Voice Activity Detection Based on Complex Exponential Atomic Decomposition and Likelihood Ratio Test

Shiwen Deng¹, ², Jiqing Han¹

¹School of Computer Science and Technology, Harbin Institute of Technology
No.92, West Da-Zhi Street, 150001, Harbin, China
²School of Mathematical Sciences, Harbin Normal University
No.50, West Hexing Road, 150080, Harbin, China
E-mail: dengswen@gmail.com

Abstract

The voice activity detection (VAD) algorithms by using Discrete Fourier Transform (DFT) coefficients are widely found in literature. However, some shortcomings for modeling a signal in the DFT can easily degrade the performance of a VAD in noise environment. To overcome the problem, this paper presents a novel approach by using the complex coefficients derived from complex exponential atomic decomposition of a signal. Those coefficients are modeled by a complex Gaussian probability distribution and a statistical model is employed to derive the decision rule from the likelihood ratio test. According to the experimental results, the proposed VAD method shows better performance than the VAD based on DFT coefficients in various noise environments.

1. Introduction

Voice activity detection (VAD), which refers to the problem of distinguishing active speech from nonspeech, plays an important role on the performance of speech processing systems such as robust speech recognition, speech enhancement, and coding systems. Various traditional VAD algorithms have been proposed based on the energy, zero-crossing rate, and spectral difference in earlier literature. However, these algorithms are easily degraded by environmental noise.

Recently, much work for improving the performance of the VADs in various high noise environments has been carried out by incorporating a statistical model and a likelihood ratio test. Those algorithms assume that the distributions of the noise and the noisy speech spectra are specified in terms of some certain parametric models such as complex Gaussian [1], complex Laplacian [2], generalized Gaussian [3], or generalized Gamma distribution. Moreover, some algorithms based on the multiple observation likelihood test (MO-LRT) and the modified maximum a posteriori (MAP) criterion are proposed in [4-6].

Most of above methods are operated in the DFT domain by classifying each sound frame into speech or noise according to the DFT coefficients, on which the performance of VAD scheme is strongly depended. However, the DFT, being a method of orthogonal basis expansion, mainly suffers two serious drawbacks. One is that a given Fourier basis is not well-suited for modeling a wide variety of signals such as speech [7-9]. The other is the problem of spectra components interference between the two components in adjacent frequency bins [7, 8]. Therefore, the performance of a VAD by using the DFT coefficients is easily degraded by environmental noise.

In this paper, we present an approach based on the conjugate subspace matching pursuit to avoid the drawbacks in the DFT. The matching pursuit is carried out in each frame by first selecting the most dominant component, then subtracting its contribution from the signal and iterating the estimation on the residual. By subtracting a component at each iteration, the next
component selected in the residual does not interfere with the previous component. Subsequently, the coefficients extracted in each frame are modeled in complex Gaussian distribution and the likelihood ratio test is employed as well. Experimental results indicate that the proposed VAD algorithm shows better results compared with the conventional algorithms based on the DFT coefficients in various noise environments.

2. Signal atomic decomposition based on conjugate subspace matching pursuit

In this section, we will briefly review the method of matching pursuit based on the conjugate subspace [7, 8]. The finite length signal, \( x \), can be approximately represented as

\[
    x \approx \sum_{k=1}^{K} \alpha_k g_{g_k},
\]

(1)

where \( K \) and \( \{ \alpha_k \} \) denote the order of decomposition and the expansion coefficients, respectively. And \( \{ g_{g_k} \} \) are the atoms chosen from a dictionary whose element consists of complex exponentials such that

\[
    g_n = S e^{j n \omega_i}, \quad n = 0, \ldots, N-1,
\]

(2)

where \( S \) is a constant in order to obtain unit-norm function, \( i \) and \( n \) are frequency and time index. Moreover, the complex exponential dictionary is denoted as \( D = \{ g_1, g_2, \ldots, g_M \} \) where \( M \) is the number of dictionary elements such that \( M > N \). And, the conjugate subspace matching pursuit is carried out as follows [7].

