POLYNOMIAL EIGENVALUE DECOMPOSITION-BASED TARGET SPEAKER VOICE ACTIVITY DETECTION IN THE PRESENCE OF COMPETING TALKERS

Vincent W. Neo[†], *Stephan Weiss*[*], *Simon W. McKnight*[†], *Aidan O. T. Hogg*[†], *Patrick A. Naylor*[†]

[†]Department of Electrical and Electronic Engineering, Imperial College London, UK
[*]Department of Electronic and Electrical Engineering, University of Strathclyde, Scotland
{vincent.neo09, s.mcknight18, a.hogg, p.naylor}@imperial.ac.uk, stephan.weiss@strath.ac.uk

ABSTRACT

Voice activity detection (VAD) algorithms are essential for many speech processing applications, such as speaker diarization, automatic speech recognition, speech enhancement, and speech coding. With a good VAD algorithm, non-speech segments can be excluded to improve the performance and computation of these applications. In this paper, we propose a polynomial eigenvalue decomposition-based target-speaker VAD algorithm to detect unseen target speakers in the presence of competing talkers. The proposed approach uses frame-based processing across multi-microphones to compute the syndrome energy, used for testing the presence or absence of a target speaker. The proposed approach is consistently among the best in F1 and balanced accuracy scores over the investigated range of signal to interference ratio (SIR) from -10 dB to 20 dB.

Index Terms— polynomial eigenvalue decomposition, target speaker voice activity detection, speaker activity detection

1. INTRODUCTION

Voice activity detection (VAD) algorithms play an essential role in speech processing applications, such as speaker diarization [1, 2], automatic speech recognition (ASR) systems [3], speech enhancement [4], and speech coding [5]. Typical VAD algorithms classify audio frames as speech or non-speech. When the VAD labels are correct, non-speech segments can be excluded to improve the application performance and computation. Conversely, mis-classification will not improve the computation when non-speech frames are labelled as 'speech present', while application performance may degrade when speech frames are labelled as 'speech absent', e.g., word deletion in ASR and noise estimation using speech frames.

Classical statistical-based VAD approaches such as [5–8] exploit the statistics of speech and noise. These approaches compute the model parameters based on the assumptions of the speech and noise distributions. However, the performance of these algorithms degrades when the assumed signal statistics are violated and the speech presence probability, which the algorithms usually exploit, is difficult to deduce analytically [9].

Machine learning-based VAD methods have also been proposed to model implicitly the data without using an explicit noisy signal model [10–12]. The VAD produced for real-time applications in the WebRTC project [13], which uses a Gaussian mixture model (GMM) trained in recognizing speech features [9], is now widely used in many systems even if they are not real-time [9]. There are many other machine learning approaches to VAD as well, including variations of methods that have also been useful for speaker recognition and diarization such as time delay neural networks (TDNNs) [14].

A major drawback of these approaches is that they do not work well when the background noise also comprises speech, such as in a restaurant [15, 16]. More recent approaches have incorporated speaker-specific information directly into the VAD employed, such as end-to-end neural speaker diarization [15], target speaker-VAD (TS-VAD) [17] and personal VAD [16]. These methods require substantial training data and are usually not designed for unseen speakers absent in the training set. In the DIHARD III speaker diarization challenge [18], many top entries used VAD with some speaker-specific information along with their diarization steps. In particular, the winning entry [19] used an iterative TS-VAD as part of their system intended to generalize to unseen speakers.

In [20], a broadband subspace approach is used to detect weak transient signals. The approach uses an iterative polynomial matrix eigenvalue decomposition (PEVD) algorithm such as the family of second-order sequential best rotation (SBR2) [21, 22] and sequential matrix diagonalization (SMD) approaches [23, 24] in the time domain or [25, 26] in the frequency domain. PEVD algorithms have been found useful for many broadband signal processing applications such as speech enhancement [27, 28], source separation [29, 30], source localizaton [31, 32] and beamforming [33].

This paper extends the work in [20] from transient signal detection to TS-VAD for unseen targets. To achieve this, we adopt a different frame-based multi-microphone approach to generate the syndrome vector. The syndrome energy, computed from the syndrome vector, is then used to test for the presence or absence of the target signal to generate a binary mask for every frame using a novel detection method. The contributions of this paper are (i) a novel PEVD-based TS-VAD algorithm and (ii) a comparison of the proposed approach with benchmark VAD approaches in simulations using realistic speech signals and measured impulse responses.

