



# Multi-channel voice activity detection based on conic constraints

Gibak Kim

School of Electronic Engineering, College of Information & Communication,  
Daegu University, Korea

imkgb27@gmail.com

## Abstract

Unlike single microphone techniques for voice activity detection (VAD), multi-microphone signal processing usually exploits the spatial information of signals received at multiple microphones. In this paper, we propose a VAD algorithm based on conic constraints to achieve robustness against the direction of arrival (DOA) estimation error. The proposed algorithm uses the phase vector as feature and detects the presence of the target speech by comparing the angles between the phase vector of the multi-microphone input signal and two mean phase vectors for target speech+interference period and interference-only period. The proposed algorithm was tested with simulation data generated by real-measured impulse response for seven uniformly distributed microphones. The simulation results showed that the proposed algorithm presents a reliable VAD metric in the presence of competing speech. The results also supported the robustness of the proposed algorithm against the DOA estimation error.

**Index Terms:** voice activity detection, phase vector, conic constraint

## 1. Introduction

VAD is an algorithm used in speech processing, wherein the presence or absence of human speech is detected from the audio samples. Though many single microphone VAD algorithms have been developed for several decades, most algorithms fail to provide reliable performance in the presence of non-stationary interfering signal which has broadband speech-like (or competing speech) spectral characteristic.

Recently, for better VAD performance, multi-microphone signal processing has been developed exploiting the spatial information of signals received at multiple microphones [1–7]. In the microphone array system, if the direction of the target speech is assumed to be known, steered response power (SRP) serves as a better VAD feature compared to VAD features extracted from single microphone signal. Several beamforming techniques have been applied to provide the SRP for the multi-channel VAD. Hoshuyama *et al.* estimated the signal-to-interference ratio (SIR) using the output powers of the fixed beamformer and the adaptive blocking matrix to control the adaptation of a robust adaptive microphone array [1]. Hoffman *et al.* proposed a GSC-based spatial VAD which is jointly included with a robust adaptive microphone array and a speech coder [2]. Other than the SRP, there have been some efforts to exploit the coherence measure for VAD feature. Specifically, Le Bouquin and Faucon introduced a technique based on the magnitude-squared coherence (MSC) [3]. Guerin developed an MSC based VAD with adaptive threshold which ensures a quasi-constant behavior in different environmental conditions and different relative microphone positions [4]. Cross power

spectrum phase (CPSP) was also used for two-microphone VAD [5, 6]. These multi-channel VAD algorithms exploit spatial information of signals using more than one microphone and perform well in high level noisy environments. However, most multi-channel VAD algorithms rely on the estimation of DOA and are sensitive to the DOA estimation error.

In this paper, we propose a new multi-channel VAD algorithm which is robust against the DOA estimation error. The proposed algorithm uses the phase vector as VAD feature to detect the target speech activity in the presence of interference [7]. Phase vectors are computed to obtain the spatial information of incoming signals and two mean phase vectors for target speech+interference period and interference-only period are estimated and the presence of the target speech is detected by comparing the angles between the phase vector of the input signal and the two mean phase vectors. To achieve the robustness against the DOA estimation error, conic constraints are imposed on the decision and update process.

This paper is organized as follows. Section 2 describes the multi-microphone signal model and the definition of the phase vector. In Section 3, conic constraints are reviewed. We propose a multi-channel VAD based on the phase vector and conic constraints in Section 4. Simulation results are provided in Section 5.

## 2. Multi-microphone signal model and phase vector

When a target speech in an  $M$ -microphone system is degraded by additive noise (interference), two hypotheses  $H_0$ ,  $H_1$  representing the absence and the presence of the target speech can be assumed as

$$\begin{aligned} H_0 : \mathbf{Y}(t, k) &= \mathbf{D}(t, k) \\ H_1 : \mathbf{Y}(t, k) &= \mathbf{X}(t, k) + \mathbf{D}(t, k) \end{aligned} \quad (1)$$

where  $\mathbf{Y}(t, k)$ ,  $\mathbf{D}(t, k)$ , and  $\mathbf{X}(t, k)$  are  $M$ -dimensional vectors of the  $k$ th discrete Fourier transform (DFT) coefficients of the observed noisy signal, interference, and target speech, respectively, at time frame  $t$ . These  $M$ -dimensional vectors are given by

