Deng and Han *EURASIP Journal on Audio, Speech, and Music Processing* 2011, **2011**:12
http://asmp.eurasipjournals.com/content/2011/1/12

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Voice activity detection based on conjugate subspace matching pursuit and likelihood ratio test

Shiwen Deng[1,2] and Jiqing Han[1*]

Abstract

Most of voice activity detection (VAD) schemes are operated in the discrete Fourier transform (DFT) domain by classifying each sound frame into speech or noise based on the DFT coefficients. These coefficients are used as features in VAD, and thus the robustness of these features has an important effect on the performance of VAD scheme. However, some shortcomings of modeling a signal in the DFT domain can easily degrade the performance of a VAD in a noise environment. Instead of using the DFT coefficients in VAD, this article presents a novel approach by using the complex coefficients derived from complex exponential atomic decomposition of a signal. With the goodness-of-fit test, we show that those coefficients are suitable to be modeled by a Gaussian probability distribution. 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 the DFT coefficients in various noise environments.

Keywords: voice activity detection, matching pursuit, likelihood ratio test, complex exponential dictionary

1 Introduction

Voice activity detection (VAD) refers to the problem of distinguishing active speech from non-speech regions in an given audio stream, and it has become an indispensable component for many applications of speech processing and modern speech communication systems [1-3] 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 [1,4,5]. However, these algorithms are easily degraded by environmental noise.

Recently, much study 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 (LRT) [6]. 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 [7], complex Laplacian [8], generalized Gaussian [9], or generalized Gamma distribution [10]. Moreover, some algorithms based on LRT consider more complex statistical structure of signals, such as the multiple observation likelihood ratio test (MO-LRT) [11,12], higher order statistics (HOS) [13,14], and the modified maximum *a posteriori* (MAP) criterion [15,16].

Most of the above methods are operated in the DFT domain by classifying each sound frame into speech or noise based on the complex DFT coefficients. These coefficients are used as features, and thus the robustness of these features has an important effect on the performance of VAD scheme. 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 [17-20]. The other is the problem of spectra components interference between the two components in adjacent frequency bins [19,20]. Figure 1 presents an example that demonstrates the drawbacks of the DFT. The DFT coefficients of a signal with five frequency components, 100, 115, 130, 160, and 200 Hz, are shown in Figure 1a and its accurate frequencies components (A, B, C, D, and E) are shown in Figure 1b. As shown in Figure 1a, first, except these frequencies components corresponding to the accurate frequencies, many other frequency components are also emerged in

* Correspondence: jqhan@hit.edu.cn
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
Full list of author information is available at the end of the article

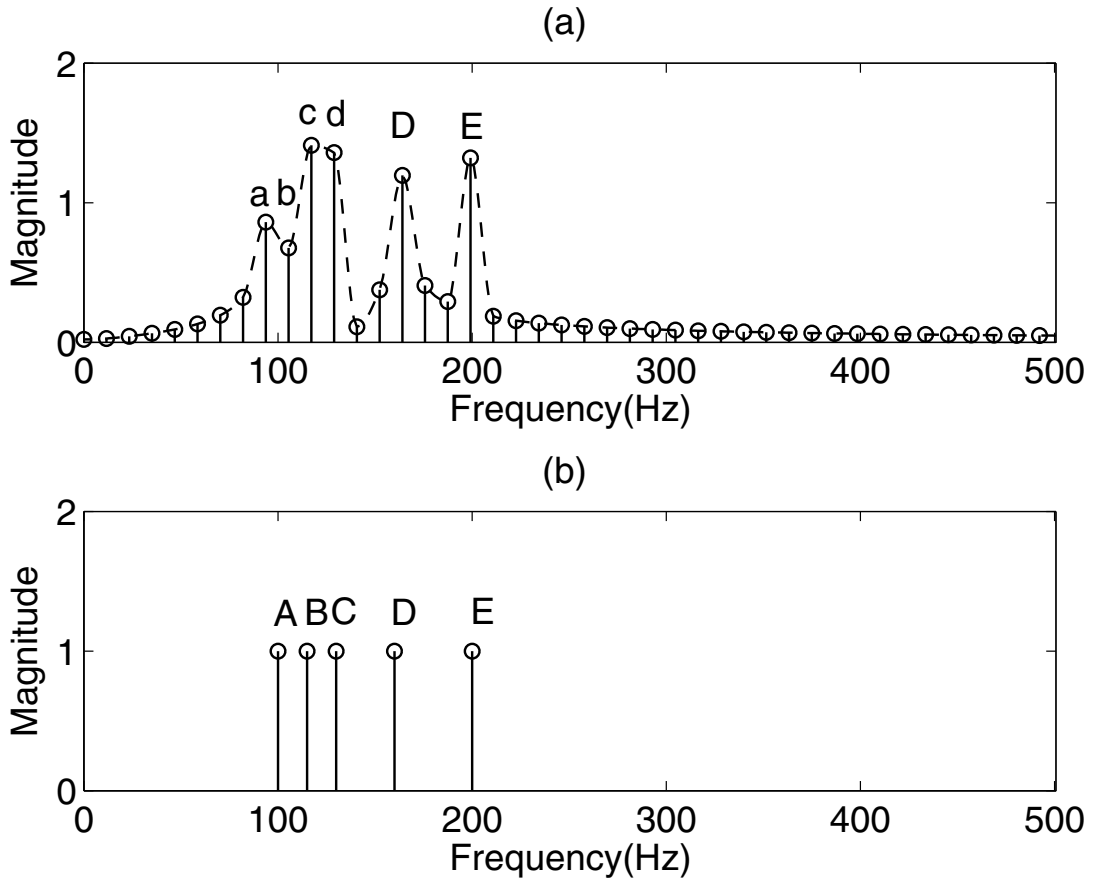

Figure 1 Drawbacks of the DFT coefficients. **(a)** The DFT coefficients of a signal with frequencies: 100, 115, 130, 160, 200 Hz; **(b)** the accurate frequency components of the signal.

the DFT coefficients all over the whole frequency bins. Second, there exists the problem of spectra components interference at a, b, c, and d frequency bins, because the corresponding accurate frequencies at A, B, C in Figure 1b are too adjacent to each other.

In this article, we present an approach for VAD based on the conjugate subspace matching pursuit (MP) and the statistical model. Specifically, the MP 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, named MP feature [21], are modeled in complex Gaussian distribution, and the LRT 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.

The rest of this article is organized as follows. Section 2 reviews the method of the conjugate subspace MP. Section 3 presents our proposed approach for VAD based the MP coefficients and statistical model. Implementation issues and the experimental results are shows in Section 4. Section 5 concludes this study.

2 Signal atomic decomposition based on conjugate subspace MP

In this section, we will briefly review the process of signal decomposition by using the conjugate subspace MP [19,20]. The conjugate subspace MP algorithm is described in Section 2.1, and the demonstration of algorithm and comparison between MP coefficients and DFT coefficients are presented in Section 2.2.

2.1 Conjugate subspace MP

Matching pursuit is an iterative algorithm for deriving compact signal approximations. For a given signal $x \in R^N$, which can be considered as a frame in a speech, the compact approximation $\hat{x}$ is given by

$$\hat{x} \approx \sum_{k=1}^{K} \alpha_k g_{\gamma_k} \tag{1}$$

where K and $\{\alpha_k\}_{k = 1,\ldots,K}$denote the order of decomposition and the expansion coefficients, respectively, and $\{g_{\gamma_k}\}_{k=1,\ldots,K}$ are the atoms chosen from a dictionary whose element consists of complex exponentials such that

$$g_i = Se^{jw_i n}, \quad n = 0, \ldots, N-1, \tag{2}$$

where i and n are frequency and time indexes, and S is a constant in order to obtain unit-norm function. The complex exponential dictionary is denoted as $\mathbf{D} = [g_1, \ldots, g_M]$ where M is the number of dictionary elements such that $M > N$. Note that, this dictionary contains the prior knowledge of the statistical structure of the signal that we are mostly interested in. Here, the prior knowledge is that speech is the sum of some complex exponential with complex weights. And hence, speech can be represented by a few atoms in dictionary, but noise is not.