2. PROBLEM FORMULATION

2.1. Signal Model

The received signal at the q-th microphone is

$$x_q(n) = \sum_{p=1}^{P} \mathbf{h}_{p,q}^T(n)\mathbf{s}_p(n) \ , \qquad (1)$$

where $\mathbf{h}_{p,q} = [h_{p,q}(0), \ldots, h_{p,q}(J)]^T$ represents the room impulse response (RIR) from the p-th source to the q-th micro-

The work of S. Weiss was supported by the Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/S000631/1, and the UK MOD University Defence Research Collaboration in Signal Processing; the work of P. A. Naylor was funded through the UK EPSRC grant no. EP/S035842/1.

978-1-6654-6867-1/22/$31.00 ©2022 IEEE

phone, modelled as a J-th order finite impulse response filter, $\mathbf{s}_p(n) = [s_p(n), \ldots, s_p(n-J)]^T$ is the p-th source signal, n is the sample index, and $[\cdot]^T$ is the transpose operator. The data vector over Q microphones is $\mathbf{x}(n) = [x_1(n), \ldots, x_Q(n)]^T \in \mathbb{R}^Q$.

Since the P source signals are not simultaneously excited all the time, the goal of a TS-VAD algorithm is to identify time segments when the p-th target source is active.

2.2. Polynomial Matrix Eigenvalue Decomposition

The space-time covariance matrix [21, 34], parameterized by time lag $\tau \in \mathbb{Z}$, is computed using

$$\mathbf{R}(\tau) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^T(n-\tau)\} \ , \tag{2}$$

where $\mathbb{E}\{\cdot\}$ is the expectation operator over n. Each element, $r_{p,q}(\tau)$, is the correlation sequence between the p-th and q-th microphone signals. This produces auto- and cross-correlation sequences on the diagonals and off-diagonals, respectively.

The z-transform of (2),

$$\boldsymbol{\mathcal{R}}(z) = \sum_{\tau=-\infty}^{\infty} \mathbf{R}(\tau) z^{-\tau} \ , \tag{3}$$

denoted by $\mathbf{R}(\tau) \circ\!\!-\!\!\bullet \boldsymbol{\mathcal{R}}(z)$, is a para-Hermitian polynomial matrix satisfying $\boldsymbol{\mathcal{R}}(z) = \boldsymbol{\mathcal{R}}^P(z) = \boldsymbol{\mathcal{R}}^H(1/z^*)$, where $[\cdot]^*$, $[\cdot]^H$, $[\cdot]^P$ are the complex conjugate, Hermitian and para-Hermitian operators respectively. The para-Hermitian matrix eigenvalue decomposition (EVD) of (3) is [34, 35]

$$\boldsymbol{\mathcal{R}}(z) = \boldsymbol{\mathcal{U}}(z)\,\boldsymbol{\Lambda}(z)\boldsymbol{\mathcal{U}}^P(z) \ , \tag{4}$$

where the columns of $\boldsymbol{\mathcal{U}}(z)$ are the polynomial eigenvectors and the elements on the diagonal matrix $\boldsymbol{\Lambda}(z)$ are the polynomial eigenvalues. Iterative PEVD algorithms based on the SBR2 [21, 22] and SMD [23, 24] are used to approximate (4) by Laurent polynomial factors.

Exploiting the orthogonality between subspaces and assuming $M = P - 1$ interferers in the absence of the target speaker, (4) can be partitioned into

$$\boldsymbol{\mathcal{R}}(z) = \begin{bmatrix} \boldsymbol{\mathcal{U}}_s(z) & \boldsymbol{\mathcal{U}}_\perp(z) \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_s(z) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_{\bar{s}}(z) \end{bmatrix} \begin{bmatrix} \boldsymbol{\mathcal{U}}_s^P(z) \\ \boldsymbol{\mathcal{U}}_\perp^P(z) \end{bmatrix} , \tag{5}$$

where $\mathbf{0}$ is a zero matrix, $\boldsymbol{\Lambda}_s : \mathbb{C} \to \mathbb{C}^{M \times M}$ contains the M principal eigenvalues of the interferer-related components with its eigenvectors on the columns of $\boldsymbol{\mathcal{U}}_s(z) : \mathbb{C} \to \mathbb{C}^{Q \times M}$ while the eigenvalues $\boldsymbol{\Lambda}_{\bar{s}} : \mathbb{C} \to \mathbb{C}^{(Q-M)\times(Q-M)}$ defines the noise floor along with the orthogonal complement subspace of the interferers spanned by the columns of $\boldsymbol{\mathcal{U}}_\perp(z) : \mathbb{C} \to \mathbb{C}^{Q \times (Q-M)}$.

