presence in various type of environment. Even if with different background noise [1].

Speech classification is a major part in speech processing and speech communication application area, i.e. speech recognition, coding and transmission. It is a complex task for noisy environment. Several voice activity detection techniques have been proposed in last decades [2]-[4]. Applications like speech and speaker recognition requires efficient feature extraction techniques where most of the voice part contains speech specific attribute. End point detection as well as silence removal are well known techniques adopted for many years for the voiced speech detection. Two widely accepted methods as Energy and Zero Crossing Rate (ZCR) have been used for silence removal [2],[5]. One of the method for voiced/nonvoiced detection based on epoch extraction is proposed[9]. Zero-frequency filtered speech signal is used to extract the instants of significant excitation (or epochs). The robustness of the method to extract epochs in the voiced regions, even with small amount of additive white noise, is used to distinguish voiced epochs from random instants detected in nonvoiced regions. A discriminative weight training to a statistical model-based voice activity detection (VAD). In this approach, the VAD decision rule is expressed as the geometric mean of optimally weighted likelihood ratios (LRs) based on a minimum classification error (MCE) method[10].

Most of the conventional VAD algorithms assume that the background noise statistics are stationary, which makes it possible to estimate the time varying noise statistics in spite of occasional presence of speech. One widely used feature for voice activity detection is the difference between speech and background noise in temporal variation of statistics. Determination of the presence of speech, in the current frame observed signal statistics compared with estimated noise statistics using some decision rules. Very sensitive VAD can be designed by formulating a decision rule that compares the estimated noise statistics and observed signal statistics. The statistical model is to optimize a VAD, where the decision rule is derived from Likelihood Ratio Test (LRT) by estimating unknown parameters using maximum likelihood criteria.

The paper is organized as follows. In section II we describe the theoretical background. Section III presents the decision rule based on LRT after detecting the end point. The result is discussed in section IV. Section V presents the conclusion.

II. THEORETICAL BACKGROUND

Speech Signal and its Basic Properties :

The speech signal is a slowly time varying signal in the sense, that, when examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary; however, over long periods of time (on the order of 1/5 seconds or more) the signal characteristics change to reflect the different speech sounds being spoken. Usually first 200 msec or more (1600 samples if the sampling rate is 8000 samples/ sec) of a speech recording corresponds to silence (or background noise) because the speaker takes some time to read when recording starts.

The features for voice detection used as classifier inputs has a major role in determination of accuracy and robustness. The features must have ability to discriminate between voiced and unvoiced speech. The selection is based on judgment about the minimum and exact features, those would provide the best performance.

Atal and Rabiner [5] presented one of the excellent method on this topic. But voiced and unvoiced decision is tied to pitch detection, that is best suitable for speech synthesis application instead of voiced speech detection.

The short time energy of vowels tends to be high and the lengthen vowels are likely to have high energy in speech signals and can be detected in a noisy environment. This enables a voice detector to decode an utterance of noisy speech by starting from appropriate beginning. After decoding the end of utterance must be determined. This can be done using the features as energy and ZCRs that shows better result and the MLE hypothesis is to consider the voiced speech.

As zero-crossing rate (ZCR) is an important parameter for end point and voiced decision, Rabiner and Schafer [6] emphasize that, if the ZCR is high, the speech signal is unvoiced else at ZCR low case, the speech signal is voiced. But it can not be said that exactly low and high value, unless it is observed from the plot.

