



Voice Activity Detection Using the Phase Vector in Microphone Array

Gibak Kim and Nam Ik Cho

School of Electrical Engineering, Seoul National University, Korea

kgeb@ispl.snu.ac.kr, nicho@snu.ac.kr

Abstract

If desired speech source is located at different position from interference, it is possible to exploit spatial selectivity for reliable speech detection. In this paper, we propose a voice activity detector (VAD) for the microphone array system, using spatial information obtained by the eigendecomposition of multi-channel correlation matrix. We use the phase vector as a measure for VAD, which is derived from the principal eigenvector. Voice activity is detected by the log likelihood ratio test under the assumption that phase vectors of speech absent and present signals have complex Gaussian distributions. The proposed algorithm is tested with the real data recorded by 8 microphones and the result shows that it performs better than GSC-based method.

Index Terms: voice activity detection, microphone array

1. Introduction

Various VAD algorithms have been developed for the applications such as speech recognition, speech enhancement, and speech coding. For the efficient bandwidth reduction in speech coding, a reduced rate can be used during speech absent periods. VAD is crucial for speech enhancement techniques which estimate noise statistics and adjust filter coefficients during speech absent intervals to avoid speech distortion. In some speech recognition systems, speech detection is accomplished inside the recognition process with non-speech models. However, the VAD still plays an important role to guarantee the speech recognition performance in noisy environment.

In order to distinguish speech from noise, numerous feature parameters have been employed such as time domain or spectral domain energy, zero-crossing rate, cepstral coefficients, spectral entropy, linear prediction coefficients, pitch, formant frequencies and the like [1]. However, most of these methods are unreliable in the presence of non-stationary or broadband speech-like noise. Recently, several researchers have introduced multi-channel algorithms for better VAD performance exploiting the spatial selectivity [2–4]. Specifically, Le Bouquin and Faucon proposed a technique based on the coherence function [2]. They assumed that the spatial correlation between the disturbing noises is weak for all frequencies of interest while the speech signals are highly correlated. They compared the averaged Magnitude Squared Coherence (MSC) with a threshold to decide whether the speech is present in the current segment or not. In [3], the author introduced the speech presence probability estimation per spectral band, assuming that every complex spectral coefficient in each snapshot forms a circular, zero mean, white complex Gaussian vector. Unlike these techniques which deal with spatially uncorrelated noise, Hoffman *et al.* used the direction information of desired speech signal and attempted to detect desired speech in the presence of coherent interference [4]. They estimated the short-term Signal-to-Interference Ratio (SIR) by dividing the GSC output power

by the estimated noise reference power. However, the desired speech signal may leak into the noise reference due to room reverberation, microphone mismatch, microphone placement uncertainty, or DOA (Direction Of Arrival) error. Though speech leakage can be reduced by some robustness techniques, the speech leakage into the noise reference tends to be higher as input SIR increases and makes it difficult to estimate the input SIR accurately.

In this paper, we use the phase vector as a measure for VAD; first, the eigenanalysis is applied to the multi-channel correlation matrix in the frequency domain and then the phase vector is derived from the principal eigenvector to represent the phase of the signal received at each microphone with respect to the first microphone. We assume that the phase vectors of speech absent and present signals have complex Gaussian distributions. The Gaussian parameters are recursively updated and the log Likelihood Ratio Test (LRT) is carried out to detect the speech presence. The proposed algorithm exploits the spatial information instead of speech distinct feature parameters and thus has the advantage that it does not need “threshold” for the decision of speech presence unlike most VAD algorithms [1–4]. In the experiments, the proposed method is compared with the GSC-based multi-channel VAD [4] and it is shown that the method yields better performance in the presence of interference.

This paper is organized as follows. In the next section, we derive the phase vector from the eigendecomposition of multi-channel input matrix and investigate its distribution. Section 3 derives a log LRT for the speech detection under the Gaussian assumption of phase vectors and presents the adaptation process for the Gaussian parameters. Section 4 shows the experimental results and section 5 concludes the paper.

