A Polynomial Subspace Projection Approach for the Detection of Weak Voice Activity

Vincent W. Neo[†], Stephan Weiss[*], 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, p.naylor}@imperial.ac.uk, stephan.weiss@strath.ac.uk

***Abstract*—A voice activity detection (VAD) algorithm identifies whether or not time frames contain speech. It is essential for many military and commercial speech processing applications, including speech enhancement, speech coding, speaker identification, and automatic speech recognition. In this work, we adopt earlier work on detecting weak transient signals and propose a polynomial subspace projection pre-processor to improve an existing VAD algorithm. The proposed multi-channel pre-processor projects the microphone signals onto a lower dimensional subspace which attempts to remove the interferer components and thus eases the detection of the speech target. Compared to applying the same VAD to the microphone signal, the proposed approach almost always improves the F1 and balanced accuracy scores even in adverse environments, e.g. -30 dB SIR, which may be typical of operations involving noisy machinery and signal jamming scenarios.**

***Index Terms*—Voice activity detection, polynomial matrix eigenvalue decomposition, multi-channel signal processing**

I. Introduction

A voice activity detection (VAD) algorithm identifies whether or not time frames contain speech. VAD is essential for many military and commercial speech processing applications such as speech enhancement [1], speech coding [2], speaker identification [3], [4], and automatic speech recognition (ASR) systems [5]. For example, speech enhancement algorithms may facilitate communication among operators in military operations where the acoustic environment is very challenging, e.g., very noisy machinery and signal jamming scenarios. Such algorithms, however, usually rely on noise estimators, which can be derived from the VAD pre-processing.

Classical statistics-based VAD approaches such [2], [6]–[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 VAD algorithms usually exploit, is difficult to deduce analytically [9]. Furthermore, during noise-only segments, rapidly changing noise can result in transient interference [10].

Machine learning-based VAD methods have also been proposed to implicitly model the data without using an explicit noisy signal model [10]–[12]. Amongst many, a VAD algorithm, which uses a Gaussian mixture model (GMM) trained in recognizing speech features, has been widely adopted for real-time applications in the WebRTC system [13]. The algorithm cannot cope with noisy environments where it becomes challenging to extract speech features, severely degrading its performance [9], [11].

In [14], a broadband subspace-based approach is used to detect weak transient signals. The approach applies a polynomial matrix eigenvalue decomposition (PEVD), which is iteratively approximated by algorithms such as the second-order sequential best rotation (SBR2) [15], [16] and sequential matrix diagonalization (SMD) [17], [18] in the time-domain or [19], [20] in the discrete Fourier transform (DFT)-domain, to generate the eigenvectors and eigenvalues. Filtering the signal through the eigenvector filterbank yields a syndrome vector, which is more discriminative towards detecting a transient signal [14].

In this work, we adapt [14] and investigate the idea of weak transient signal detection for VAD. The novel contributions of this paper are (i) a subspace-projection approach for VAD instead of the syndrome vector approach used for weak transient signal detection in [14], (ii) the use of realistic speech signals and measured room impulse responses (RIRs) and (iii) a comparison of the proposed approach against benchmark VAD algorithms in adverse environments. We first describe the goal of a VAD algorithm and provide a review of PEVD in Section II. The proposed method based on a multi-channel polynomial subspace projection is presented in Section III. Simulations and results are discussed in Section IV and Section V concludes our findings.

II. Problem Formulation

A. 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 RIR from the p-th source to the q-th microphone, 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

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.

sample index, and $[\cdot]^T$ is the transpose operator. The data vector over Q microphones is $\mathbf{x}(n) = [x_1(n), \ldots, x_Q(n)]^T$.

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

B. Polynomial Matrix Eigenvalue Decomposition

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

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

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 eigenvalue decomposition (EVD) of (3) is [21], [22]

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

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 [15], [16] and SMD [18], [23] are used to approximate (4) by Laurent polynomial factors.