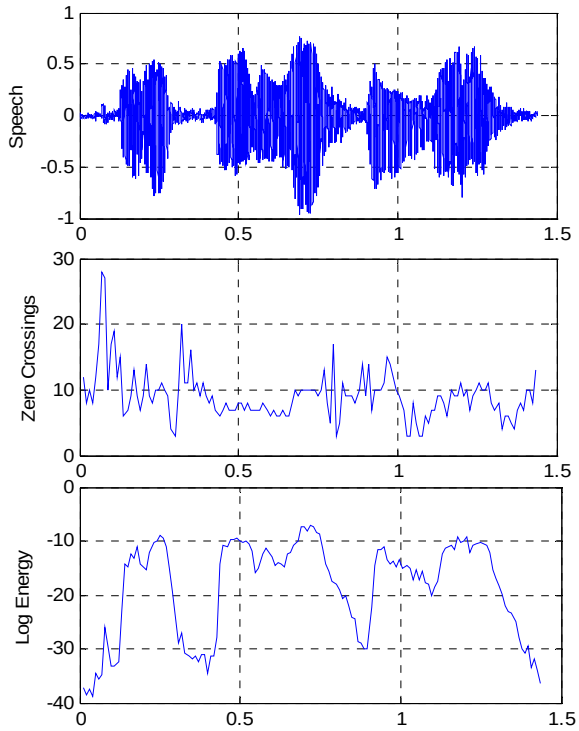

Fig 1.Noisy Speech Signal, its corresponding ZCR and Log Energy

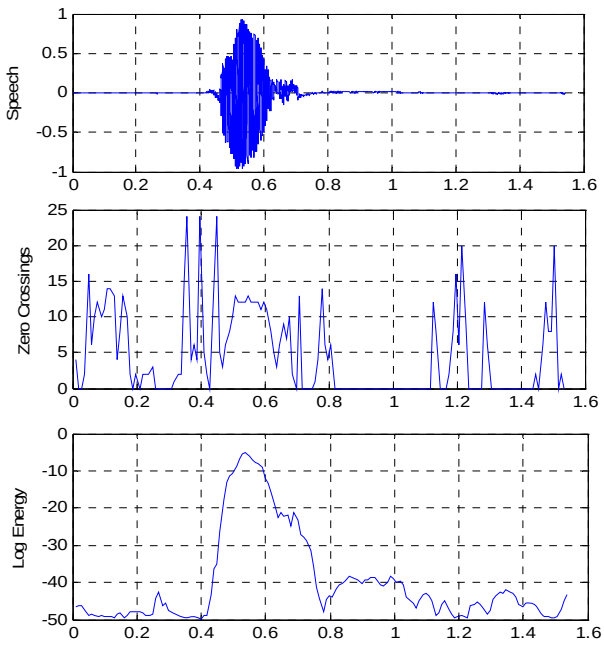

Fig 2. Clean Speech Signal, its corresponding ZCR and Log Energy

III. DECISION RULES FOR VAD BASED ON LRT

The decision rule can be formulated by the decision statistics. This quantity majors the difference between noise and observed signal statistics and decision threshold. In this paper it is derived from generalized Likelihood Ratio Test, assuming that the noise statistics are given a priori by the noise statistics estimator [7]-[8].

Low bit rate speech coders operate on frame basis as a result voice activity detection also perform for each frame of L samples that is assumed to be stationary. The statistical model in which the speech and noise signal are Gaussian random process that are independent of each other. Hence the Discrete Fourier Transform (DFT) co-efficient vectors of each process are asymptotically independent Gaussian random variables. For L-dimensional co-efficient vectors of speech, noise and noisy speech are denoted by S, N and X and for kth elements Sk, Nk and Xk respectively. Two hypotheses for VAD to consider for each frame are

H_0 : speech absent : X = N (1-a)

H_1 : speech present : X = N + S (1-b)

where S, N, and X are L dimensional discrete. Fourier transform (DFT) coefficient vectors of speech, noise and noisy speech. Then the probability density functions conditioned on H_0 and H_1 are given by

$$p(X \mid H_o) = \prod_{k=0}^{L-1} \frac{1}{\pi \lambda_N(k)} \exp\left\{ -\frac{|X_k|^2}{\lambda_N(k)} \right\} \quad (2)$$

$$p(X|H_1)=\prod_{k=0}^{L-1}\frac{1}{\pi[\lambda_N(k)+\lambda_S(k)]}\exp\left\{-\frac{|X_k|^2}{\lambda_N(k)++\lambda_S(k)}\right\} \quad (3)$$