2. Phase vector and its distribution

When additive noise degrades speech in an M -microphone system, we assume two hypotheses H_0 , H_1 representing speech absence and presence as

$$\begin{aligned} H_0 : \mathbf{y}(t, k) &= \mathbf{n}(t, k) \\ H_1 : \mathbf{y}(t, k) &= \mathbf{x}(t, k) + \mathbf{n}(t, k) \end{aligned} \quad (1)$$

where $\mathbf{y}(t, k)$, $\mathbf{n}(t, k)$, and $\mathbf{x}(t, k)$ are M -dimensional vectors of the k th DFT coefficient for observed signal, noise, and speech, respectively, at time instance t . These vectors are given by

$$\mathbf{y}(t, k) = [y_1(t, k) \ y_2(t, k) \ \cdots \ y_M(t, k)]^T \quad (2)$$

$$\mathbf{n}(t, k) = [n_1(t, k) \ n_2(t, k) \ \cdots \ n_M(t, k)]^T \quad (3)$$

$$\mathbf{x}(t, k) = [x_1(t, k) \ x_2(t, k) \ \cdots \ x_M(t, k)]^T \quad (4)$$

in which $y_i(t, k)$, $n_i(t, k)$, and $x_i(t, k)$ are DFT coefficients at the i th microphone. By applying the eigendecomposition to the multi-channel correlation matrix ($\mathbf{R}_y(t, k) =$

10.21437/Interspeech.2007-737

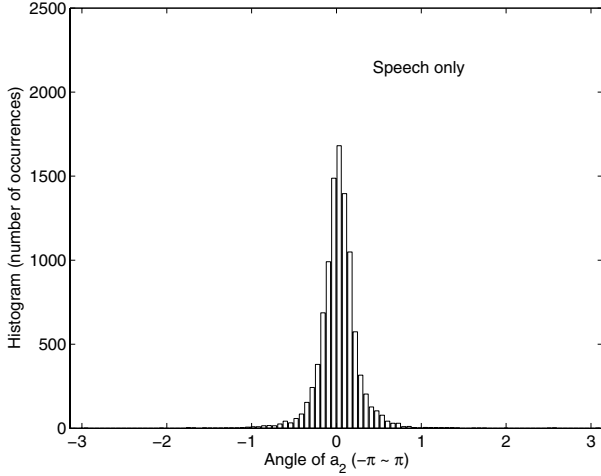


Figure 1: Histogram of angle (a_2 : the 2nd element of the phase vector) for speech signals (without interference).

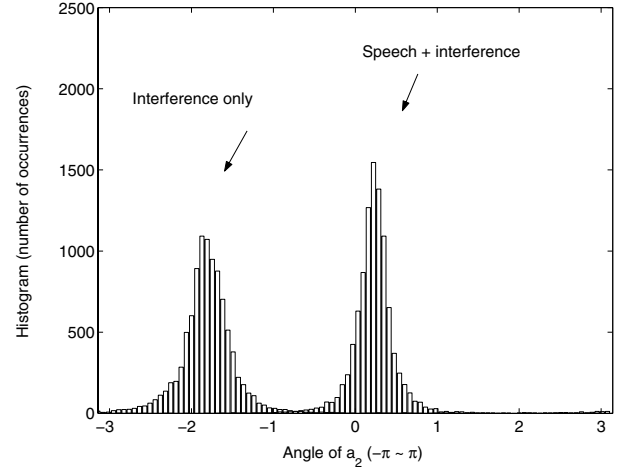


Figure 3: Histograms of angle (a_2 : the 2nd element of the phase vector) for speech absent and present signals (direction of interference= 0° , $SIR=15dB$).

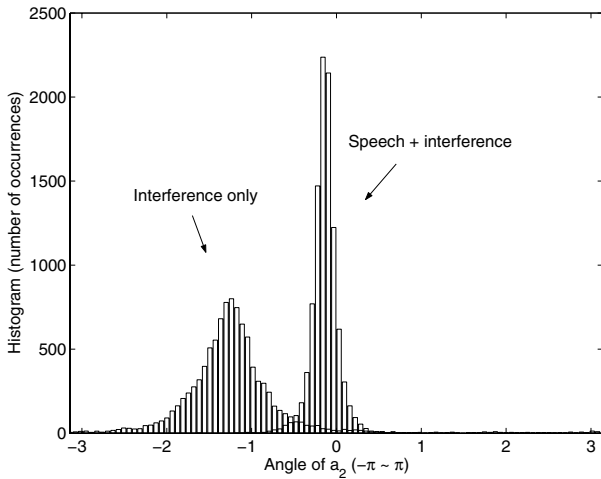


Figure 2: Histograms of angle (a_2 : the 2nd element of the phase vector) for speech absent and present signals (direction of interference= 45° , $SIR=15dB$).

