A Source-Filter Model-Based Unvoiced Speech Detector for Speech Coding

Qian Wang, Xin Du, Weikang Gu
Dept. Information Science and Electronic Engineering of Zhejiang University
Hangzhou, China
e-mail: wangqian_826@zju.edu.cn

***Abstract*—A novel and easy to realize approach for Voice Activity Detection (VAD) is proposed based on the source-filter speech model in the application of Linear-Prediction-structure speech coding. Generalized Likelihood Ratio Test (GLRT) is adopted to formulate the voice activity detector which contains an unvoiced speech detector and a voiced speech detector respectively. By exploiting the linear predictive analysis coefficients and the pitch information produced in the speech coder, the two separate detectors work effectively in uncorrelated noise without increasing computational complexity significantly. Experimental results show that the unvoiced speech detector outperforms conventional algorithm used in speech coding under various noisy conditions.**

Key words-Voice Activity Detection； Generalized Likelihood Ratio Test； Speech Coding； Linear Prediction； Unvoiced Speech Detector

I. INTRODUCTION

Voice Activity Detection (VAD) has become an attractive research topic since the last century with wide application in speech coding[1], speech enhancement[2], and mass storage of speech or audio[3] etc. A speech coder can reduce the computational cost and the average transmission bit rate by correctly classifying speech and noise[5]. Taking G.729B, the bit stream only contains 15 bits for a non active voice frame while that of an active frame is 80 bits.

Speech can be classified into two types: voiced speech and unvoiced speech. Voiced speech signal always contains a pitch, which is a conspicuous characteristic differs from unvoiced speech and noise signal. And it also has a large signal-to-noise ratio (SNR) in usual speech communication environment which is favorable for detecting voiced speech from noise. Unvoiced speech signal, the other type of the speech signal, can be modeled as a noise-like source modulated by the vocal tract. The noise-like property and the relatively low energy make the feature extraction of unvoiced speech has to work in the disadvantageous circumstances.

A general VAD algorithm is composed of a feature extraction module and a classification module. In the past two decades, various feature extraction algorithms have been proposed [6]-[9]. G729B is a well-known method, in which non active voice frames were separated by employing line spectral frequencies, short-time energy, low-frequency energy and zero-crossing rate as features. High-Order Statistics (HOS) [8], adopt the property that High-Order Moments of white Gaussian noise are zero while those of voiced speech are non zero. And some further approach based on HOS feature has been described in[6,7]. The High-Order Moments of unvoiced speech adjacent to the voiced speech are non zero, which is a characteristic of the unvoiced speech. So the HOS-based features e.g. log-kurtosis (LK)[7] and skewness-to-kurtosis ratio (SKR)[8] are claimed to be useful to detect unvoiced speech signals.

Considering the stochastic/deterministic model of the source of speech signal and uncertain existence of voiced or unvoiced speech, this paper presents a voice activity detector using a time-domain speech production model associated with the traditional vocal tract model of linear predictive filtering. Because of the existence of unknown parameters of vocal tract, acoustic source and additive noise, a generalized likelihood ratio test (GLRT) is integrated in VAD problem. Meanwhile, the LP coefficients from speech coder are adopted to avoid the additional calculation.

The remainder of this paper is organized as follows. In Section 2 the stochastic/deterministic model of speech signal with vocal tract filter is explained. Section 3 deduces the GLRT detector of speech in uncorrelated noise. Optimized and practicable detectors of unvoiced speech and voiced speech are generated respectively. Comparative experimental results are shown in Section 4, followed by the conclusions and future work in Section 5.

II. STOCHASTIC/DETERMINISTIC SPEECH MODEL

Human speech is originated from the airflow generated by lungs, which is then sent to the laryngeal. The laryngeal modulates the airflow and generates the acoustical source of speech: periodic pulse by vibration of vocal tract for voiced speech or noise-like signal when vocal tract no longer vibrating for unvoiced speech. A reasonable model of the source of speech signal can be established as:

$$r[m]=\theta\sum_{i=l}^{K}A_i\cos(2\pi f_0 im+\varphi_i)+(1-\theta)g[m],m=0,1,...,M \qquad (1)$$

Where， M is the number of sampling points in a piece of speech signal. Parameters A_i , φ_i , $i=1,2,...,K$ are the amplitudes and the phases of each harmonics, whose fundamental frequency is denoted by f_0 . These three parameters consist of the deterministic part of source signal. The latter part of (1), also the stochastic part, stands for the source of unvoiced speech by g[n], which is always modeled as a Gaussian random signal with zero-mean and variance σ_s^2 .