Where $\lambda_N(k)$ and $\lambda_S(k)$ denote the variances of N_k and S_k, respectively. The likelihood ratio for the kth frequency band is

$$\Lambda_k \equiv \frac{p(X_k|H_1)}{p(X_k|H_0)}=\frac{1}{1+\xi_k}\exp\left\{\frac{\gamma_k\xi_k}{1+\xi_k}\right\} \quad (4)$$

Where $\xi_k \equiv \lambda_N(k)/\lambda_S(k)$ and $\gamma_k \equiv \frac{|X_k|^2}{\lambda_N(k)}$ and are called a priori and a posteriori signal-to-noise ratios, respectively. The decision rule is established from the geometric mean of the likelihood ratios for the individual frequency bands, and that is given as

$$\log\Lambda=\frac{1}{L}\sum_{k=0}^{L-1}\log\Lambda_k \underset{H_0}{\overset{H_1}{\lessgtr}} \eta \quad (5)$$

To estimate the unknown parameters ξ_k's, the ML estimator can be derived as

$$\xi_k^{(ML)}=\gamma_k-1 \quad (6)$$

Substituting eq-(6) into eq-(5) and applying likelihood ratio test yields

$$\log\hat{\Lambda}^{(ML)}=\frac{1}{L}\sum_{k=0}^{L-1}\{\gamma_k-\log\gamma_k-1\} \underset{H_0}{\overset{H_1}{\lessgtr}} \eta \quad (7)$$

The LHS of eq-(6) can not be smaller than zero. It implies that the likelihood ratio is biased to H_1. The MLE method reduces the fluctuation of the estimated likelihood ratios in noisy periods.

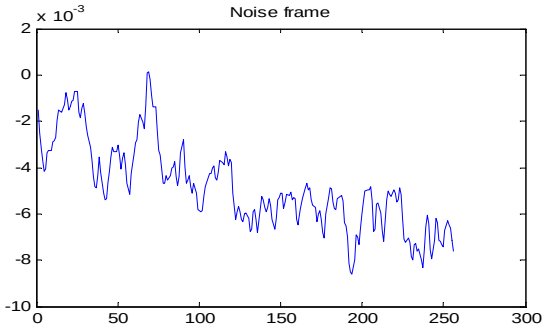

Fig 3. First Noise frame of windowed Speech signal with window length=256.

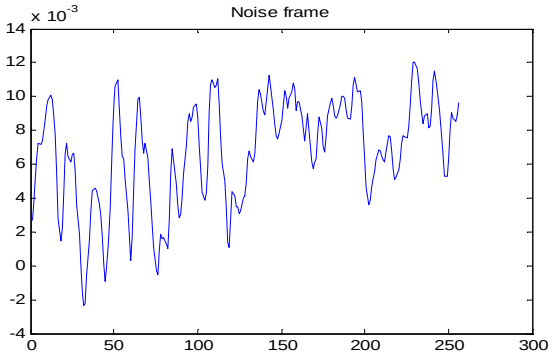

Fig 4. Second Noise frame of windowed Speech signal with window length=256.

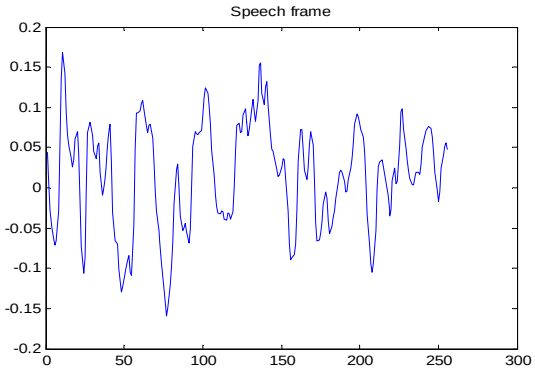

Fig 5. Windowed Unvoiced Speech signal with window length=256.