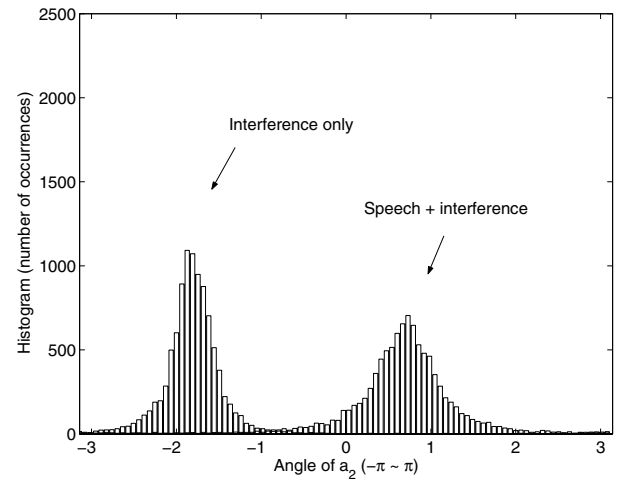


Figure 4: Histograms of angle (a_2 : the 2nd element of the phase vector) for speech absent and present signals (direction of interference= 0° , $SIR=5dB$).

$E[\mathbf{y}(t, k)\mathbf{y}(t, k)^H]$), the spatial subspace decomposition can be obtained. For instance, if a signal from the far-field single signal source is arrived at the microphones, the principal eigenvector (corresponding to the largest eigenvalue) of the correlation matrix corresponds to the array manifold vector of the far-field signal. The eigendecomposition of the correlation matrix is given by

$$\begin{aligned} \mathbf{R}_y(t, k) &= \mathbf{Q}(t, k)\mathbf{D}(t, k)\mathbf{Q}(t, k)^H \\ &= \sum_{i=1}^M d_i(t, k)\mathbf{q}_i(t, k)\mathbf{q}_i(t, k)^H \end{aligned} \quad (5)$$

in which $\mathbf{Q}(t, k)$ is the unitary eigenvector matrix consisting of eigenvectors $\mathbf{q}_i(t, k)$ and $\mathbf{D}(t, k)$ is a diagonal matrix as $\mathbf{D}(t, k) = \text{diag}\{d_1(t, k) \ d_2(t, k) \ \cdots \ d_M(t, k)\}$. By incorporating with the principal eigenvector $\mathbf{q}_1(t, k) = [q_1(t, k) \ q_2(t, k) \ \cdots \ q_M(t, k)]^T$ and normalizing the eigen-

vector by its first element as

$$\begin{aligned} \mathbf{q}'_1(t, k) &= \mathbf{q}_1(t, k)/q_1(t, k) \\ &= [1 \ q_2(t, k)/q_1(t, k) \ \cdots \ q_M(t, k)/q_1(t, k)]^T \\ &\triangleq [1 \ q'_1(t, k) \ q'_2(t, k) \ \cdots \ q'_{M-1}(t, k)]^T, \end{aligned} \quad (6)$$

the phase vector is described as

$$\begin{aligned} \mathbf{a}(t, k) &= [a_1 \ a_2 \ \cdots \ a_{M-1}]^T \\ &\triangleq \left[\frac{q'_1(t, k)}{|q'_1(t, k)|} \ \frac{q'_2(t, k)}{|q'_2(t, k)|} \ \cdots \ \frac{q'_{M-1}(t, k)}{|q'_{M-1}(t, k)|} \right]^T. \end{aligned} \quad (7)$$