$$\mathbf{Y}(t, k) = [Y_1(t, k) \ Y_2(t, k) \ \cdots \ Y_M(t, k)]^T \quad (2)$$

$$\mathbf{D}(t, k) = [D_1(t, k) \ D_2(t, k) \ \cdots \ D_M(t, k)]^T \quad (3)$$

$$\mathbf{X}(t, k) = [X_1(t, k) \ X_2(t, k) \ \cdots \ X_M(t, k)]^T \quad (4)$$

in which  $Y_i(t, k)$ ,  $D_i(t, k)$ , and  $X_i(t, k)$  are the DFT coefficients at the  $i$ th microphone and  $[\cdot]^T$  denotes the transpose operation.

To obtain the spatial information of the multi-microphone signal, we apply the eigen-decomposition to the  $M \times M$  correlation matrix  $(\mathbf{R}_{\mathbf{Y}\mathbf{Y}}(t, k) = E \{ \mathbf{Y}(t, k) \mathbf{Y}(t, k)^H \})$  where  $E \{ \cdot \}$

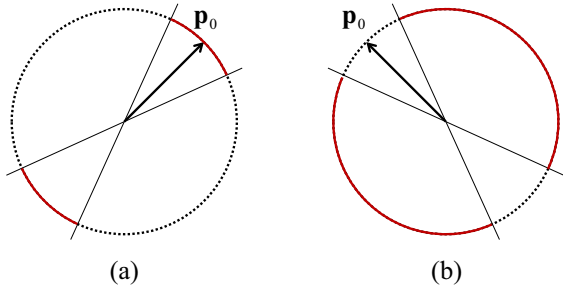


Figure 1: *Graphical description of conic constraints (solid arcs) represented in 2-dimensional real vector space. (a) conic constraint (b) exterior conic constraint.*

denotes the statistical expectation and  $[\cdot]^H$  is the Hermitian operator and take the principal eigenvector (corresponding to the largest eigenvalue). For instance, if a far-field signal is arrived at multiple microphones, the principal eigenvector of the correlation matrix corresponds to the steering vector of the far-field signal which represents the direction of the incoming signal [8].

The eigen-decomposition of the correlation matrix is given by

$$\begin{aligned} \mathbf{R}_{\mathbf{Y}\mathbf{Y}}(t, k) &= \mathbf{Q}(t, k)\mathbf{\Lambda}(t, k)\mathbf{Q}(t, k)^H \\ &= \sum_{i=1}^M \lambda_i(t, k)\mathbf{q}_i(t, k)\mathbf{q}_i(t, k)^H \end{aligned} \quad (5)$$

where  $\mathbf{Q}(t, k)$  is the unitary eigenvector matrix consisting of eigenvectors  $(\mathbf{q}_1(t, k), \mathbf{q}_2(t, k), \dots, \mathbf{q}_M(t, k))$ , and  $\mathbf{\Lambda}(t, k)$  is a diagonal matrix of which diagonal elements  $(\lambda_1(t, k) \geq \lambda_2(t, k) \geq \dots \geq \lambda_M(t, k))$  are the eigenvalues of the correlation matrix. We normalize the principal eigenvector  $\mathbf{q}_1(t, k) = [q_{1,1}(t, k) \ q_{1,2}(t, k) \ \dots \ q_{1,M}(t, k)]^T$  by its first element as

$$\begin{aligned} \bar{\mathbf{q}}_1(t, k) &\triangleq \frac{\mathbf{q}_1(t, k)}{q_{1,1}(t, k)} \\ &= \begin{bmatrix} \frac{q_{1,2}(t, k)}{q_{1,1}(t, k)} & \frac{q_{1,3}(t, k)}{q_{1,1}(t, k)} & \dots & \frac{q_{1,M}(t, k)}{q_{1,1}(t, k)} \end{bmatrix}^T \\ &= [1 \ \bar{q}_{1,1}(t, k) \ \bar{q}_{1,2}(t, k) \ \dots \ \bar{q}_{1,M-1}(t, k)]^T. \end{aligned} \quad (6)$$

Finally, the  $(M-1)$ -dimensional phase vector is defined as

$$\mathbf{p}(t, k) \triangleq \begin{bmatrix} \bar{q}_{1,1}(t, k) & \bar{q}_{1,2}(t, k) & \dots & \bar{q}_{1,M-1}(t, k) \\ |\bar{q}_{1,1}(t, k)| & |\bar{q}_{1,2}(t, k)| & \dots & |\bar{q}_{1,M-1}(t, k)| \end{bmatrix}^T. \quad (7)$$

The  $i$ th element of the phase vector represents the phase of the signal received at the  $(i+1)$ th microphone with respect to the first microphone [7].