The conjugate subspace MP is a method of subspace pursuit. In the subspace pursuit, the residual of a signal is projected into a set of subspaces, each of which is spanned by some atoms from the dictionary, and the most dominant component in the corresponding subspace is selected and subtracted from the residual. Each of the subspaces in the conjugate subspace MP is the two-dimensional subspace spanned by an atom and its complex conjugate. With the given complex dictionary, the conjugate subspace MP is operated as follows.

Let r_k denotes the residual signal after k - 1 pursuit iterations, and the initial condition is $r_0 = x$. At the kth iteration, the new residual r_{k+1} is given by

$$r_{k+1} = r_k - 2Re\{\alpha_k g_{\gamma_k}\}, \tag{3}$$

where α_k is a complex coefficient, $Re\{\cdot\}$ denotes the real part of a complex value, and g_{γ_k} is the atom selected from the dictionary $\mathbf{D}$ given by

$$g_{\gamma_k} = \arg\max_{g \in D}(\mathrm{Re}\{< g, r_k >^* \alpha_k\}), \tag{4}$$

where the superscript * denotes conjugate transpose. The projection coefficient of the residual r_k over the conjugate subspace span $\{g, g^*\}$, α_k, is obtained by

$$\alpha_k = \frac{1}{1 - |c|^2}(< g, r_k > - c < g, r_k >^*), \tag{5}$$

where g^* is the complex conjugate of g and $c =< g, g^* >$is the conjugate cross-correlation coefficient. 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

$$\hat{x} \approx 2\sum_{k=1}^{K} \mathrm{Re}\{\alpha_k g_{\gamma_k}\}, \tag{6}$$

where $\{\alpha_k\}_{k=1}^{K}$ are referred to as the complex MP coefficients of atomic decomposition.

2.2 Demonstration of algorithm and comparison between MP coefficients and DFT coefficients

In this section, we present an example to demonstrate the procedure of the decomposition and compare the MP coefficients with DFT coefficients. Let $x[m]$ be the original signal defined by a sum of five sinusoids as follows

$$x[m] = \sum_{i=1}^{5} \cos(2\pi m f_i / F_s), \quad \text{for } m = 1, 2, \ldots$$

where F_s = 4, 000 Hz is the sample frequency, and the frequencies f_1, f_2, ..., f_5 are 100, 115, 130, 160, and 200 Hz, respectively.

The noisy signal $y[m]$ is given by $y[m] = x[m] + n$, where n is the uncor-related additive noise. Figure 2a shows a 256 sample segment selected by a Hamming window from $y[m]$, the corresponding DFT coefficients are shown in Figure 2b,c that shows the accurate frequency components of $x[m]$. The procedure of the MP decomposition of five iterations is shown in Figure 3. In each iteration, the component with the maximum of Re $\{< g, r_k >^* \alpha_k\}$ is selected as shown in the left column in Figure 3, and, the corresponding α_k is the MP coefficient in the kth iteration. The extracted components $2\mathrm{Re}\{\alpha_k g_{\gamma k}\}$ at the kth iteration is shown in the right column in Figure 3 and is subtracted from the current residual r_k to obtain the next residual r_{k+1} according to Equation (3). After five iterations, we can obtain five MP coefficients α_1, ..., α_5, whose magnitudes are shown in Figure 2d.

As shown in Figure 2, the MP coefficients accurately capture all the frequency components of the original signal $x[m]$ from the noisy signal $y[m]$, but the DFT coefficients only capture two frequency components of $x[m]$.

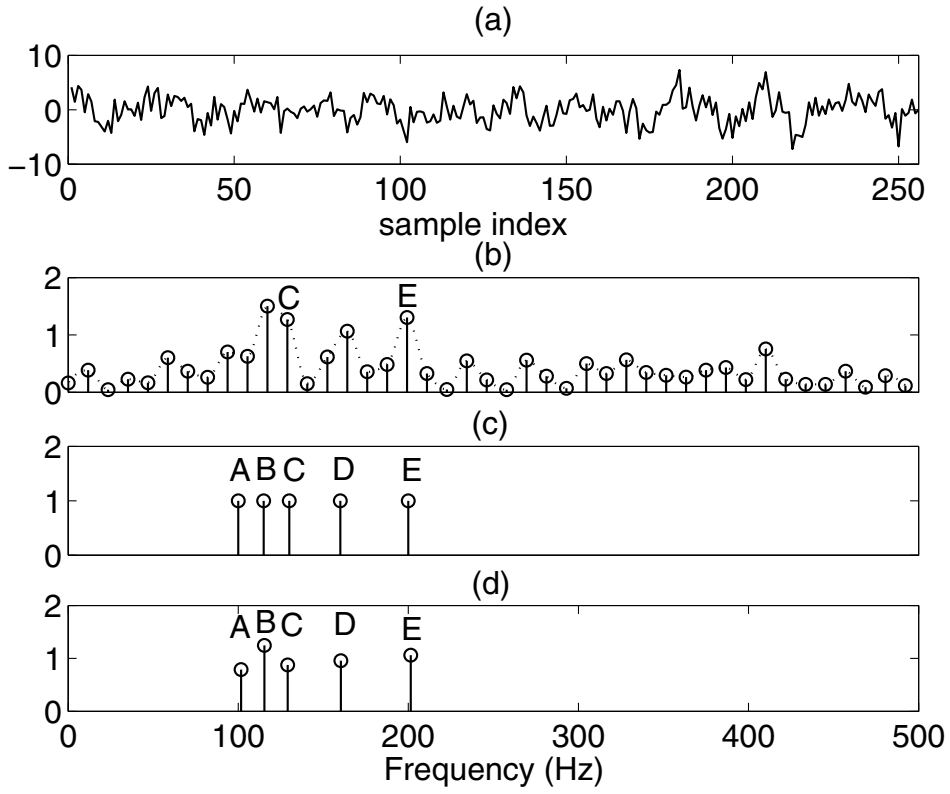

Figure 2 Decomposition of a noisy signal by DFT and the conjugate subspace MP. **(a)** The noisy signal; **(b)** the DFT coefficients of the noisy signal; **(c)** the accurate frequency components of the original signal; **(d)** the MP coefficients of the noisy signal after five iterations.

On the other hand, the MP coefficients well represent the frequency components without the problem of the spectra components interference, such as these components at A, B, and C shown in Figure 2d, but the DFT coefficients fail to do this even in the noise-free case. Therefore, the MP coefficients are more robust that the DFT coefficients, and are not sensitive to the noise.

3 Decision rule based on MP coefficients and LRT

In this section, the VAD based on the MP coefficients and LRT is presented in Section 3.1. To test the distribution of the MP coefficients, a goodness-of-fit test (GOF) for those coefficients is provided in Section 3.2. More details about the MP feature are discussed in Section 3.3.

3.1 Statistical modeling of the MP coefficients and decision rule

Assuming that the noisy speech x consists of a clean speech s and an uncorrelated additive noise signal n, that is

$$x = s + n \tag{7}$$

Applying the signal atomic decomposition by using the conjugate MP, the noisy MP coefficient extracted from x at each pursuit iteration has the following form

$$\alpha_k = \alpha_{s,k} + \alpha_{n,k}, \quad k = 1, \ldots, K, \tag{8}$$

where $\alpha_{s,k}$ and $\alpha_{n,k}$ are the MP coefficients of clean speech and noise, respectively. The variance of the noisy MP coefficient α_k is given by

$$\lambda_k = \lambda_{s,k} + \lambda_{n,k}, \quad k = 1, \ldots, K. \tag{9}$$

where $\lambda_{s,k}$ and $\lambda_{n,k}$ are the variances of MP coefficients of clean speech and noise, respectively.