The i th element of this vector represents the phase of the signal received at the $(i+1)$ th microphone with respect to the first microphone. We assume that the density of phase vector is complex Gaussian function both in the speech absent and

present periods. Fig. 1 shows the histogram in terms of the angle of a_2 (the 2nd element of the phase vector), when the speech signal is received at the front of the microphone array (90°) with reverberation time of 150 ms (no other interfering signal exists). It shows that the mean angle of a_2 is near 0 for speech-only signal. But if there exists an interfering signal, the mean of the noisy speech signal deviates from that of the speech-only signal and the variance also changes according to the direction of interference and the SIR. Fig. 2 - Fig. 4 illustrate the variations of the distribution of noisy speech signal according to the direction of interference and the SIR.

3. Log LRT and Gaussian parameters update

With the Gaussian pdf assumption of phase vectors, the probability density functions conditioned on H_0 and H_1 are given by

$$p(\mathbf{a}(t, k)|H_0) = \frac{1}{(2\pi)^M |\boldsymbol{\Sigma}_n(t, k)|} \cdot \exp \left\{ - \left(\mathbf{a}(t, k) - \mathbf{m}_n(t, k) \right)^H \cdot \boldsymbol{\Sigma}_n^{-1}(t, k) (\mathbf{a}(t, k) - \mathbf{m}_n(t, k)) / 2 \right\} \quad (8)$$

$$p(\mathbf{a}(t, k)|H_1) = \frac{1}{(2\pi)^M |\boldsymbol{\Sigma}_{x+n}(t, k)|} \cdot \exp \left\{ - \left(\mathbf{a}(t, k) - \mathbf{m}_{x+n}(t, k) \right)^H \cdot \boldsymbol{\Sigma}_{x+n}^{-1}(t, k) (\mathbf{a}(t, k) - \mathbf{m}_{x+n}(t, k)) / 2 \right\} \quad (9)$$

where \mathbf{m}_n , \mathbf{m}_{x+n} , $\boldsymbol{\Sigma}_n$, and $\boldsymbol{\Sigma}_{x+n}$ denote the mean vectors and covariance matrices of the phase vector for the speech absent (n) and present signals ($x+n$), respectively. We compute the log likelihood ratio to detect the voice activity as

$$\Lambda(t) = \sum_{k=1}^K \log \frac{p(\mathbf{a}(t, k)|H_1)}{p(\mathbf{a}(t, k)|H_0)} \quad (10)$$

where K is the number of DFT points. If the log likelihood ratio is greater than 0, the speech presence is detected. There is no bias in the log likelihood and thus no need to update the threshold.

The mean vector and covariance matrix of noise are derived from the initial $N_{init.ns}$ frames, which are assumed to be speech absent, i.e., noise-only. The covariance matrix of the speech present signal is set to be that of noise, and the mean vector for the speech present signal is initially chosen as the array manifold vector of desired speech signal which is assumed to be known. As previously mentioned, the phase vector of the noisy speech signal deviates from the array manifold vector according to the direction of interference and the SIR (Fig. 2 - Fig. 4). In some cases, we can reasonably assume the invariance of the interference direction, but the SIR is usually highly time varying and thus the mean vectors and covariance matrices need to be adaptively updated. To update the parameters, the segmental MAP (Maximum A posteriori Probability) estimation is employed by using the *a posteriori* probability for weighting the adaptation [5]. The *a posteriori* probabilities are estimated by the current Gaussian parameters and the Bayes'

theorem as

$$P(\lambda_{x+n}(t, k)|\mathbf{a}(t, k)) = \frac{p(\mathbf{a}(t, k)|\lambda_{x+n}(t, k))}{p(\mathbf{a}(t, k)|\lambda_{x+n}(t, k)) + p(\mathbf{a}(t, k)|\lambda_n(t, k))} \quad (11)$$

$$P(\lambda_n(t, k)|\mathbf{a}(t, k)) = \frac{p(\mathbf{a}(t, k)|\lambda_n(t, k))}{p(\mathbf{a}(t, k)|\lambda_{x+n}(t, k)) + p(\mathbf{a}(t, k)|\lambda_n(t, k))} \quad (12)$$