Exploiting the orthogonality between subspaces and assuming L signal components, (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, $\boldsymbol{\Lambda}_s : \mathbb{C} \to \mathbb{C}^{L\times L}$ contains the L principal eigenvalues of the signal-related components with its eigenvectors on the columns of $\boldsymbol{\mathcal{U}}_s(z) : \mathbb{C} \to \mathbb{C}^{Q\times L}$ while the eigenvalues $\boldsymbol{\Lambda}_{\bar{s}} : \mathbb{C} \to \mathbb{C}^{(Q-L)\times(Q-L)}$ defines the noise floor along with the orthogonal complement or noise-only subspace spanned by the columns of $\boldsymbol{\mathcal{U}}_\perp(z) : \mathbb{C} \to \mathbb{C}^{Q\times(Q-L)}$.

III. Polynomial Subspace Projection Approach for Voice Activity Detection

A. Polynomial Subspace Projection

Typically, VAD algorithms operate directly on the microphone signals. In the presence of strong interfering signals, however, the performance of these algorithms degrades, as will be investigated in Section IV.

Assuming that the first few frames contain only the interferer components, the space-time covariance matrix in (2) can be estimated using [24], [25]. After computing the PEVD of (2), the orthogonal complement subspace $\boldsymbol{\mathcal{U}}_\perp(z)$ is generated according to (5).

In [14], a syndrome vector is computed by filtering the microphone signals through the eigenvector $\boldsymbol{\mathcal{U}}_\perp(z) \bullet\!\!-\!\!\circ\, \mathbf{U}_\perp(n)$. This syndrome vector is used to detect the entry of a new target source that may be weaker in power than the L interferers, assumed to be stationary for a period of time. The syndrome energy increases in the presence of a new source which is likely to protrude into the subspace $\boldsymbol{\mathcal{U}}_\perp(z)$. Furthermore, since $\mathbf{U}_\perp(n)$ is designed to be causal [26] and may introduce bulk delays to the microphone signals for signal alignment, the syndrome vector may no longer be temporally aligned with the microphone signals. Hence, the syndrome vector cannot be directly used to generate a VAD mask for the microphone signals.

Instead of generating a syndrome vector in [14], a polynomial subspace projection $\boldsymbol{\mathcal{P}}(z) = \boldsymbol{\mathcal{U}}_\perp(z)\boldsymbol{\mathcal{U}}_\perp^P(z) \in \mathbb{C}^{Q\times Q}$ is performed on the microphone signals $\mathbf{x}(n)$ to project them onto a reduced $(Q-L)$ dimensional subspace. This will generate time signals $\mathbf{y}(n)$ with a reduction in energy contributions of the estimated L interferer components using

$$\mathbf{y}(n) = \sum_k \sum_m \mathbf{U}_\perp(k)\, \mathbf{U}_\perp^H(k-m)\, \mathbf{x}(n-m) \; . \tag{6}$$

Note that L is the estimated rank of the interferer components. In general, because of errors incurred in estimating (2) and spectral majorization obtained by using PEVD algorithms such as SBR2 and SMD, leakage occurs across the subspace, i.e., some signal components leak into $\boldsymbol{\mathcal{U}}_\perp(z)$. More notably, in the context of dereverberation [27], the direct-path and early reflections are captured by the subspace associated with the first principal eigenvalue while the late reverberant components are observed in the other subspaces [28]. While an over-estimation of L may be advantageous in minimizing the energy spread of the interferer components, the projection of the target signal onto a lower $(Q-L)$ dimensional subspace may not yield significant components in $\mathbf{y}(n)$.

B. Voice Activity Detection on Projected Component

In order to detect a change point due to an emerging target speaker in the syndrome vector, a VAD algorithm [2] can be applied to the q-th processed signal $y_q(n)$ to generate a more reliable binary mask $m_q(n)$ than the microphone signal $x_q(n)$ which contains some interferer components. The segments containing the target source are then extracted using

$$\hat{s}_q(n) = m_q(n) \cdot y_q(n) \; , \tag{7}$$

where $\hat{s}(n)$ is the estimated target speech in the q-th processed signal, and $m_q(n)$ takes on the value 0 or 1 since it is binary. The proposed method is summarized in Algorithm 1.