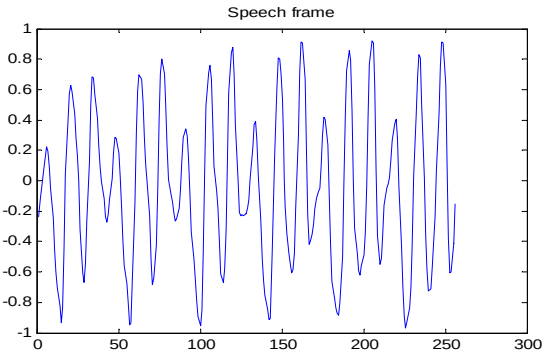

Fig 6. Windowed Voiced Speech signal with window length=256.

IV. RESULT AND DISCUSSION

Samples of speech from 12 different subjects at 12 different levels of conditions were collected. The speech data spoken by the subjects were sampled at 8000 Hz. The samples are

divided into frames and various features are extracted from these framed samples with the end point detection. When the VAD is used in an LPC-based speech coder, the noise can be estimated, but requires large amount of computation. Energy in voiced sample is greater than the unvoiced sample of speech. And the unvoiced speech is a periodic or random in nature. ZCR as a demarcation rule specifying that if ZCR of a speech segment exceeds 50, that can be labeled as unvoiced segment whereas any segment showing ZCR at about 12 is considered to be voiced one. These two methods taking together results 70% accuracy with respect to manually labeled speech sample. In order to make noisy environments, we added the vehicular noise to the clean speech by varying the SNR as 5dB, 10dB, and 15dB respectively. The performance of the model is evaluated by the error probability (P_E) which is the sum of false alarm and missing probabilities. The corresponding error probabilities of the proposed method is 8.5, 7.8 and 6.4 respectively for SNR of 5dB, 10dB, and 15dB for vehicular noise. The proposed algorithm shows a better result in various conditions.

V. CONCLUSION

The proposed method results in a better performance than existing results. This shows the efficacy in time-varying noise case also. The MLE method improves the performance of the VAD by yielding smoother estimates of the *a priori* SNR.

REFERENCES

[1] K. Srinivasan, and A. Gersho, "Voice activity detection for cellular networks," in Proc. IEEE Speech Coding Workshop, pp. 85-86, Oct 1993.

[2] D. G. Childers, M. Hahn, and J. N. Larar, "Silence and voiced/unvoiced/mixed excitation classification of speech," in IEEE Trans. ASSP, vol. 37(11), pp.1771-1774, Nov 1989.

[3] L. R. Rabiner and M. R. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," in Proc. ICASSP, pp. 323–326, 1977.

[4] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in Proc. TENCON, pp. 321-324, 1993.

[5] B. Atal, and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition," Acoustics, Speech, and Signal Processing, IEEE Trans., vol. 24(3), pp. 201-212, Jun 1976.

[6] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signal," Prentice-Hall, Inc., Englewood Cliffs, New Jersy, 1978.

[7] J. Sohn, and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaption," in Proc. Int. Conf. Acoustics, Speech, and Signal Processing, pp. 365-368, 1998.

[8] J. Sohn, N. S. Kim, and W. Sung, "A Statistical Model-Based voice activity detection," in IEEE Signal Processing Letters, vol. 6(1), pp. 1-3, 1999.

[9] N. Dhananjaya and B. Yegnanarayana, "Voiced/Nonvoiced Detection Based on Robustness of Voiced Epochs", IEEE Signal Processing Letters, vol. 17, No. 3, MARCH 2010.

[10] Sang-Ick Kang, Q-Haing Jo, and Joon-Hyuk Chang," Discriminative Weight Training for a Statistical Model-Based Voice Activity Detection", IEEE SIGNAL PROCESSING LETTERS, VOL. 15, 2008.