The K-dimensional MP coefficient vectors of speech, noise, and noisy speech are denoted as $\boldsymbol{\alpha}_s$, $\boldsymbol{\alpha}_n$, and $\boldsymbol{\alpha}$ with their kth elements $\alpha_{s,k}$, $\alpha_{n,k}$, and α_k, respectively. Given two hypotheses H_0 and H_1, which indicate speech absence and presence, we assume that

$$H_0 : \boldsymbol{\alpha} = \boldsymbol{\alpha}_n$$
$$H_1 : \boldsymbol{\alpha} = \boldsymbol{\alpha}_n + \boldsymbol{\alpha}_s$$

For implementation of the above statistical model, a suitable distribution of the MP coefficients is required.

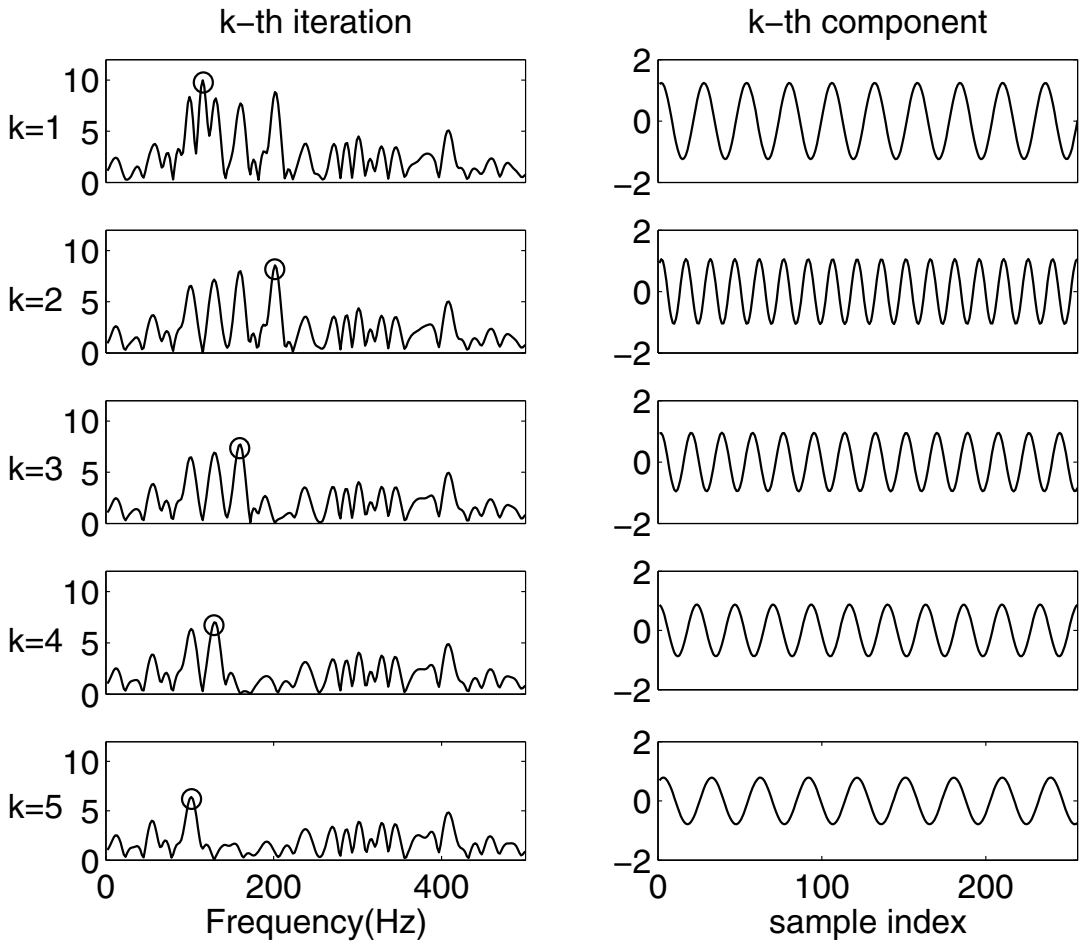

Figure 3 Five iterations of the MP for a noisy signal. The left column shows each iteration of the MP and the selected component is marked by a open circles; the right column shows the corresponding signal component extracted at each iteration.

In this article, we assume that the MP coefficients of noisy speech and noise signal are asymptotically independent complex Gaussian random variables with zero means. We also assume that the variances of the MP coefficient of noise, $\{\lambda_{n,k}, k = 1, ..., K\}$ are known. Thus, the probability density functions (PDFs) conditioned on H_0, and H_1 with a set of K unknown parameters $\Theta = \{\lambda_{s,k}, k = 1, ..., K\}$, are given by

$$p(\boldsymbol{\alpha}|H_0) = \prod_{k=1}^{K} \frac{1}{\pi\lambda_{n,k}} \exp\left\{-\frac{|\alpha_k|^2}{\lambda_{n,k}}\right\} \tag{10}$$

$$p(\boldsymbol{\alpha}|\Theta, H_1) = \prod_{k=1}^{K} \frac{1}{\pi\left(\lambda_{n,k} + \lambda_{s,k}\right)} \exp\left\{-\frac{|\alpha_k|^2}{\lambda_{n,k} + \lambda_{s,k}}\right\} \tag{11}$$

The maximum likelihood estimate $\hat{\Theta} = \{\hat{\lambda}_{s,k}, k = 1, \ldots, K\}$ of Θ is obtained by

$$\hat{\Theta} = \arg\max_{\Theta}\{\log p(\boldsymbol{\alpha}|\Theta, H_1)\}, \tag{12}$$

and equals

$$\hat{\lambda}_{s,k} = |\alpha_k|^2 - \lambda_{n,k}, \quad k = 1, \ldots, K. \tag{13}$$

By substituting Equation (13) into Equation (11), the decision rule using the likelihood ratio is obtained as follows

$$\begin{aligned}\Lambda_g &= \tfrac{1}{K}\log\tfrac{p(\boldsymbol{\alpha}|\hat{O},H_1)}{p(\boldsymbol{\alpha}|H_0)} \\ &= \frac{1}{K}\sum_{k=1}^{K}\left\{\tfrac{|\alpha_k|^2}{\lambda_{nk}} - \log\tfrac{|\alpha_k|^2}{\lambda_{nk}} - 1\right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta\end{aligned} \tag{14}$$

where η denotes a threshold value.

3.2 GOF test for MP coefficients

The MP coefficients are considered to follow a Gaussian distribution in section above. To test this, we carried out a statistical fitting test for the noisy MP coefficients conditioned on both hypotheses under various noise conditions. To this end, the Kolomogorov-Sriminov (KS) test [22], which serves as a GOF test, is employed to guarantee a reliable survey of the statistical assumption.

With the KS test, the empirical cumulative distribution function (CDF) F_α is compared to a given distribution function F, where F is the complex Gaussian function. Let $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, ..., \alpha_N\}$ be a set of the MP coefficients extracted from the noisy speech data, and the empirical CDF is defined by

$$F_\alpha = \begin{cases} 0, & z < \alpha_{(1)} \\ \dfrac{n}{N}, & \alpha_{(n)} \le z < \alpha_{(n+1)}, \quad n = 1, \ldots, N \\ 1, & z \le \alpha_{(N)} \end{cases} \tag{15}$$

where $\alpha(n)$, $n = 1, ..., N$ are the order statistics of the data $\boldsymbol{\alpha}$. To compute the order statistics, the elements of $\boldsymbol{\alpha}$ are sorted and ordered so that $\alpha_{(1)}$ represents the smallest element of $\boldsymbol{\alpha}$ and $\alpha_{(N)}$ is the largest one.