### 3. Conic constraints

Conic constraints are quadratic constraints which have been often used for detecting a signal of interest or estimating the steering direction of a signal in a background noise for radar or sonar system [9, 10]. With imperfect knowledge of the signal and interference space, conic constraints are often imposed for robust detection or estimation. Typical constraints are:

- Conic Constraint:

$$\Omega_1 = \left\{ \mathbf{p} \in \mathbb{C}^M : \frac{|\mathbf{p}^H \mathbf{p}_0|^2}{\|\mathbf{p}\|^2 \|\mathbf{p}_0\|^2} \geq \gamma \right\} \quad (8)$$

where  $\mathbf{p}_0$  is the nominal vector and  $0 \leq \gamma \leq 1$ . This is equivalent to limiting the minimum squared cosine angle between the estimated ( $\mathbf{p}$ ) and the nominal ( $\mathbf{p}_0$ ) vectors, namely  $\mathbf{p}$  has to lie in a conic region with axis  $\mathbf{p}_0$  and whose aperture is ruled by  $\gamma$ . A graphical description of the constraint set  $\Omega_1$  is given in Fig. 1 (a) when 2-dimensional real vectors are assumed.

- Exterior Conic Constraint:

$$\Omega_2 = \left\{ \mathbf{p} \in \mathbb{C}^M : \frac{|\mathbf{p}^H \mathbf{p}_0|^2}{\|\mathbf{p}\|^2 \|\mathbf{p}_0\|^2} \leq \gamma \right\}. \quad (9)$$

This is equivalent to limiting the maximum squared cosine angle between the estimated and the nominal phase vectors, namely  $\mathbf{p}$  has to lie outside a conic region whose axis is the vector  $\mathbf{p}_0$ . It arises when it becomes necessary to discriminate between an interfering signal lying in a conic region and a target signal lying in the complement of the quoted region, namely the exterior of the cone. A graphical description of the constraint set  $\Omega_2$  is given in Fig. 1 (b) when 2-dimensional real vectors are assumed.

In this paper, we exploit conic constraints to achieve robustness against the DOA estimation error for the multi-channel VAD in the presence of interference.

## 4. Proposed VAD

In this section, we propose a multi-channel VAD algorithm based on the phase vector and conic constraints. First, the short-time Fourier transform is applied to the multi-microphone signal received at the microphone array. To obtain the phase vector of the multi-microphone signal at every frequency bin, we need to perform the eigen-decomposition to the estimated correlation matrix as shown in Eq. 5. In this paper, instead of computing the eigen-decomposition to obtain the principal eigenvector of the correlation matrix, we use the PASTd algorithm (Projection Approximation Subspace Tracking with deflation) proposed by Yang [11]. While the eigen-decomposition of an  $n$ -by- $n$  correlation matrix requires  $O(n^3)$  operations, the PASTd algorithm requires only  $O(n)$  operation when we only need the principal eigenvector.

The metric for target speech detection for time-frequency region  $(t, k)$  is defined by comparing the two squared cosine angles with conic constraints as:

$$G(t, k) = \begin{cases} 1, & F_X(\mathbf{p}(t, k)) > F_D(\mathbf{p}(t, k)) \\ & \text{and } F_D(\mathbf{p}(t, k)) \leq \gamma_D(k) \\ & \text{and } F_X(\mathbf{p}(t, k)) \geq \gamma_X(k) \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

$$F_X(\mathbf{p}(t, k)) = \frac{|\mathbf{p}(t, k)^H \mathbf{p}_X(t, k)|^2}{\|\mathbf{p}(t, k)\|^2 \|\mathbf{p}_X(t, k)\|^2} \quad (11)$$

$$F_D(\mathbf{p}(t, k)) = \frac{|\mathbf{p}(t, k)^H \mathbf{p}_D(t, k)|^2}{\|\mathbf{p}(t, k)\|^2 \|\mathbf{p}_D(t, k)\|^2}. \quad (12)$$

where  $\mathbf{p}_X(t, k)$  and  $\mathbf{p}_D(t, k)$  are the mean phase vectors for target speech+interference period and interference-only period, respectively.