3. POLYNOMIAL EVD-BASED TARGET SPEAKER VOICE ACTIVITY DETECTION ALGORITHM

We extend the work in [20] and propose two modifications including (i) frame-based processing for syndrome generation and (ii) a different detection approach for TS-VAD.

3.1. Frame-based Syndrome Generation

The multi-microphone signals are first processed using L frames, each of size T, assuming no overlap between frames. Therefore, the data vector for the i-th sample index in the ℓ-th frame $\mathbf{x}_\ell(i) \in \mathbb{R}^Q$ can be written as

$$\mathbf{x}_\ell(i) = \mathbf{x}(\ell T + i) \ , \quad i = 0, \ldots, T-1 \ . \tag{6}$$

The first L_I frames is assumed to contain only the interferers, and the space-time covariance in (2) can be estimated using [36, 37]. After computing its PEVD, the orthogonal complement subspace $\boldsymbol{\mathcal{U}}_\perp(z)$ based on (5) is generated. For each frame, a syndrome vector $\mathbf{y}_\ell(i) \in \mathbb{C}^{(Q-M)}$ is computed by filtering the data vector through the eigenvector $\boldsymbol{\mathcal{U}}_\perp(z) \bullet\!\!-\!\!\circ \mathbf{U}_\perp(n)$ using

$$\mathbf{y}_\ell(i) = \sum_\nu \mathbf{U}_\perp^H(-\nu)\mathbf{x}_\ell(i-\nu) \ . \tag{7}$$

The syndrome vector $\mathbf{y}_\ell(i)$, whose change in statistics indicates the presence of the target speaker, is similar to the projection of $\mathbf{x}_\ell(i)$ onto a smaller $(Q-M)$-dimensional subspace. This projection removes components associated with the interferer $\boldsymbol{\mathcal{U}}_s(z)$. In the interferer-only case, the energy of $\mathbf{y}_\ell(i)$ is expected to be smaller than $\mathbf{x}_\ell(i)$ if there are only M interferer signals, as seen in (5).

3.2. Target Speaker Voice Activity Detection

Instead of a decimated subspace detector in [20], a simple hard thresholding mechanism is used in our proposed VAD design. The energy of the syndrome data $\mathbf{y}_\ell(i)$ for the ℓ-th time frame can be calculated using

$$\xi_\ell = \sum_{i=0}^{T-1} \|\mathbf{y}_\ell(i)\|_2^2 \ , \tag{8}$$

where $\|\cdot\|_2$ is the Euclidean-norm of a vector. When the target speaker begins to talk, some of the signal components associated with the target is likely to protrude into the orthogonal complement subspace $\boldsymbol{\mathcal{U}}_\perp(z)$ and result in a large value for (8). In this case, changes in the syndrome energy become more easily detectable than the energy of the microphone signals.

A threshold value ξ_I indicating the absence of the target speaker can be calculated by averaging across the L_I interferer-only frames. Any deviation from ξ_I is subsequently used to detect the presence of the target on a frame-by-frame basis. For the ℓ-th time frame, a binary mask $m(\ell)$ is generated based on

$$m(\ell) = \begin{cases} 1, & \xi_\ell > \xi_I \ , \\ 0, & \text{otherwise} \ . \end{cases} \tag{9}$$

The mask aims to identify frames containing the target. Inevitably, this may also include the interferer when both talkers are speaking. The target source $s(n)$ is extracted by applying the mask using

$$\hat{s}(n) = [m(\frac{n}{T}) * p_T(n)] \cdot x_1(n) \ , \tag{10}$$

where $m(t) = 0$ unless $t \in \mathbb{Z}$ such that $n = \ell T$, $p_T(n)$ is a Dirichlet or rectangular window of length T, $x_1(n)$ is the first reference microphone, and $*$ denotes the linear convolution operator.