For simulating the noisy environments, the white and factory noises from the NOISEX'92 database are added to a clean speech signal at 0 dB SNR. With the noisy speech, the mean and variance are calculated and substituted into the Gaussian distribution. Figure 4 shows the comparison of the empirical CDF and Gaussian function. As can be seen, the empirical CDF curves of noisy speech signal are much closed to that of the Gaussian CDF under both the white and factory noise conditions. Therefore, the Gaussian distribution is suitable for modeling the MP coefficients.

3.3 Obtaining MP features

As mentioned before, the DFT coefficients suffer several shortcomings for modeling a signal and exposing the signal structure. We use the MP coefficients, $\{\alpha_k\}_{k=1}^K$, obtained by the 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 [17] and are insensitive to environment noise. Therefore, the MP coefficients as a new feature for VAD are more suitable for the classification task than DFT coefficients.

With the decomposition of a speech signal by using the conjugate MP, the MP feature also captures the harmonic structures of the speech signal. Such harmonic components can be viewed as a series of sinusoids, which are buried in noise, with different amplitude, frequency, and phase. The kth harmonic component h_k extracted from the kth pursuit iteration has the following form

$$h_k = A_k \cos(\omega_k + \phi_k) = 2\mathrm{Re}\{\alpha_k g_{\gamma_k}\} \tag{16}$$

where A_k, ω_k, and φ_k are the amplitude, frequency, and phase of the sinusoidal component h_k, respectively. Those harmonic structures are prominent in a signal when the speech is present but not when noise only.

In a practical implementation, the procedure for extracting MP feature is described as follows. Assuming the input signal is segmented into non-overlapping frames, each frame is decomposed by conjugate subspace MP. Thus, the complex MP coefficients of a given frame are obtained. Instead of requiring a full reconstruction of a signal, the goal of MP is to extract MP coefficients. These coefficients capture the most characters of a signal so that the VAD detector based on them can detect whether the speech is present or not. Naturally, the selection of iteration number K depends on the number of sinusoidal components in a speech signal.

4 Experiments and results

4.1 Noise statistic update

To implement the VAD scheme, the variance of the noise MP coefficients requires to be estimated, which are assumed to be known in Equation (14). We assume that the signal consists of noise only during a short initialization period, and the initial noise characteristics are learned. The background noise is usually non-stationary, and hence the estimation requires to be adaptively updated or tracked. The update is performed frame by frame by using the minimum mean square error (MMSE) estimation.

Since the signal is frame-processed, we use the superscript (m) to refer to the mth frame so that $\lambda_{n,k}^{(m)}$ and $\alpha_k^{(m)}$ denote $\lambda_{n,k}$ and α_k, respectively. Given the noisy MP coefficients $\alpha_k^{(m)}$ at the mth frame, the optimal estimate of the variance of the noise MP coefficients $\lambda_{n,k}^{(m)}$ under MMSE is given by

$$\begin{aligned} \hat{\lambda}_{n,k}^{(m)} &= E(\lambda_{n,k}^{(m)} | \alpha_k^{(m)}) \\ &= E(\lambda_{n,k}^{(m)} | H_0) P(H_0 | \alpha_k^{(m)}) + E(\lambda_{n,k}^{(m)} | H_1) P(H_1 | \alpha_k^{(m)}) \end{aligned} \tag{17}$$

where

$$E(\lambda_{n,k}^{(m)} | H_0) = |\alpha_k^{(m)}|^2 \tag{18}$$

$$E(\lambda_{n,k}^{(m)} | H_1) = \hat{\lambda}_{n,k}^{(m-1)} \tag{19}$$

and $\hat{\lambda}_{n,k}^{(m-1)}$ is the estimate in the previous frame. Based on the total probability theorem and Bayes rule, the posterior probabilities of H_0 and H_1 given α_k in

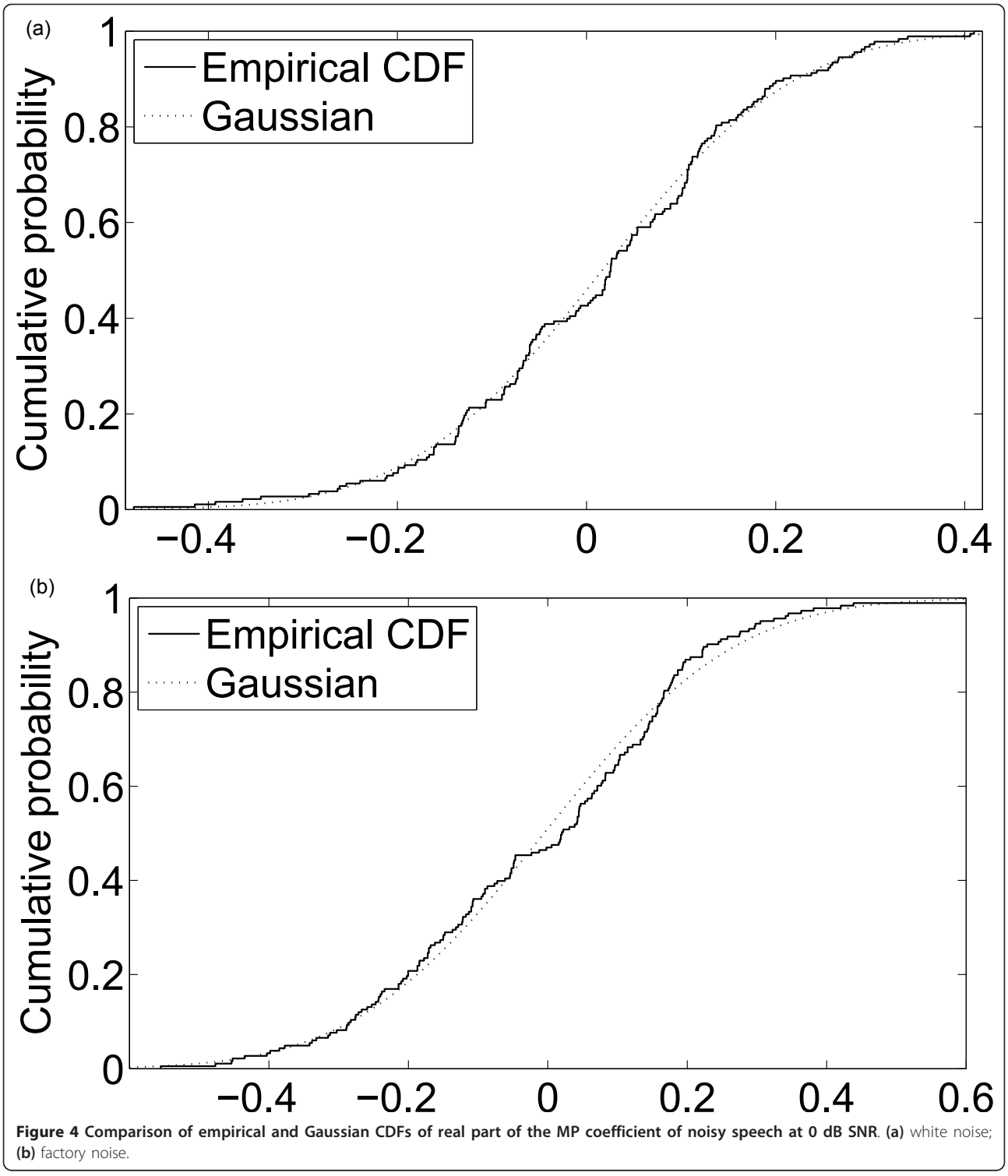

Figure 4 Comparison of empirical and Gaussian CDFs of real part of the MP coefficient of noisy speech at 0 dB SNR. (a) white noise; **(b)** factory noise.