<Initialization and constants>

1. Initialization of  $\mathbf{p}_D(t, k)$ : the initial mean phase vector for interference-only period is calculated using the initial

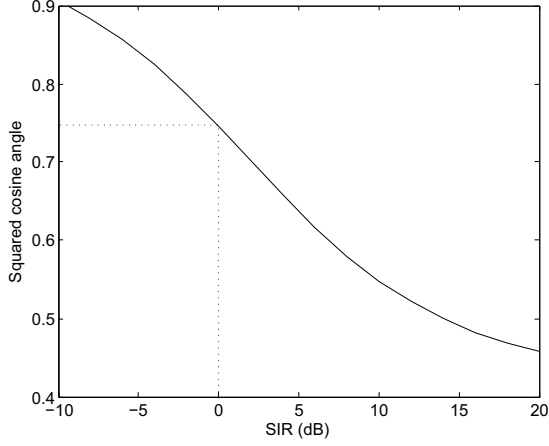


Figure 2: The squared cosine angle between the phase vector of interference-corrupted signal and the phase vector derived from the steering vector of the target signal according to SIR. The curve was obtained by calculating angles from simulated data.

time frames ( $N_D$  frames  $\sim 0.5$  secs) which are assumed to be interference-only period as

$$\mathbf{p}_D(t, k) = \frac{1}{N_D} \sum_{t=1}^{N_D} \mathbf{p}(t, k). \quad (13)$$

2. Initialization of  $\mathbf{p}_X(t, k)$ : with the knowledge of the direction of the target speech, the mean phase vector for target speech+interference period is initially chosen as the phase vector derived from the steering vector of the target speech.
3.  $\gamma_D(k)$ : the constant for exterior conic constraint,  $\gamma_D(k)$  is computed using the mean and the standard deviation of  $F_D(\mathbf{p}(t, k))$  during the initial interference-only period as

$$\gamma_D(k) = \mu - \frac{4}{N_D} \sqrt{\sum_{t=1}^{N_D} (F_D(\mathbf{p}(t, k)) - \mu)^2} \quad (14)$$

where  $\mu = \frac{1}{N_D} \sum_{t=1}^{N_D} F_D(\mathbf{p}(t, k))$ .

4.  $\gamma_X(k)$ :  $\gamma_X(k)$  is chosen as 0.75 which is the squared cosine angle between  $\mathbf{p}(t, k)$  and the phase vector derived from the steering vector of the target signal at 0 dB SIR (see Fig. 2).

Finally, we decide whether the current time frame is target speech present ( $H_1$ ) or absent ( $H_0$ ) by comparing a threshold ( $T$ ) and the averaged value of the metrics  $G(t, k)$  across frequency bins as

$$\Gamma(t) \underset{H_0}{\overset{H_1}{\gtrless}} T \quad (15)$$

$$\Gamma(t) = \frac{1}{l_2 - l_1 + 1} \sum_{k=l_1}^{l_2} G(t, k) \quad (16)$$

where  $l_1$  and  $l_2$  are the frequency bins corresponding to 500 Hz and 6000 Hz, respectively.

The EM (Expectation Maximization) MAP (Maximum *a posteriori*) adaptation of the mean phase vectors are performed

by weighting the contribution of the input phase vector with the *a posteriori* probability as [7]

$$\mathbf{p}_X(t) = \frac{1}{c_X(t)} (\beta c_X(t-1) \cdot \mathbf{p}_X(t-1) + P(F_X(\mathbf{p}(t)) > F_D(\mathbf{p}(t)) | \mathbf{p}(t)) \cdot \mathbf{p}(t)) \quad (17)$$

,if  $F_X(\mathbf{p}(t)) > F_D(\mathbf{p}(t))$  and  $F_X(\mathbf{p}(t)) > \gamma_X$

$$\mathbf{p}_D(t) = \frac{1}{c_D(t)} (\beta c_D(t-1) \cdot \mathbf{p}_D(t-1) + P(F_D(\mathbf{p}(t)) > F_X(\mathbf{p}(t)) | \mathbf{p}(t)) \cdot \mathbf{p}(t)) \quad (18)$$