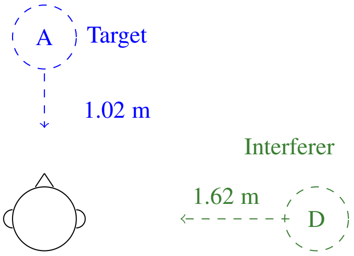


Fig. 1: Cafeteria setup from Kayser database [39].

4. SIMULATION AND RESULTS

4.1. Setup

The target and interferer speech signals were taken from the VCTK corpus [38] and the 2-channel RIRs were taken from the in-ear recordings in the Kayser database [39]. The speech and interferer signals were separately convolved with the RIRs before being added together at each microphone. The signal to interference ratio (SIR) [40] at the first microphone, taken to be the reference, was varied from -10 dB to 20 dB. The target speaker and directional interferer are respectively 1.02 m in front and 1.62 m to the right of the listener, at positions A and D shown in Fig. 1 [39]. The directional interferer in all experiments was the same male talker who spoke continuously without pauses.

The proposed PEVD approach was compared against Sohn [5] and WebRTC approaches [13]. WebRTC [13] operates at modes 0–3 from the least to the most aggressive setting. The microphone signals were processed in 30 ms frames. The time support used for (3) in the proposed PEVD-based approach was also 30 ms. The SMD algorithm was used for computing PEVD and the parameters are based on [27]. The first 500 ms were assumed to contain only the interferer signals and were therefore used for calculating (5).

4.2. Ground Truth Labels

A similar procedure described in [41] is used to establish the ground truth (GT) labels. The RIR from the target to the first microphone, chosen as the reference, is truncated approximately 5 ms after the direct-path peak. The truncation is necessary to ensure that the target speech is time aligned with the microphone signals while minimizing reverberation. The anechoic target speech signal is then convolved with the truncated RIR to generate the target speech in $x_1(n)$. The VAD algorithm mode 3 [13] is applied to the target signal to generate the ground truth VAD labels as shown in Fig. 2. Informal listening examples for the VAD outputs are available [42].

4.3. Evaluation Measures

The counts for the ground truth and predicted labels are tabulated in Table 1 using a confusion matrix [43]. The absence or presence of speech is indicated by the label ‘0’ or ‘1’. A true positive (TP) and a true negative (TN) are obtained when both labels are ‘1’ and ‘0’ respectively. A false negative (FN) occurs when the predicted label is ‘0’ but the ground truth is ‘1’ while a false positive (FP) happens when the predicted label is ‘1’ but the ground truth is ‘0’. This allows the use of F1, true positive rate (TPR), true negative rate (TNR), and balanced accuracy (BACC) scores defined as [43]

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\,,\ \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}\,,$$
$$\mathrm{F1} = \frac{\mathrm{TP}}{\mathrm{TP}+0.5\times(\mathrm{FP}+\mathrm{FN})}\,,\ \mathrm{BACC} = \frac{\mathrm{TPR}+\mathrm{TNR}}{2}\,. \quad (11)$$

4.4. Experiments and Discussions

4.4.1. Experiment 1: Female Target and Male Interferer

The male interferer spoke continuously at a SIR of 5 dB. The first 500 ms was used by the PEVD approach to generate the orthogonal complement of the interferer subspace $\boldsymbol{\mathcal{U}}_\perp(z)$ and to calculate the threshold value ξ_I. The syndrome vector is generated for each frame by filtering the data vector through $\boldsymbol{\mathcal{U}}_\perp(z)$ in (7) to derive the syndrome energy ξ. The evolution of the syndrome energy is plotted along with the target and received signals in Fig 2(a). The syndrome energy follows the envelope of the target signal, and a thresholding mechanism can be used to design a detector.

The results for the different VAD algorithms are shown in Fig 2(a) and Table 1(a). The proposed PEVD approach is the best performing algorithm with F1 and BACC scores of 0.820 and 0.667, respectively, while Sohn comes second. The PEVD approach can better identify the absence of a target source as shown by the missed detections before 1 s, between 6–8 s and after 14 s by other methods in Fig 2(a) and scores the best in TN with 72 frames. The plot shows that G0 and G3 tend to label the frames as containing the target speech, e.g., binary mask values of 1, resulting in high FP.