Published by Atlantis Press, Paris, France.

Vocal tracts, including mouth and nasal cavity, color the source by shaping the frequency components, which produces the human speech. Although uniform tube model is a fundamental and ideal model of the vocal tract, practical vocal tracts could also be modeled as a filter with some formants. Linear predicting (LP) filter, h[n], an all-pole filter, has been widely applied to model the vocal tracts and denoted in Z-domain by:

$$H(z) = \frac{1}{1 + \sum_{i=1}^{L} a_i z^{-i}} \tag{2}$$

Where L is the order of linear predictive analysis filter and $(a_1, a_2, ..., a_L)$ are the linear predictive coefficients. In brief, (1) and (2) formulate an effective model of general speech in time domain and are then incorporated in the proposed detector.

III. Generalized Likelihood Ratio Test of Speech Signal

Assuming speech signal is degraded by additive noise, then the voice activity detection could be described as a binary hypotheses:

$$\begin{aligned} H_0 &: Speech\ Absence\ : X = N \\ H_1 &: Speech\ \mathrm{Pr}esence : X = S + N \end{aligned} \tag{3}$$

Where $S = (s[1], s[2], ..., s[M])^T$ and $N = (n[1], n[2], ..., n[M])^T$ are clean speech and noise presented by different harmonic components respectively. A reasonable assumption of noise N is that it is obeying a Gaussian distribution: $N \sim N(0, \sigma_N^2 I)$. Incorporating the source-filter model deduced in previous section, the hypotheses come to:

$$\begin{aligned} H_0 &: X = N \\ H_1 &: X = S + N = A^{-1}R + N = BR + N \end{aligned} \tag{4}$$

R is the source of speech of signal and matrix A is determined by L linear predictive coefficients defined in (2) while B is the inverse matrix of A. Revisiting (1) and using the Trigonometrical Transform, R could be represented in a vector form by:

$$R = \theta H\beta + (1-\theta)G \tag{5}$$

Where

$$H = \begin{bmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ \cos(2\pi f_0) & \cdots & \cos(2\pi K f_0) & \sin(2\pi f_0) & \cdots & \sin(2\pi K f_0) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos(2\pi f_0(M-1)) & \cdots & \cos(2\pi K f_0(M-1)) & \sin(2\pi f_0(M-1)) & \cdots & \sin(2\pi K f_0(M-1)) \end{bmatrix}$$

$\beta = [p_1\ p_2 \cdots p_K\ q_1\ q_2 \cdots p_K]$, and $p_i = A_i \cos\varphi_i$, $q_i = -A_i \sin\varphi_i$. So the hypotheses can be characterized as follows:

$$p(\mathbf{x} \mid H_0) = \frac{1}{(2\pi)^{N/2} \det(\sigma_N^2 I)^{1/2}} \exp\{-\frac{\mathbf{x}^T\mathbf{x}}{\sigma_N^2}\} \tag{6}$$

$$p(\mathbf{x} \mid H_1) = \frac{1}{(2\pi)^{N/2} \det(\sigma_R^2 BB^T + \sigma_N^2 I)^{1/2}} \tag{7}$$
$$\exp\{-(\mathbf{x} - \theta BH\beta)^T (\sigma_R^2 (1-\theta) BB^T + \sigma_N^2 I)^{-1})(\mathbf{x} - \theta BH\beta)\}$$

Using the generalized likelihood ratio test, the detector is considered as:

$$L(\mathbf{x}) = \max_{B,H,\beta,\theta} \log \frac{p(\mathbf{x} \mid H_1)}{p(\mathbf{x} \mid H_0)} > \gamma \tag{8}$$

As θ is zero when speech signal is absent and equals one when speech signal is present respectively, a further form of the detector is:

$$L(\mathbf{x}) = \max_{\theta}\{\max_{B,H,\beta} \log \frac{p(\mathbf{x} \mid H_1)}{p(\mathbf{x} \mid H_0)}\} = \max\{L_1(\mathbf{x}) = L(\mathbf{x}; \theta = 1), L_0(\mathbf{x}) = L(\mathbf{x}; \theta = 0)\}$$