,if  $F_D(\mathbf{p}(t)) > F_X(\mathbf{p}(t))$  and  $F_D(\mathbf{p}(t)) > \gamma_D$

where

$$P(F_X(\mathbf{p}(t)) > F_D(\mathbf{p}(t)) | \mathbf{p}(t)) = \frac{F_X(\mathbf{p}(t))}{F_X(\mathbf{p}(t)) + F_D(\mathbf{p}(t))} \quad (19)$$

and

$$P(F_D(\mathbf{p}(t)) > F_X(\mathbf{p}(t)) | \mathbf{p}(t)) = \frac{F_D(\mathbf{p}(t))}{F_X(\mathbf{p}(t)) + F_D(\mathbf{p}(t))}. \quad (20)$$

The smoothing factor,  $\beta$  is experimentally chosen as 0.8 and  $c_X(t)$ ,  $c_D(t)$  are smoothed with  $\beta$  as

$$c_X(t) = \beta c_X(t-1) + P(F_X(\mathbf{p}(t)) > F_D(\mathbf{p}(t)) | \mathbf{p}(t)) \quad (21)$$

$$c_D(t) = \beta c_D(t-1) + P(F_D(\mathbf{p}(t)) > F_X(\mathbf{p}(t)) | \mathbf{p}(t)) \quad (22)$$

with  $c_X(1) = c_D(1) = N_D/2$ . The frequency bin index  $k$  is omitted for brevity for Eqs. 17-22.

## 5. Simulation results

We tested the proposed algorithm in the presence of competing speech. For competing speech interference, we prepared a news clip recorded at 16 kHz sampling rate. The multi-microphone signals were generated by the convolution of sound sources with acoustic impulse responses. The impulse responses were obtained from the RWCP Sound Scene Database [12] which were measured 2 m away from the center of the microphone array in real environments. The reverberation time was around 300 ms. The microphone array is a linear type and has 7 microphones with 5.66 cm uniform intervals. The multi-microphone target signal was created by convolving the target source with the impulse response measured in front of the microphone array. In the same way, for the interference, multi-microphone competing speech was generated as coming at the angle of 40° to the direction of the target signal (see Fig. 3). The multi-channel interference corrupted the target signal at the SIR level of 0 dB.

The simulation results in the presence of competing speech are displayed in Fig. 4. The top panel shows the target speech captured by the center microphone. The noisy speech signal corrupted by the directional interference (competing speech) at 0 dB SIR is shown in the second panel. The third panel shows the result of the CPSP after steering toward the direction of the target speech [5,6] for the purpose of comparison. The proposed VAD metric (as per Eq. 16) is shown in the bottom panel. The true DOA for the target speech was 90° as shown in Fig. 3 and the estimated DOA was assumed to be 90°, 85°, and 80° in this simulation. Both the CPSP and the proposed VAD metric performed well when there was no DOA estimation error (when the estimated DOA is 90°). While the CPSP cannot be used

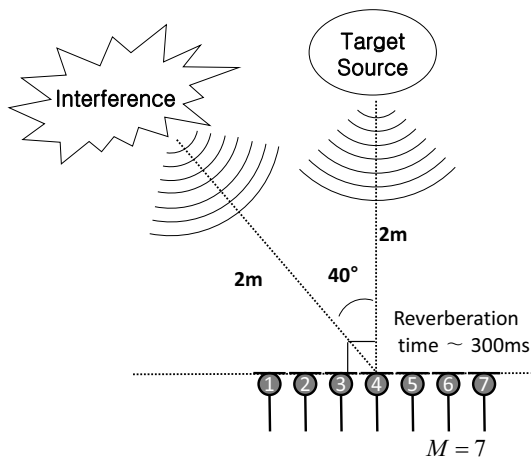


Figure 3: Simulation environment: target source and interference arriving at a linear microphone array.

as a VAD metric any more when there is  $5^\circ$  mismatch for the DOA estimation, the proposed VAD metric still presented reliable performance with  $10^\circ$  error in the DOA estimation (when the estimated DOA was  $80^\circ$ ).