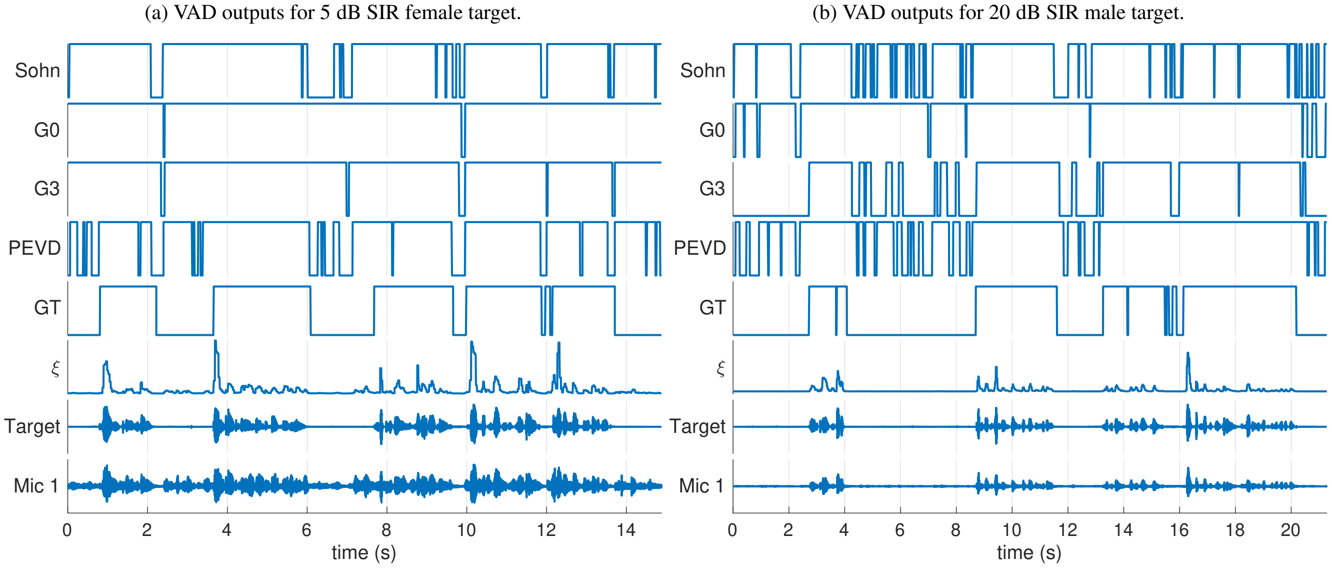
4.4.2. Experiment 2: Male Target and Male Interferer

The target speaker and interferer in this experiment are both males at 20 dB SIR. Changes in the syndrome energy ξ with the target speech as well as the received signal in the first microphone are shown in Fig. 2(b). Compared to the female target example in Experiment 1, the PEVD-based VAD for the male target speaker gives a relatively high FP score. This might result from subspace leakage since both male speakers are likely to be more similar than a pair of male and female speakers. G3 gives the best improvement in F1 and BACC scores, followed by PEVD, while G0 performs the worst.

When the SIR worsens to -10 dB, Sohn performs better than PEVD in F1 score by 0.01, while PEVD scores better in BACC by 0.047. Compared to Sohn, PEVD can better identify the absence than the presence of the target speaker in frames. This result is also consistent with the other SIR and the female target talker. The WebRTC methods, G0 and G3, do not perform well, and they tend to label frames as containing the target speaker, resulting in very low TN and high FP scores.

5. CONCLUSION

In this work, we have proposed a novel PEVD-based TS-VAD for detecting speech from a target speaker in the presence of competing talkers. We have introduced two main modifications to an earlier approach for the detection of weak transient signals, namely, frame-based processing and a thresholding mechanism that generates a binary mask. The energy of the syndrome matrix at each frame has been shown to follow the overall envelope of the target signal and can be used to detect its presence. The proposed approach is consistently among the best in F1 and BACC scores over the investigated range of SIR from -10 dB to 20 dB.



(a) VAD outputs for 5 dB SIR female target.

(b) VAD outputs for 20 dB SIR male target.