And the two likelihood ratio functions could be summarized as:

$$L_0(\mathbf{x}) = \max_{B} \log \frac{p(\mathbf{x} \mid H_1, \theta = 0)}{p(\mathbf{x} \mid H_0)} = \max_{B}\{-\mathbf{x}^T((\sigma_R^2 BB^T + \sigma_N^2 I)^{-1} - (\sigma_N^2 I)^{-1})\mathbf{x}\} \tag{9}$$

$$L_1(\mathbf{x}) = \max_{B,H,\beta} \log \frac{p(\mathbf{x} \mid H_1, \theta = 1)}{p(\mathbf{x} \mid H_0)} = \max_{B,H,\beta}\{-(\sigma_N^2)^{-1}(\mathbf{x} - BH\beta)^T(\mathbf{x} - BH\beta) - (-(\sigma_N^2)^{-1}\mathbf{x}^T\mathbf{x})\} \tag{10}$$

The interesting results reveal that the detector could be factored into simple problems consisted of (9) and (10). Fortunately, Matrix B and H do not need a great quantity of computation because the estimation of pitch and linear predictive coefficients is completed by the LP-structure speech coder. The process is shown in the figure:

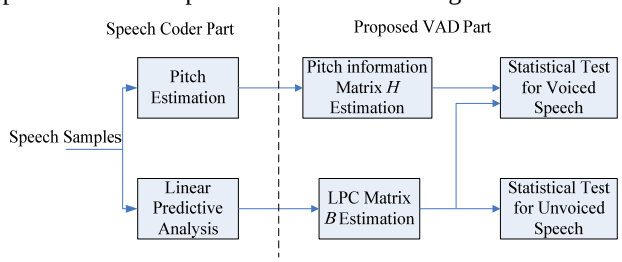

Figure 1. The Proposed Detector Cooprated with LP-Structure Speech Coder.

A. L1 — Detector of Voiced Speech

With the Gaussian assumption and model linearity, the minimum variance unbiased (MVU) estimator of unknown parameters β should be directly calculated by:

$$\hat{\beta} = ((BH)^T BH)^{-1}(BH)^T \mathbf{x}$$

Adopting the discussions above, the voiced speech detector can be described as:

$$L_1(\mathbf{x}) = \mathbf{x}^T BH((BH)^T BH)^{-1}(BH)^T \mathbf{x} \quad (11)$$

B. *L0 — Detector of Unvoiced Speech*

T_0, a replacement of L_0 is used for convenience:

$$T_0(\mathbf{x}) = \sigma_N^2 \mathbf{x}^T [\frac{1}{\sigma_N^2} I - (\sigma_R^2 BB^T + \sigma_N^2 I)^{-1}]\mathbf{x} > 2\sigma_N^2 \gamma'$$

In a speech coding application, SNR of the unvoiced speech is relatively low while prediction gain is large, and thus variance of source of unvoiced speech is far smaller than that of the additive noise: $\sigma_R^2 << \sigma_N^2$ .Exploiting that property and omitting the irrelevant variables and parameters, the detector can be finally simplified:

$$T_0(\mathbf{x}) = (B^T \mathbf{x})^T (B^T \mathbf{x}) \quad (12)$$

The demonstration and deduction would not be elaborated herein.

IV. Experimental Results

To prove the effectiveness of proposed unvoiced speech detector, this section will provide our experimental results with comparison with other feature extraction method by Receiver Operating Characteristic (ROC) curves.

According to the analysis of two sub-detectors, L_1 and L_0 detectors, are designed to detect voiced speech signal and unvoiced speech signal, respectively. L_1 detector is well-known to be an effective one to detect the harmonic signal, which represents most of voiced speech. Thus, the experiment only focus on evaluating the performance of L_0 detector with the database composed of unvoiced speech signals. Speech signals are uttered by males, females and children including multiple kinds of languages like English, Chinese, French, and Japanese. Exactly, one thousand unvoiced speech frames are manually separated from the natural speech samples and mixed with Gaussian noise with different SNR.