Equation 17 are derived as follows

$$P(H_0|\alpha_k^{(m)}) = \frac{p(\alpha_k^{(m)}|H_0)P(H_0)}{p(\alpha_k^{(m)}|H_0)P(H_0) + p(\alpha_k^{(m)}|H_1)P(H_1)} = \frac{1}{1+\varepsilon\Lambda_k^{(m)}} \quad (20)$$

$$P(H_1|\alpha_k^{(m)}) = \frac{\varepsilon\Lambda_k^{(m)}}{1+\varepsilon\Lambda_k^{(m)}} \quad (21)$$

where $\varepsilon = P(H_1) = P(H_0)$ and $\Lambda_k^{(m)} = p(\alpha_k^{(m)}|H_1)/p(\alpha_k^{(m)}|H_0)$. Since the decision is made by observing all the K MP coefficients, we replace the LRT at the kth MP coefficient $\Lambda_k^{(m)}$ with their geometric mean $\Lambda_g^{(m)}$ in Equation (14).

Then the update formula of the variances of noise MP coefficients is given by

$$\hat{\lambda}_{n,k}^{(m)} = \frac{1}{1+\varepsilon\Lambda_g^{(m)}}|\alpha_k^{(m)}|^2 + \frac{\varepsilon\Lambda_g^{(m)}}{1+\varepsilon\Lambda_g^{(m)}}\hat{\lambda}_{n,k}^{(m-1)}. \quad (22)$$

4.2 Experimental results

In this section, the experimental results of our method are presented. To implement the proposed method, the dictionary **D** is the fundamental ingredient for decomposing a signal. The atoms of the dictionary are generated according to Equation (2), and the number of atoms is set to be $2N$, where $N = 256$. Thus, the complex exponential dictionary **D** is a $N \times 2N$ complex matrix, and is used in the following experiments. To demonstrate the effectiveness of the proposed VAD, a test signal (Figure 5b) is created by adding white noise to a clean speech (Figure 5a) at 0 dB SNR, and is divided into non-overlapping frames with the frame length 256. The atomic decomposition based on the conjugate subspace MP is operated on the test signal. The likelihood ratios and the results of VAD calculated with Equation (14) are shown in Figure 5c,d, respectively. As can be seen, even at such a low SNR, the results also correctly indicate the speech presence and thus verify the effectiveness of MP coefficients in VAD.

The selection of the iteration number K in the MP has an important effect on the performance of the proposed method and the computational cost. As shown in Figure 6, the performances of the VAD in various K are measured in terms of the the the receiver operating characteristic (ROC) curves, which show the trade-off between the false alarm probability (P_f) and speech detection probability (Pd). It is clearly shown that the increasing of K improves the performance of the VAD. A larger K, however, implies an increased computational cost. Figure 7 shows the decrease of the average errors, defined by $P_e = (P_f + 1 - P_d)=2$, against the increase of K in white, vehicle, and babble noise at 0 dB. The average errors in three noises remain unchange when the value

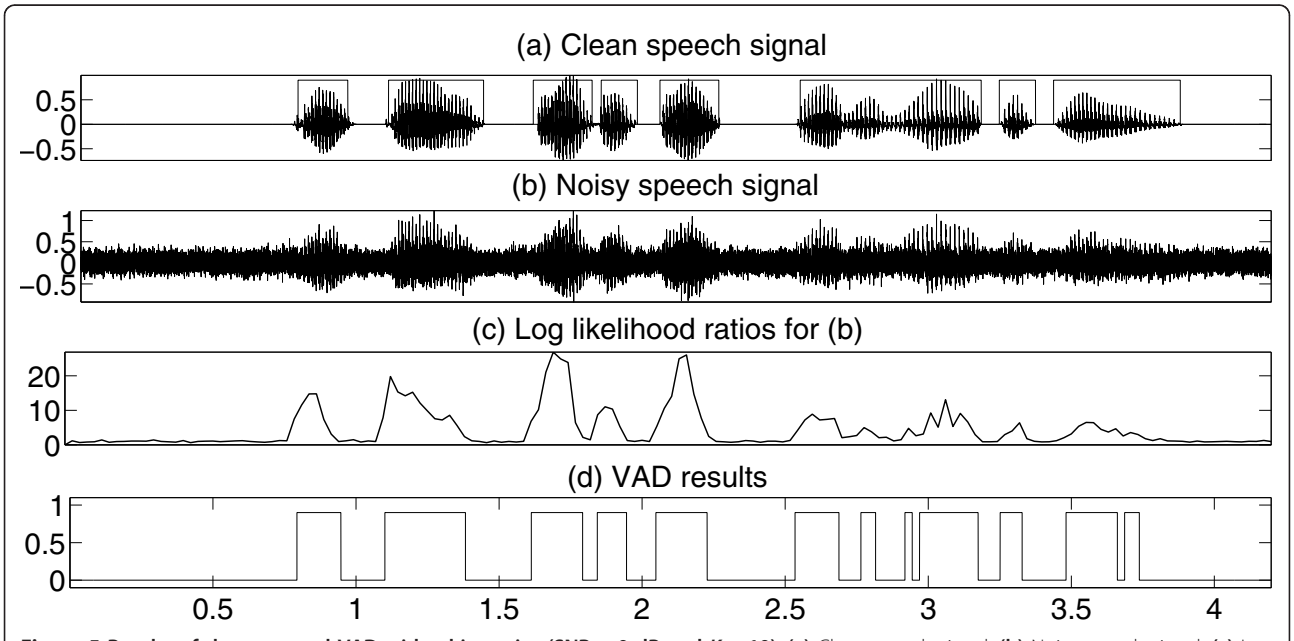

Figure 5 Results of the proposed VAD with white noise (SNR = 0 dB and K = 10). **(a)** Clean speech signal. **(b)** Noisy speech signal. **(c)** Log likelihood ratio for **(b)**. **(d)** VAD results.