Fig. 2: Comparison of VAD outputs $m_1(n)$ using Sohn VAD [5], WebRTC Modes 0 and 3 (G0, G3) [13], and the proposed PEVD-based VAD. The ground truth (GT), syndrome $y(n)$, target and microphone 1 signals are plotted for reference. The VAD outputs are plotted for (a) 5 dB SIR target female speaker, and (b) 20 dB SIR target male speaker.

Table 1: Confusion matrix and scores for VAD output on target speaker in the presence of a competing talker at various SIR.

(a) Female target in SIR = 5 dB

Metric	Sohn	G0	G3	PEVD
TP	295	313	310	294
TN	43	4	10	72
FP	139	178	172	110
FN	18	0	3	19
F1	0.790	0.779	0.780	**0.820**
BACC	0.589	0.511	0.523	**0.667**

(b) Male target in SIR = 20 dB

Metric	Sohn	G0	G3	PEVD
TP	343	357	350	356
TN	84	36	277	120
FP	267	316	75	232
FN	14	0	7	1
F1	0.709	0.693	**0.895**	0.753
BACC	0.600	0.551	**0.884**	0.669

(c) Male target in SIR= -10 dB

Metric	Sohn	G0	G3	PEVD
TP	338	360	354	294
TN	57	0	4	128
FP	273	330	326	202
FN	24	2	8	68
F1	**0.695**	0.684	0.679	0.685
BACC	0.553	0.497	0.495	**0.600**

6. REFERENCES

[1] S. W. McKnight, A. O. T. Hogg, V. W. Neo, and P. A. Naylor, “A study of salient modulation domain features for speaker identification,” in *Asia-Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf. (APSIPA)*, Dec. 2021, pp. 705–712.

[2] A. O. T. Hogg, C. Evers, A. H. Moore, and P. A. Naylor, “Overlapping speaker segmentation using multiple hypothesis tracking of fundamental frequency,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1479–1490, Mar. 2021.

[3] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, ser. Signal Processing Series. New Jersey: Prentice Hall, 1993.

[4] S. Doclo and M. Moonen, “GSVD-based optimal filtering for single and multimicrophone speech enhancement,” *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[5] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[6] 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,” Int. Telecommun. Union (ITU-T), Recommendation, Jun. 2012.

[7] S. Gazor and W. Zhang, “A soft voice activity detector based on a Laplacian–Gaussian model,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 498–505, Sep. 2003.

[8] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[9] J. H. Ko, J. Fromm, M. Philipose, I. Tashev, and S. Zarar, “Limiting numerical precision of neural networks to achieve real-time voice activity detection,” in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2018, pp. 2236–2240.

[10] Z.-L. Zhang and D. Wang, “Boosting contextual information for deep neural network based voice activity detection,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 2, pp. 252–264, Feb. 2016.

[11] A. Ivry, B. Berdugo, and I. Cohen, “Voice activity detection for transient noisy environment based on diffusion nets,” *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 254–264, May 2019.

[12] J. W. Shin, J.-H. Chang, and N. S. Kim, “Voice activity detection based on statistical and machine learning approaches,” *Comput. Speech and Language*, vol. 24, no. 2010, pp. 515–530, Mar. 2009.

[13] Google, "WebRTC Voice Activity Detector," 2021. [Online]. Available: https://github.com/wiseman/py-webrtcvad

[14] Y. Bai, J. Yi, J. Tao, Z. Wen, and B. Liu, "Voice activity detection based on time-delay neural networks," in *Asia-Pacific Signal and Inform. Process. Assoc. Annual Summit and Conf. (APSIPA)*, Nov. 2019, pp. 1173–1178.

[15] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop*, 2019, pp. 296–303.

[16] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I. L. Moreno, "Personal VAD: Speaker-conditioned voice activity detection," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2020, pp. 433–439.

[17] I. Medennikov *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, Oct. 2020, pp. 274–278.

[18] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third DIHARD challenge evaluation plan," Dec. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2006.05815

[19] Y. Wang *et al.*, "USTC-NELSLIP system description for DIHARD-III challenge," Mar. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2103.10661

[20] S. Weiss, C. Delaosa, J. Matthews, I. K. Proudler, and B. A. Jackson, "Detection of weak transient signals using a broadband subspace approach," in *Sensor Signal Process. for Defence Conf. (SSPD)*, Sep. 2021.