Given the initial condition \( r_0 = x \), let \( r_k \) denote the residual signal after \( k-1 \) iterations. At the \( k \)-th stage, the new residual is given by

\[
    r_{k+1} = r_k - 2 \text{Re}\{ \alpha_k g_{g_k} \},
\]

(3)

where \( g_{g_k} \in D \) such that

\[
    g_{g_k} = \arg \max_{g \in D} \left\{ \text{Re}\{ \langle g, r_k \rangle^* \alpha_k \} \right\}.
\]

(4)

The projection coefficient of the residual \( r_k \) over the conjugate subspace \( \text{span}\{ g, g^* \} \) is obtained by

\[
    \alpha_k = \frac{1}{1 - |c|^2} \left( \langle g, r_k \rangle - c \langle g, r_k \rangle^* \right),
\]

(5)

where \( g^* \) is the conjugate of \( g \) and \( c = \langle g, g^* \rangle \) is the conjugate cross-correlation.

To obtain atomic decomposition of a signal, the MP iteration is continued until a halting criterion is met. After \( K \) iterations, the decomposition of \( x \) corresponds to the estimate

\[
    x \approx 2 \sum_{k=1}^{K} \text{Re}\{ \alpha_k g_{g_k} \},
\]

(6)

where \( \{ \alpha_k \}_{k=1}^{K} \) are the complex MP coefficients of atomic decomposition.

3. Decision rule based on MP coefficients and LRT

In this section, the VAD based on the MP coefficients and LRT is first presented in subsection 3.1. Then, more details of the MP coefficients are discussed in subsection 3.2.

3.1. Statistical model and decision rule

Assuming that the noisy speech \( x \) consists of a clean speech \( s \) and an additive noise \( n \), that is

\[
    x = s + n.
\]

(7)

The MP coefficient extracted from \( x \) at each iteration has the following form

\[
    \alpha_k = \alpha_{s,k} + \alpha_{n,k}, \quad k = 1, \ldots, K,
\]

(8)

where \( \alpha_{s,k} \) and \( \alpha_{n,k} \) are the MP coefficient of clean speech and noise, respectively.

The \( K \)-dimensional MP coefficient vectors of speech, noise, and noisy speech are denoted as \( \mathbf{a}_s, \mathbf{a}_n \), and \( \mathbf{a} \) with their \( k \)-th elements \( \alpha_{s,k}, \alpha_{n,k} \), and \( \alpha_k \), respectively. Given two hypotheses \( H_0 \) and \( H_1 \), which indicate speech absence and presence, we assume that

\[
    H_0: \mathbf{a} = \mathbf{a}_s, \text{ speech absence}
\]

\[
    H_1: \mathbf{a} = \mathbf{a}_n + \mathbf{a}_s, \text{ speech presence}.
\]

The MP coefficients of noisy speech and noise signal are asymptotically independent complex Gaussian random variables. Moreover, the MP coefficients of clean speech and noise signal have zero mean and their variance satisfy

\[
    \lambda_k = \lambda_{s,k} + \lambda_{n,k}, \quad k = 1, \ldots, K.
\]

(9)

where \( \lambda_s, \lambda_{s,k} \) and \( \lambda_{n,k} \) are the variance of MP coefficients of noisy signal, clean speech, and noise, respectively. Thus, the probability density functions conditioned on \( H_0 \) and \( H_1 \) are given by

\[
    p(\mathbf{a} \mid H_0) = \prod_{k=1}^{K} \frac{1}{\pi \lambda_{s,k}} \exp \left\{ -\frac{\left| \alpha_{s,k} \right|^2}{\lambda_{s,k}} \right\}
\]

(10)

\[
    p(\mathbf{a} \mid H_1) = \prod_{k=1}^{K} \frac{1}{\pi \left( \lambda_{s,k} + \lambda_{n,k} \right)} \exp \left\{ -\frac{\left| \alpha_{s,k} \right|^2}{\lambda_{s,k} + \lambda_{n,k}} \right\}.
\]

(11)
where $\lambda_{n,k}$ can be estimated through the noise statistic estimation procedure. In addition, the maximum likelihood estimate of $\hat{\lambda}_{n,k}$, $\hat{\lambda}_{n,k}$, is obtained by

$$
\hat{\lambda}_{n,k} = \left| |\alpha_{n,k}|^2 - \lambda_{n,k} \right|.
$$

(12)

Furthermore, the decision rule using log likelihood ratio is defined by

$$
\log \Lambda = \frac{1}{K} \sum_{k=1}^{K} \left\{ \log \sigma_{n,k}^2 - \log \frac{\sigma_{n,k}^2}{\lambda_{n,k}} - 1 \right\} \geq \eta, \tag{13}
$$

where $\eta$ denotes a threshold value.