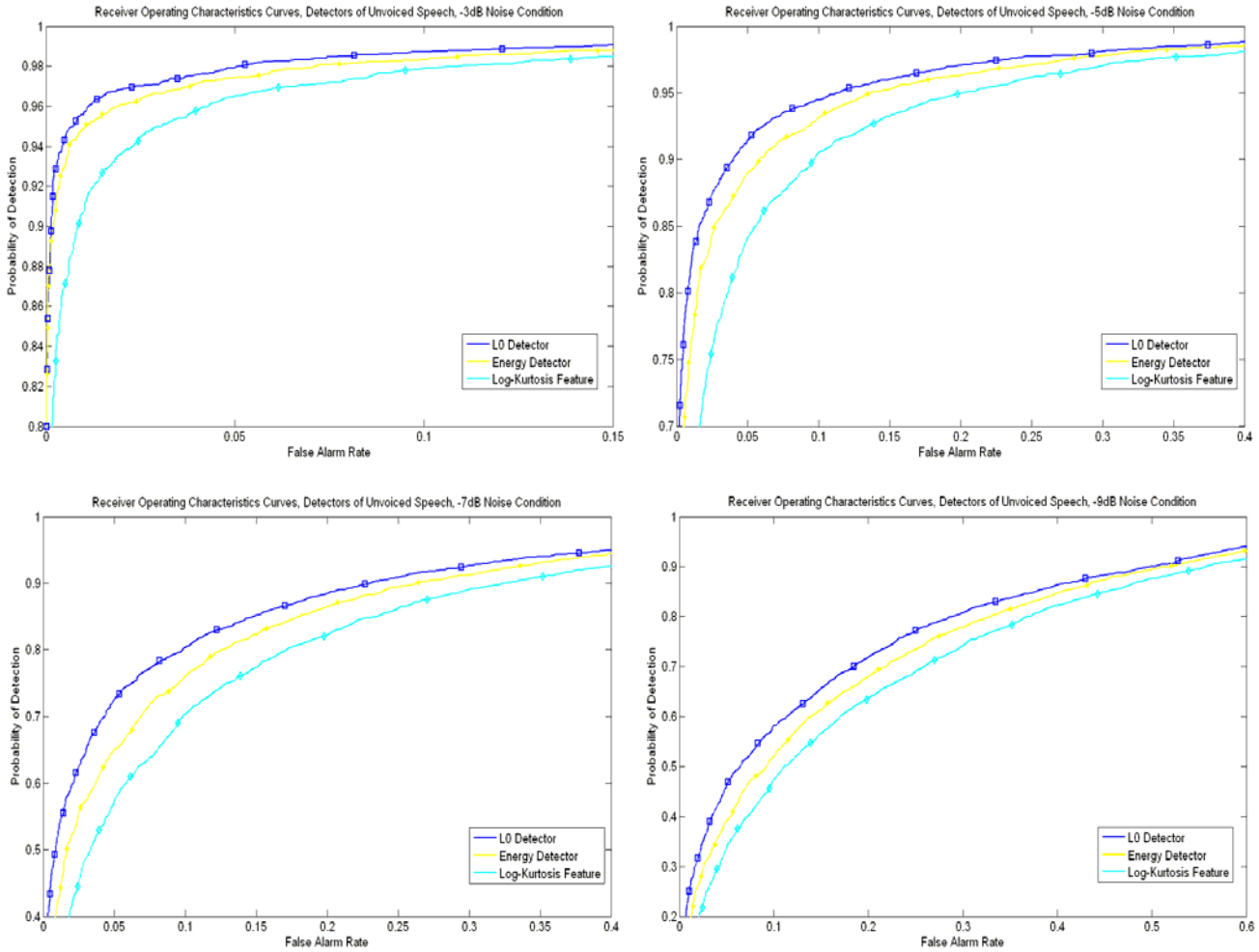

Figure 2. ROC curves of propose *L0* detector, log-kurtosis feature, and energy detector of unvoiced speech signal at -3dB, -5dB, -7dB and -9dB SNR.

ROC curves depict the relationship between false alarm rate and probability of detection in a certain SNR. Probability of detection refers to the probability of correctly detecting the unvoiced speech frames while false alarm rate refers to the probability of wrongly classifying the noise frames as the unvoiced speech frame. Three methods, proposed L_0 detector, log-kurtosis feature[7] and energy detector[5], were compared together. And the additive noise is the white noise from NOISEX92 database[12].