[21] J. G. McWhirter, P. D. Baxter, T. Cooper, S. Redif, and J. Foster, "An EVD algorithm for para-hermitian polynomial matrices," *IEEE Trans. Signal Process.*, vol. 55, no. 5, pp. 2158–2169, May 2007.

[22] Z. Wang, J. G. McWhirter, J. Corr, and S. Weiss, "Multiple shift second order sequential best rotation algorithm for polynomial matrix EVD," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2015, pp. 844–848.

[23] V. W. Neo and P. A. Naylor, "Second order sequential best rotation algorithm with Householder transformation for polynomial matrix eigenvalue decomposition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2019, pp. 8043–8047.

[24] S. Redif, S. Weiss, and J. G. McWhirter, "Sequential matrix diagonalisation algorithms for polynomial EVD of parahermitian matrices," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 81–89, Jan. 2015.

[25] F. K. Coutts, K. Thompson, I. K. Proudler, and S. Weiss, "An iterative DFT-based approach to the polynomial matrix eigenvalue decomposition," in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, 2018, pp. 1011–1015.

[26] S. Weiss, I. K. Proudler, and F. K. Coutts, "Eigenvalue decomposition of a parahermitian matrix: Extraction of analytic eigenvalues," *IEEE Trans. Signal Process.*, vol. 69, pp. 722–737, 2021.

[27] V. W. Neo, C. Evers, and P. A. Naylor, "Enhancement of noisy reverberant speech using polynomial matrix eigenvalue decomposition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3255–3266, Oct. 2021.

[28] V. W. Neo, C. Evers, and P. A. Naylor, "Polynomial matrix eigenvalue decomposition of spherical harmonics for speech enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Jun. 2021, pp. 786–790.

[29] S. Redif, S. Weiss, and J. G. McWhirter, "Relevance of polynomial matrix decompositions to broadband blind signal separation," *Signal Process.*, vol. 134, pp. 76–86, May 2017.

[30] V. W. Neo, C. Evers, and P. A. Naylor, "Polynomial matrix eigenvalue decomposition-based source separation using informed spherical microphone arrays," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, Oct. 2021, pp. 201–205.

[31] W. Coventry, C. Clemente, and J. Soraghan, "Enhancing polynomial MUSIC algorithm for coherent broadband sources through spatial smoothing," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2017, pp. 2448–2452.

[32] A. O. T. Hogg, V. W. Neo, C. Evers, S. Weiss, and P. A. Naylor, "A polynomial eigenvalue decomposition MUSIC approach for broadband sound source localization," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, Oct. 2021.

[33] S. Weiss, S. Bendoukha, A. Alzin, F. K. Coutts, I. K. Proudler, and J. Chambers, "MVDR broadband beamforming using polynomial matrix techniques," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2015, pp. 839–843.

[34] S. Weiss, J. Pestana, and I. K. Proudler, "On the existence and uniqueness of the eigenvalue decomposition of a parahermitian matrix," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2659–2672, May 2018.

[35] S. Weiss, J. Pestana, I. K. Proudler, and F. K. Coutts, "Corrections to "On the Existence and Uniqueness of the Eigenvalue Decomposition of a Parahermitian Matrix"," *IEEE Trans. Signal Process.*, vol. 66, no. 23, pp. 6325–6327, Dec. 2018.

[36] C. Delaosa, J. Pestana, N. J. Goddard, S. Somasundaram, and S. Weiss, "Sample space-time covariance matrix estimation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2019, pp. 8033–8037.

[37] C. Delaosa, J. Pestana, N. J. Goddard, S. Somasundaram, and S. Weiss, "Support estimation of a sample space-time covariance matrix," in *Sensor Signal Process. for Defence Conf. (SSPD)*, 2019.

[38] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus design, collection and data analysis of a large regional accent speech database," in *Conf. Asian Spoken Language Research and Evaluation*, Nov. 2013.

[39] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP J. on Advances in Signal Process.*, vol. 2009, no. 1, p. 298605, Jul. 2009.

[40] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[41] S. Braun and I. Tashev, "On training targets for noise-robust voice activity detection," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2021, pp. 421–425.

[42] V. W. Neo, "PEVD-based speaker activity detection in the presence of competing talkers," Apr. 2022. [Online]. Available: https://vwn09.github.io/research/pevd-tsvad-iwaenc

[43] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, Aug. 2018.