### 3.2. MP coefficients

As mentioned before, the DFT coefficients suffer several shortcomings for modeling a signal and exposing the signal structure. We use the MP coefficients, $\{\alpha_{n,k}\}$, obtained by MP as the new feature for discriminating speech and nonspeech. With the advantage of the atomic decomposition, MP coefficients can capture the characteristics of speech [8] and are insensitive to environment noise. Therefore, the MP coefficients are more suitable for the classification task than DFT coefficients.

MP coefficients, which are obtained in Eq. (6), reveal the harmonic structures of a signal. Such harmonic components can be viewed as a sinusoid and has the following form

$$
A \cos(\omega_i + \phi) = 2 \text{Re}\left\{\alpha_i g_n\right\}, \tag{14}
$$

where $A$ is the amplitude of the sinusoidal component, $\omega_i$ and $\phi$ are its frequency and phase, respectively. Furthermore, those structures are prominent in a signal when the speech is present but not when noise only.

The extraction process for MP coefficients is given as follows. Assuming the input signal is decomposed into non-overlapping frames, each frame is decomposed by conjugate subspace matching pursuit. Thus, the complex MP coefficients corresponding to a frame are obtained. Instead of requiring a full reconstruction of a signal, the goal of MP is the extraction of MP coefficients which capture the most characters of a signal in order to detect whether or not the speech is present. The selection of iteration number $K$ depends on the number of sinusoidal components in a speech signal.

### 4. Experiments and analysis

In this section, the experimental results of our method are presented. Fig. 1(a) and (b) illustrate the statistical properties of MP coefficients of the clean speech signal corrupted by white noise and factory noise from the NOISEX database at 0dB SNR. As can be seen, the empirical cumulative distribution function (CDF) curves of noisy speech signal are much closed to the Gaussian CDF and insensitive to noise.

To examine the effectiveness of the proposed VAD, a test signal (Fig. 2(b)) is created by adding white noise to a clean speech (Fig. 2(a)) at an SNR of 0dB. The test signal is divided into non-overlapping frames in which frame length is 256 points. The atomic decomposition based on the conjugate subspace MP is operated on the test signal. The likelihood ratios and the results for VAD calculated with Eq. (13) are shown in Fig. 2(c)-(d), respectively. As can be seen, even at such a low SNR, the results also correctly indicate the speech presence and verify the effectiveness of MP coefficients in VAD.

The receiver operating characteristic (ROC) curves show the trade-off between the reduction of the false alarm probability ($P_f$) and the increase in detection probability ($P_d$) in Fig. 2. It is used to evaluate the performance of the proposed VAD in the different selection of iteration number $K$ and to compare the proposed method with the VAD using DFT coefficients and G.729 VAD. In this analysis, the test speech material to verify the effectiveness of VAD.
The performances of the VAD algorithms are summarized in Table 1 where the false alarm probability $P_f$ and detection probability $P_d$ are obtained by two VAD methods in various noise environments. The experimental results show that the VAD based on MP coefficients outperforms the ones based on DFT in all of the testing conditions and it can be concluded that the MP coefficients are more robust to background noise than DFT.

5. Conclusion

This paper presents a robust voice activity detection algorithm, which effectively avoids drawbacks in the DFT by using the conjugate subspace MP and the LRT. It has been shown that the MP coefficients perform better than the DFT coefficients when the iteration number in the atomic decomposition of a signal is larger than 10. Furthermore, the experimental results show that the proposed approach outperforms the VAD the method using DFT coefficients in various noise environments.

References