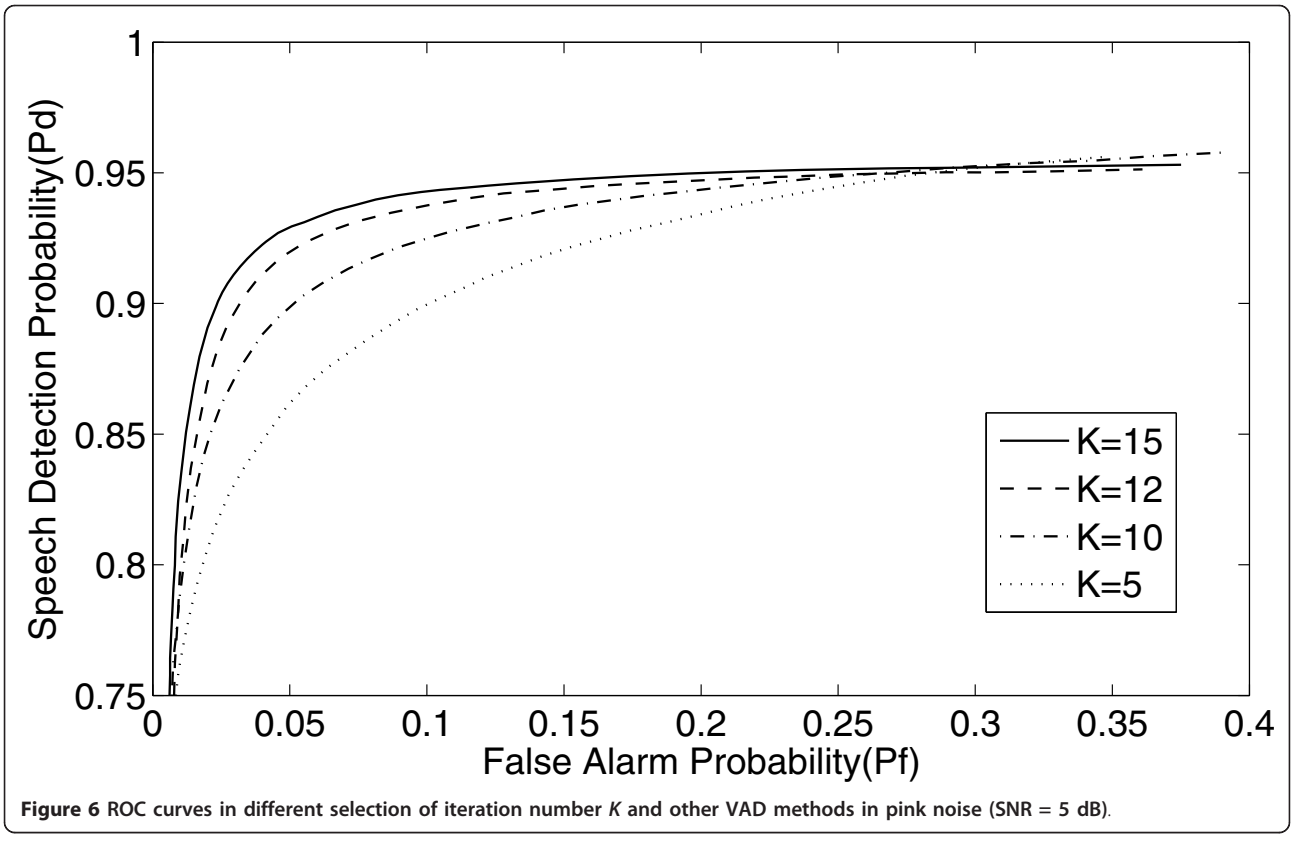

Figure 6 ROC curves in different selection of iteration number *K* and other VAD methods in pink noise (SNR = 5 dB).

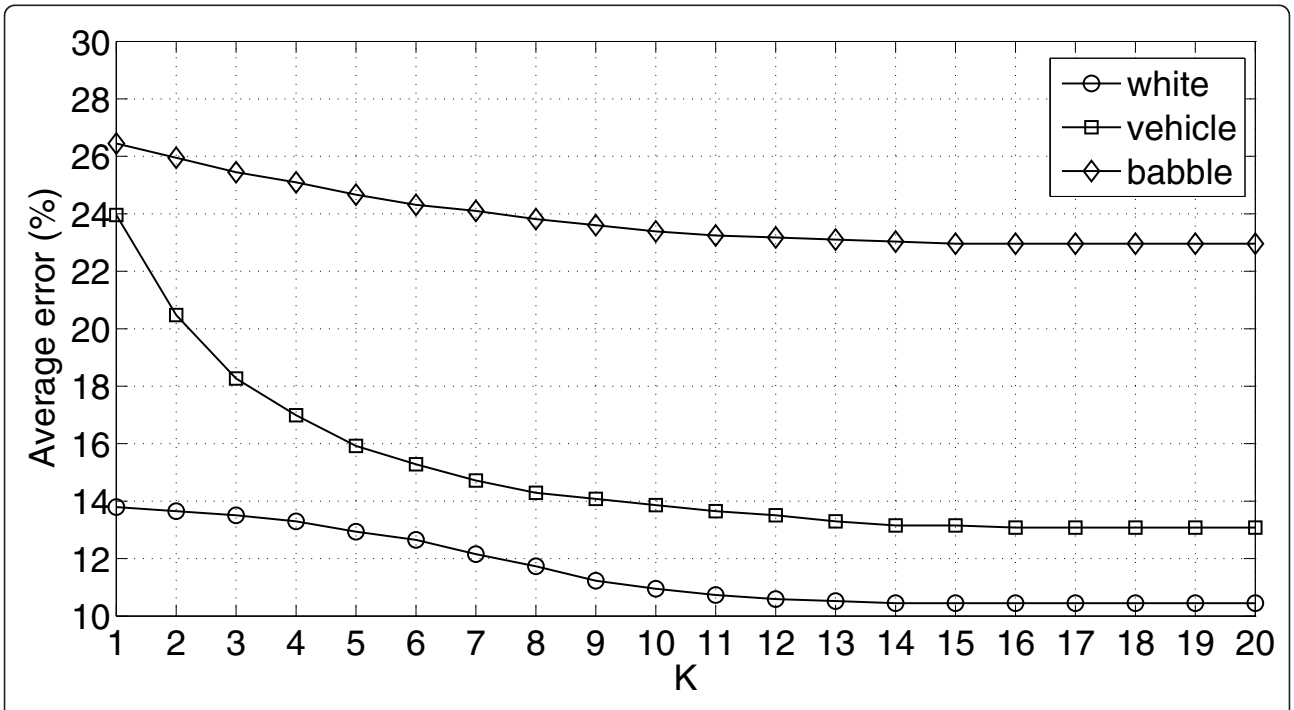

Figure 7 Average error for speech detection when increasing the iteration number *K* in the atomic decomposition in white, vehicle, and babble noise (SNR = 0 dB).

of K is larger than 15. Therefore, a reasonable value of K is equal to 15 so as to yield a good trade-off between the computational cost and the performance.

Based on the ROC curves, we evaluated the performances of the proposed LRT VAD based on the MP coefficients (LRT-MP) by comparing with the popular LRT VADs based on DFT coefficients, including Gaussian (LRT-Gaussian) [7], Laplacian (LRT-Laplacian) [8], and Gamma (LRT-Gamma) [10]. The test speech material used for the comparison is a clean speech of 135 s connected from 30 utterances selected from TIMIT database. The reference decisions are made on the clean speech by labeling manually at every 10 ms frame. To simulate the noise environments, the noise signal from NOI-SEX'92 database is added to the test speech at 5 dB SNR. For fair comparison, we do not consider any hang over during the detection, as these can be added in a heuristic way after the design of the decision rule. Figures 8, 9, and 10 shows the ROC curves of these VADs in the white, vehicle, and babble noise environments at 5 dB. It was observed that the proposed approach outperforms other VADs in three noise conditions. These results indicate that the MP coefficients can capture harmonic structure of speech that is insensitive to noise. In more detail, the performances of the proposed method compared with the LRT-Laplacian, which has a better performance than the LRT-Gaussian and LRT-Gamma, are summarized in Table 1, under white, vehicle, and babble noise conditions. The experimental results show that the VAD based on MP coefficients outperforms the ones based on the DFT in all of the testing conditions, and it can be concluded that the MP coefficients are more robust to background noise than the DFT.

5 Conclusion

In this article, we present a novel approach for VAD. The method is based on the complex atomic decomposition of a signal by using the conjugate subspace MP. With the decomposition, the complex MP coefficients are obtained, and modeled as the complex Gaussian distribution which is a suitable one according to the results of GOF test. Based on the statistical model, the decision rule for VAD is derived by incorporating the LRT on it. In a practical implementation, the decision is made frame by frame in a frame-processed signal.