## 6. Conclusions

A new multi-channel VAD was proposed using the spatial information of the observed signal. The proposed algorithm uses the phase vector as VAD feature and detects the presence of target speech based on conic constraints. The proposed algorithm was evaluated by the simulated multi-channel data generated using multi-channel impulse responses measured in real environments. The simulation results demonstrated that the proposed algorithm can detect the presence of target speech reliably in the presence of competing speech at SIR level as low as 0 dB. The results also showed that the proposed algorithm is less sensitive to the DOA estimation error.

## 7. 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 Knowledge Economy (MKE). This work was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0014143).

## 8. References

- [1] O. Hoshuyama, B. Begasse, A. Sugiyama, and A. Hirano, "A real-time robust adaptive microphone array controlled by an SNR estimate," in *Proc. of IEEE Intern. Conf. on Acoust., Speech, Signal Processing*, 1998, pp. 3605–3608.
- [2] M. W. Hoffman, Z. Li, and D. Khataniar, "GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 175–179, 2001.
- [3] R. L. Bouquin-Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech Commun.*, vol. 16, no. 3, pp. 245–254, 1995.
- [4] A. Guerin, "A two-sensor voice activity detection and speech enhancement based on coherence with additional enhancement of low frequencies using pitch information," in *Proc. of European Signal Processing Conf.*, 2000.

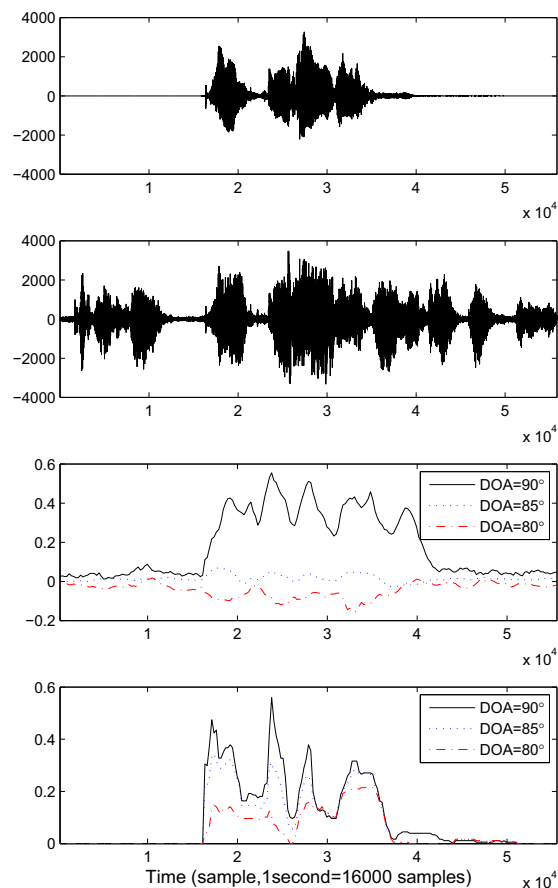


Figure 4: Simulation results in the presence of competing speech. Top panel shows the target speech captured by the center microphone. The second panel shows the noisy signal contaminated by competing speech at 0 dB SIR. The third and bottom panel show the CSP and the proposed VAD metric as per Eq. 16, respectively. The true DOA was  $90^\circ$  and three different cases were tested with estimated DOAs of  $90^\circ$ ,  $85^\circ$ , and  $80^\circ$ .

- [5] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," in *Proc. of Interspeech*, 2003, pp. 501–504.
- [6] G. Kim and N. I. Cho, "Two-microphone voice activity detection in the presence of coherent interference," in *Proc. of Interspeech*, 2006, pp. 1686–1689.
- [7] —, "Voice activity detection using phase vector in microphone array," *Electronics Letters*, vol. 43, no. 14, pp. 783–784, July 2007.
- [8] H. V. Trees, *Detection, estimation and modulation theory, Part IV: Optimum Array Processing*. New York: Wiley, 2002.
- [9] S. Ramprasad, T. W. Parks, and R. Shenoy, "Signal modeling and detection using cone classes," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 329–338, 1996.
- [10] A. De Maio, Y. Huang, D. P. Palomar, S. Zhang, and A. Farina, "Fractional QCQP with applications in ML steering direction estimation for radar detection," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 172–185, 2011.
- [11] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 95–107, 1995.
- [12] "RWCP sound scene database in real acoustical environments," Real World Computing Partnership, (c)1998-2001.