with $\lambda_{x+n} = \{\mathbf{m}_{x+n}, \boldsymbol{\Sigma}_{x+n}\}$, $\lambda_n = \{\mathbf{m}_n, \boldsymbol{\Sigma}_n\}$, and equal *a priori* probabilities i.e., $P(\lambda_{x+n}) = P(\lambda_n) = 0.5$. The EM (Expectation Maximization) MAP adaptation of mean vector is performed by weighting the contribution of the input phase vector with the *a posteriori* probability:

$$\mathbf{m}_{x+n}(t, k) = \frac{1}{C_{x+n}(t)} [\beta C_{x+n}(t-1, k) \mathbf{m}_{x+n}(t-1, k) + P(\lambda_{x+n}(t, k)|\mathbf{a}(t, k)) \mathbf{a}(t, k)] \quad (13)$$

$$\mathbf{m}_n(t, k) = \frac{1}{C_n(t)} [\beta C_n(t-1, k) \mathbf{m}_n(t-1, k) + P(\lambda_n(t, k)|\mathbf{a}(t, k)) \mathbf{a}(t, k)] \quad (14)$$

where $C_{x+n}(t, k)$ and $C_n(t, k)$ are smoothed with smoothing factor β as

$$C_{x+n}(t, k) = \beta C_{x+n}(t-1, k) + P(\lambda_{x+n}(t, k)|\mathbf{a}(t, k)) \quad (15)$$

$$C_n(t, k) = \beta C_n(t-1, k) + P(\lambda_n(t, k)|\mathbf{a}(t, k)) \quad (16)$$

with $C_{x+n}(1, k) = C_n(1, k) = N_{init.ns}/2$. The covariance matrices are updated in a similar way. Fig. 5 describes The adaptation curves in the presence of DOA error. The Euclidean distances between the input phase vector and the speech mean vector are plotted, in which the speech signal begins at 64th frame. As DOA error increases, the mean vector approaches the input phase vector more slowly. The adaptation curves are also shown as a function of $N_{init.ns}$ in Fig. 6. The mean vector is slowly modified by the adaptation process in the case of large $N_{init.ns}$.

4. Experimental results

We evaluate the proposed method using a uniform linear array of 8 omnidirectional microphones with the distance between adjacent microphones of 5cm. Speech signal is received at the front of the microphone array (90°) while the music is played from the direction of the left side of the microphone array (0°). The sampling rate is 16kHz and 512-point FFT is applied to the windowed data with 256 sample overlap. The proposed method is compared with GSC-based multi-channel VAD proposed in [4]. $N_{init.ns}$ and β are set to be 30 (corresponding to 480ms) and 0.99, respectively. The principal eigenvector is computed by PAST (Projection Approximation Subspace Tracking) method with low computational complexity [6]. Moreover, since the proposed method need to calculate a single eigenvector corresponding to the largest eigenvalue instead of all the eigenvectors, only $O(M)$ operations are required at every update.

Fig. 7 describes performance evaluation of the proposed method compared to the GSC-based method. In Fig. 7 (c), the

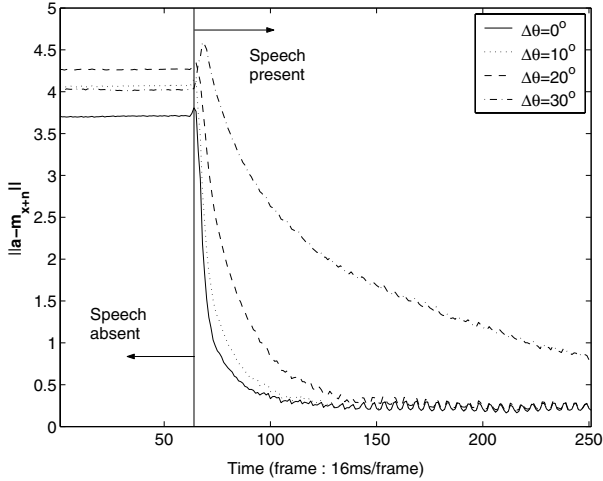


Figure 5: Adaptation curve of mean vector according to DOA errors.