The advantage of the proposed approach is that the MP coefficients are insensitive to the environmental noise, and hence the performance of VAD is robust in high noise environments. Note that, the advantage with MP coefficients is obtained at the cost of computational cost, which is proportional to the iteration number. An online detection can be implemented when the iteration number is smaller than 20. Furthermore, the

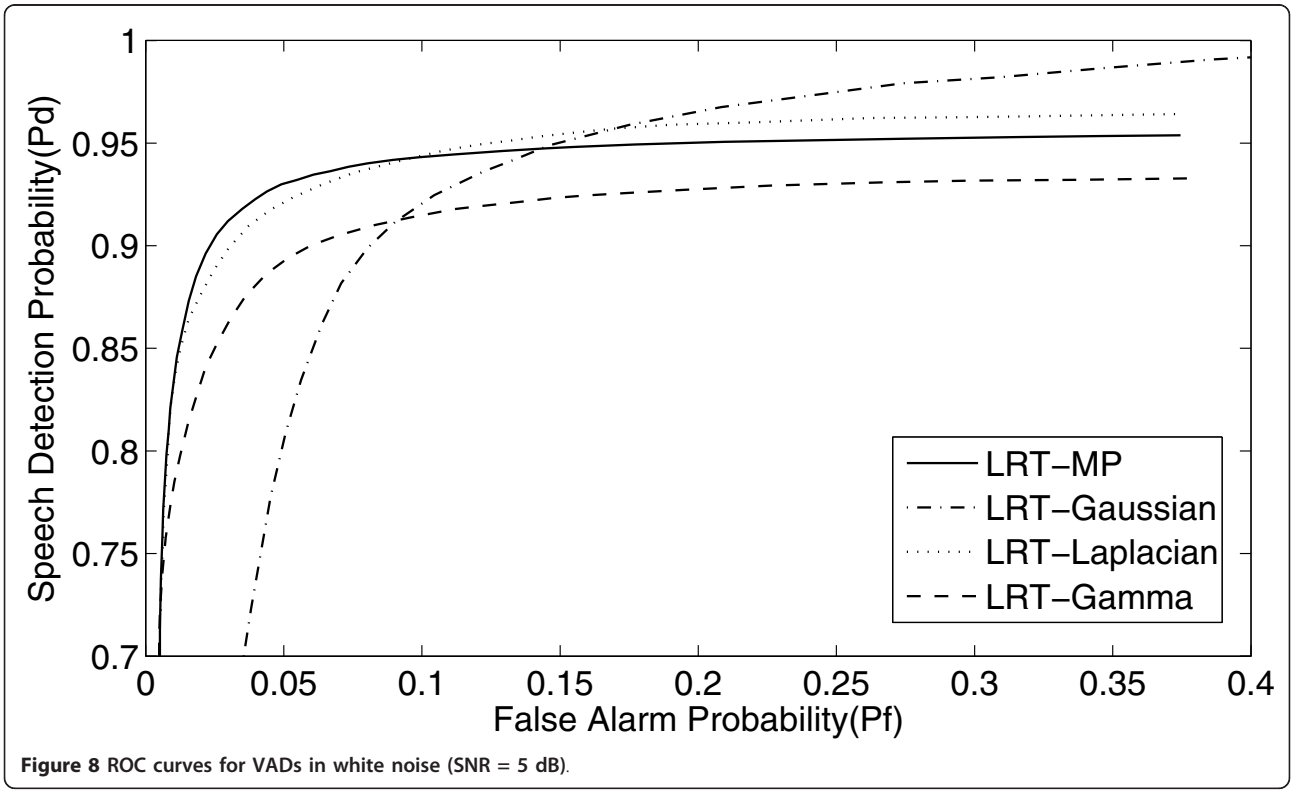

Figure 8 ROC curves for VADs in white noise (SNR = 5 dB).

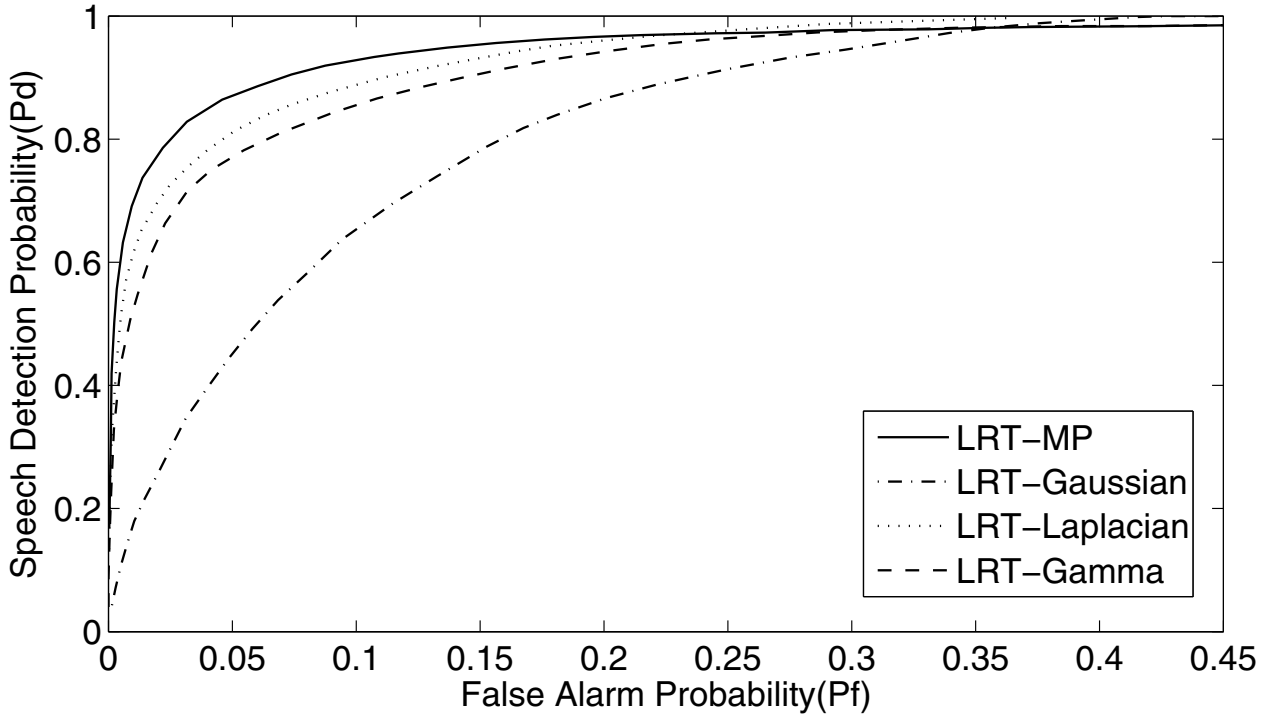

Figure 9 ROC curves for VADs in vehicle noise (SNR = 5 dB).