From four figures, proposed detector nearly increases the probability of detection in all SNR conditions and the false alarm rates are lower compared to the other two metrics. In fact, the traditional energy detector could be simplified from proposed L_0 detector by approximately adopting $BB^T = I$:

$$T'(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$$

It is just the traditional detector with the assumption that unvoiced speech is a white Gaussian process. In this view, the significant improvement of proposed L_0 detector over tradition energy detector lies in the incorporation of a more proper model (vocal tract information) for the unvoiced speech in the detection problem.

V. CONCLUSION

In this paper, we proposed an unvoiced speech detector, which is applicable to Linear Prediction structure speech coder. It revisits voiced or unvoiced speech production model: the stochastic/deterministic Gaussian source signal modulated by vocal tract model, which is involved with some unknown parameters including linear predictive coefficients, source parameters, pitch information, and noise parameters. According to the maximum likelihood estimation rule, the statistical test could be divided into two sub-detectors, and the outperformance of the unvoiced speech detector has been verified in previous section. By exploiting the Linear Predictive coefficients and pitch information which are products of the speech coder, the proposed detector works effectively with little increase in computation.

ACKNOWLEDGMENT

This work was supported by Zhejiang Provincial Natural Science Foundation of China under Grant LY12F01019 and LY12F01020, and Science Technology Department of Zhejiang Province under Grant 2010R50006

REFERENCES

[1] S.A. Wibowo, K. Usman, “Voice Activity Detection G729B Improvement technique using K-nearest Neighbor Method,” International Conference on Distributed Framework and Applications, 2010, pp. 1-5.

[2] Y. Ephraim, D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Trans. Acoustic, Speech and Signal Processing, vol. 32(6) pp. 1109-1121, 1984. doi: 10.1109/TASSP.1984.1164453.

[3] D. Zeddelmann, “A Feature-based Approach to Noise Robust Speech Detection,” Speech Communication; 10. ITG Symposium; Proceedings of Topic(s): Fields, Waves & Electromagnetics, , 2012, pp: 1 – 4.

[4] J.W. Shin, H.J. Kwon, S.H. Jin, N.S. Kim, ”Voice Activity Detection Based on Conditional MAP Criterion,” IEEE Signal Processing Letters, 2008, vol. 15, pp. 257-260. doi: 10.1109/LSP.2008. 917027.

[5] ITU-T G.729 Annex B: A silence compression scheme for G.729 optimized for terminal fonforming to Recommendation V.70, 1996.

[6] D. Cournapeau, T. Kawahara, “Evaluation of Real-time Voice Activity Detection based on High Order Statistics,” Interspeech 2007: 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, pp. 2945-2948, 2007.

[7] K. Li, M.N.S. Swamy, M.O. Ahmad, “An Improved Voice Activity Detection Using Higher rder Statistics,” IEEE Trans. Speech and Audio Processing, vol. 13(5) ppp. 965-974, 2005. doi: 10.1109/TSA.2005. 851955.

[8] E. Nemer, R. Goubran, S. Mahmoud, “Robust Voice Activity Detection Using Higher-Order Statistics in the LPC Residual Domain,” IEEE Trans. Speech and Audio Processing, 9(3):217-231, 2001. doi: 10.1109/89.905996.

[9] R. Tucker, ”Voice Activity Detection Using a Periodicity Measure,” IEE Proceedings-I, vol. 139(4) pp. 377-380, 2002.

[10] T. Petsatodis, C. Boukis, F. Talantzis, Z.H. Tan, R. Prasad, “Convex Combination of Multiple Statistical Models With Application to VAD,” IEEE trans. Audio, Speech, and Language Processing, 2011, vol. 19(8) pp. 2314-2327. doi: 10.1109/TASL.2011.2131131.

[11] K.Sakhnov, E. Verteletskaya, B. Simak, “Low-complexity Voice Activity Detector Using Periodicity and Energy Ratio,” 16th IEEE Int. Conf. on Systems, Signals and Image Processing (IWSSIP), 2009, pp.1-5. doi: 10.1109/IWSSIP.2009.5367799.

[12] NOISEX92 databaase, http://spib.ece.rice.edu/spib/select_noise.html.