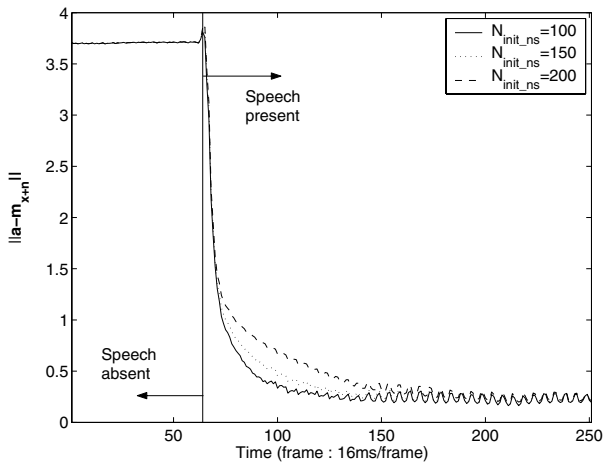


Figure 6: Adaptation curve of mean vector according to N_{init_ns} .

short term SIR estimate and threshold are illustrated with solid and dotted line, respectively. A basic threshold T_1 is nominally initialized to 0dB and recursively updated during non-speech activity. Speech is detected if $SIR \geq T_1 + \Delta T$ (typically, $\Delta T=5\text{dB}$) (Fig. 7 (d)). Fig. 7 (e) and (f) show the results of the proposed method with the log likelihood ratio and detected voice segments, which demonstrate the better performance of the proposed method compared to the GSC-based method. Also note that the proposed method does not need to update the threshold which is required for the GSC-based method.

5. Conclusions

We have proposed a VAD for microphone array, which exploits the spatial information of speech and noise source by employing the phase vector. Under the assumption of complex Gaussian distributions of speech and noise phase vectors, speech period is detected by the log LRT. The proposed algorithm exploits the spatial information instead of speech distinct feature parameters and has the advantage that it does not need “threshold” for

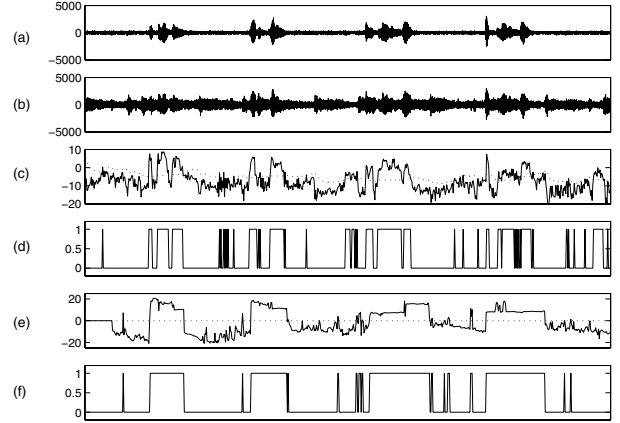


Figure 7: Signal waveforms at the first microphone and VAD results. (a) signal waveform in the absence of interference. (b) signal waveform in the presence of music interference. (c) the estimated SIR based on GSC (solid line) and threshold (dotted line). (d) the detected voice activity from the GSC-based SIR estimate. (e) the log likelihood ratio of proposed method (solid line) and 0 as a threshold (dotted line). (f) the detected voice activity from the log likelihood ratio of the proposed method.

the decision of speech presence unlike most VAD algorithms. Experimental results show that the proposed method performs better than GSC-based method in the presence of interference.

6. Acknowledgements

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

7. References

- [1] B.-F. Wu, “Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, September 2005.
- [2] R. Le Bouquin and G. Faucon, “Study of a voice activity detector and its influence on a noise reduction system”, *Speech communication* vol. 16, pp. 245-254, 1995.
- [3] I. Potamitis, “Estimation of speech presence probability in the field of microphone array,” *IEEE Signal Processing Letters*, vol. 11, no. 12, pp. 956-959, December 2004.
- [4] M. Hoffman, Z. Li, and D. Khataniar, “GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech,” *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 2, pp. 175-179, March 2001.
- [5] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291-298, April 1994.
- [6] B. Yang, “Projection approximation subspace tracking,” *IEEE Trans. on Signal Processing*, vol. 43, no. 1, pp. 95-107, January 1995.