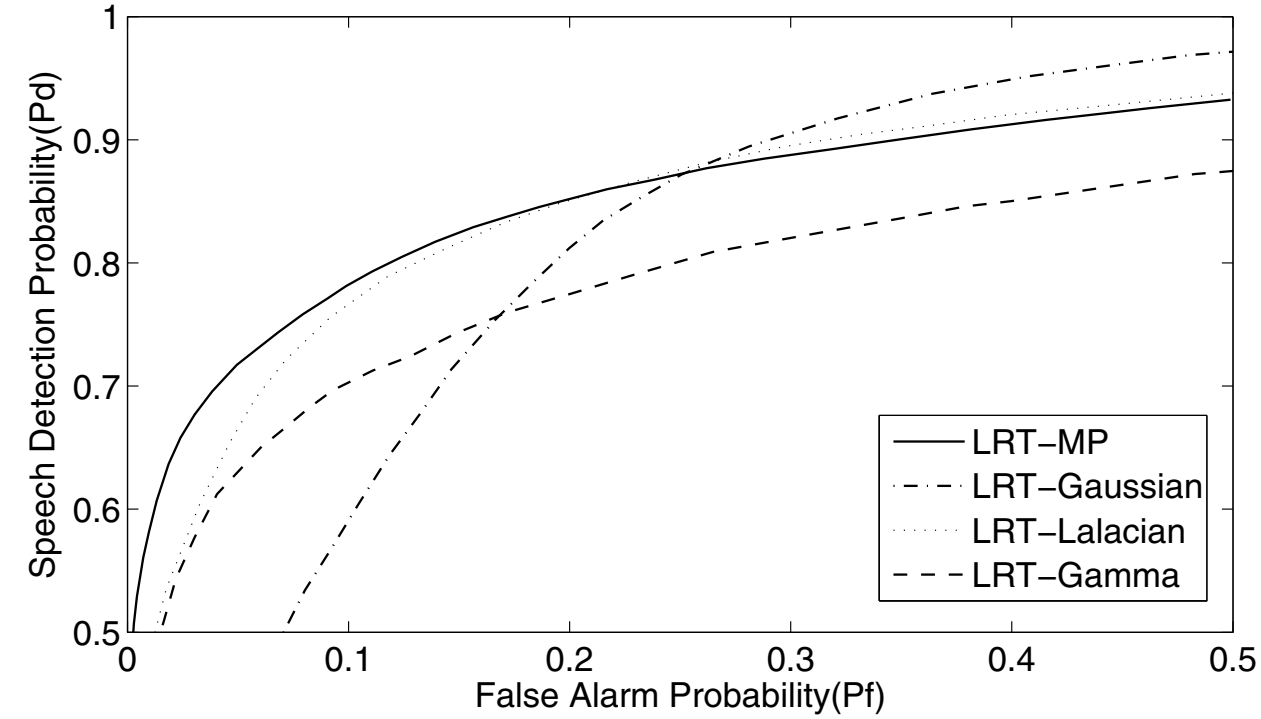

Figure 10 ROC curves for VADs in babble noise (SNR = 5 dB).

Table 1 Performance evaluation in different noise conditions

Environments		LRT-MP		LRT-Laplacian	
Noise	SNR (dB)	P_d (%)	P_f (%)	P_d (%)	P_f (%)
White	0	87.9	10.7	88.7	10.3
	5	94.3	9.9	94.2	9.7
	10	96.4	9.5	95.8	9.6
	20	97.2	9.4	96.8	9.2
Vehicle	0	85.3	10.9	80.3	11.4
	5	93.3	10.7	89.7	10.5
	10	95.4	9.1	92.5	10.2
	20	97.2	8.8	95.2	9.3
Babble	0	63.3	11.1	58.7	11.9
	5	79.3	11.1	78.9	11.7
	10	84.2	9.3	80.6	10.4
	20	87.4	9.1	83.7	9.6

experimental results show that the proposed approach outperforms the traditional VADs based on DFT coefficients in white, vehicle, and babble noise conditions.

Acknowledgements
This study was supported by the Natural Science Foundation of China (No. 61071181 and 91120303).

Author details
[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China [2]School of Mathematical Sciences, Harbin Normal University, Harbin, China

Competing interests
The authors declare that they have no competing interests.

Received: 29 June 2011 Accepted: 21 December 2011
Published: 21 December 2011

References

1. A Benyassine, E Shlomot, HY Su, D Massaloux, C Lamblin, JP Petit, ITU-T Recommendation G.729, Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. IEEE Commun Mag. **35**(9):64–73 (1997). doi:10.1109/35.620527
2. K Itoh, M Mizushima, Environmental noise reduction based on speech/non-speech identification for hearing aids. Proc Int Conf Acoust, Speech, and Signal Process. **1**, 419–422 (1997)
3. N Virag, Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans Speech Audio Process. **7**(2):126–137 (1999). doi:10.1109/89.748118
4. K Woo, T Yang, K Park, C Lee, Robust voice activity detection algorithm for estimating noise spectrum. Electron Lett. **36**(2):180–181 (2000). doi:10.1049/el:20000192
5. M Marzinzik, B Kollmeier, Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Trans Speech Audio Process. **10**(6):341–351 (2002). doi:10.1109/TSA.2002.803420
6. SM Kay, *Fundamentals of Statistical Signal Processing*. (Prentice-Hall, Englewood Cliffs, 1998)
7. J Sohn, NS Kim, W Sung, A statistical model-based voice activity detection. IEEE Signal Process Lett. **6**(1):1–3 (1999). doi:10.1109/97.736233
8. JH Chang, JW Shin, NS Kimm, Likelihood ratio test with complex Laplacian model for voice activity detection. *Proc Eurospeech*. (Geneva, Switzerland, 2003), pp. 1065–1068
9. JW Shin, JH Chang, NS Kim, Voice activity detection based on a family of parametric distributions. Pattern Recogn Lett. **28**(11):1295–1299 (2007). doi:10.1016/j.patrec.2006.11.015
10. JW Shin, JH Chang, HS Yun, NS Kim, Voice activity detection based on generalized gamma distribution. Proc IEEE Internat Conf on Acoustics, Speech, and Signal Processing **1**, 781–784 (2005). Corfu, Greece 17-19
11. J Ramirez, JC Segura, C Benitez, L Garcia, A Rubio, Statistical voice activity detection using a multiple observation likelihood ratio test. IEEE Signal Process Lett. **12**(10):689–692 (2005)
12. JM Gorriz, J Ramirez, EW Lang, CG Puntonet, Jointly Gaussian PDF-based likelihood ratio test for voice activity detection. IEEE Trans Speech Audio Process. **16**(8):1565–1578 (2008)
13. J Ramirez, JM Gorriz, JC Segura, CG Puntonet, AJ Rubio, Speech/non-speech discrimination based on contextual information integrated bispectrum LRT. IEEE Signal Process Let. **13**(8):497–500 (2006)
14. JM Gorriz, J Ramirez, CG Puntonet, JC Segura, Generalized LRT-based voice activity detector". IEEE Signal Process Lett. **13**(10):636–639 (2006)
15. JW Shin, HJ Kwon, NS Kim, Voice activity detection based on conditional MAP criterion. IEEE Signal Process Lett. **15**, 257–260 (2008)
16. Shiwen Deng, Jiqing Han, A modified MAP criterion based on hidden Markov model for voice activity detection. Proc Int Conf Acoust, Speech, Signal Process 5220–5223 (2011). Prague 22-27
17. SG Mallat, Z Zhang, Matching pursuit in a time-frequency dictionary. IEEE Trans Signal Process. **41**(12):3397–3415 (1993). doi:10.1109/78.258082
18. M Goodwin, Matching pursuit with damped sinusoids. Proc IEEE Internat Conf on Acoustics, Speech, and Signal Processing **3**, 2037–2040 (1997). Munich, Germany 21-24
19. M Goodwin, M Vetterli, Matching pursuit and atomic signal models based on recursive filter banks. IEEE Trans Signal Process. **47**(7):1890–1902 (1999). doi:10.1109/78.771038
20. MR McClure, L Carin, Matching pursuits with a wave-based dictionary. IEEE Trans Signal Process. **45**(12):2912–2927 (1997). doi:10.1109/78.650250
21. D Shiwen, H Jiqing, Voice activity detection based on complex exponential atomic decomposition and likelihood ratio test. *20th Int Conf Pattern Recognition, ICPR 2010*. (Istanbul, Turkey, 2010), pp. 89–92
22. RC Reininger, JD Gibson, Distributions of the two dimensional DCT coefficients for images. IEEE Trans Commun. **31**(6):835–839 (1983). doi:10.1109/TCOM.1983.1095893

doi:10.1186/1687-4722-2011-12
Cite this article as: Deng and Han: **Voice activity detection based on conjugate subspace matching pursuit and likelihood ratio test.** *EURASIP Journal on Audio, Speech, and Music Processing* 2011 **2011**:12.