IV. Simulation and Results

A. Setup

Measurements of the speech signals and $Q = 8$ channel cafeteria RIRs were taken from the VCTK corpus [29] and

Algorithm 1 Polynomial Subspace Projection-Based VAD.

```
Inputs: $\mathbf{x}(n) \in \mathbb{R}^Q, L.$
    $\mathbf{R}(\tau) \leftarrow E\{\mathbf{x}(n)\mathbf{x}^T(n)\}$ // interferer-only frames, see (2)
    $\boldsymbol{\mathcal{R}}(z) \leftarrow \mathcal{Z}\{\mathbf{R}(\tau)\}$ // see (3)
    $\boldsymbol{\mathcal{U}}(z), \boldsymbol{\Lambda}(z) \leftarrow \text{PEVD } \{\boldsymbol{\mathcal{R}}(z)\}$ // use any [15]–[18]
    $\mathbf{y}(n) \leftarrow \text{project}\{\mathbf{U}_{\perp}(n), \mathbf{x}(n)\}$ // see (6)
    $m_q(n) \leftarrow \text{VAD}\{y_q(n)\}$ // apply VAD [2] on q-th signal
    $\hat{s}_q(n) \leftarrow m_q(n) \cdot x_q(n)$ // extract target activity, see (7)
    return $\hat{s}_q(n)$.
```

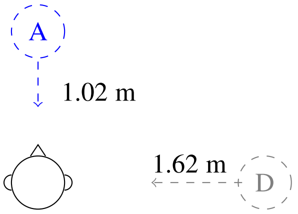


Fig. 1: Experiment setup in the cafeteria from [30].

Kayser database [30], respectively. The interferer signals comprising F16 cockpit and destroyer engine room noise were extracted from the Noisex database [31]. If necessary, signals were resampled to match the sampling rates of 48 kHz. The speech and interferer signals were separately convolved with the RIRs before being added together at each microphone. The source-to-interferences ratio (SIR) [32] at the first microphone, taken to be the reference, was varied from -30 dB to 20 dB. The target speaker and directional interferer are respectively 1.02 m in front (along the y-axis) and 1.62 m to the right (along the x-axis) of the listener, at positions A and D in Fig. 1 [30].

The VAD algorithms used include Sohn's approach [2] and the approach used by WebRTC [13]. WebRTC operates at modes 0–3 from the least to the most aggressive setting. The microphone signals were processed in 30 ms frames. The first 15 frames were assumed to contain only the interferer signals and were therefore, used for calculating (5). We also applied [2] to the projected signal $\mathbf{y}(n)$ to investigate if there is any advantage of pre-processing with (6) using different rank estimates, $L = 1, 2, 5, 7$.

B. Ground Truth Labels

A similar procedure described in [33] 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 anechoic target speech signal is then convolved with the truncated RIR to generate the target speech in $x_1(n)$. The truncation is necessary to ensure that the target speech is time aligned with the microphone signals while minimizing reverberation. The VAD algorithm Mode 3 [13] is applied to the target signal to generate the ground truth VAD labels as shown in Fig. 2(a). For the short target speech used later in Experiment 2 shown in Fig. 2(a)(ii), the positive label '1' at approximately 2.8 s corresponds to a bilabial sound made with both lips [5], as also observed in informal listening examples [34]. In this paper, results for only the first microphone are presented.

C. Evaluation Measures

The counts for the ground truth and predicted labels are tabulated using a confusion matrix [35]. The absence or presence of speech is indicated by the label '0' or '1'. True positive (TP) and true negative (TN) are obtained when both labels are '1' and '0' respectively. False negative (FN) occurs when the predicted label is '0' but the ground truth is '1' while 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 [35]

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FP}}\,, \ \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FN}}\,,$$

$$\text{F1} = \frac{\text{TP}}{\text{TP} + 0.5 \times (\text{FP} + \text{FN})}\,, \ \text{BACC} = \frac{\text{TPR} + \text{TNR}}{2}\,. \tag{8}$$

D. Results and Discussions

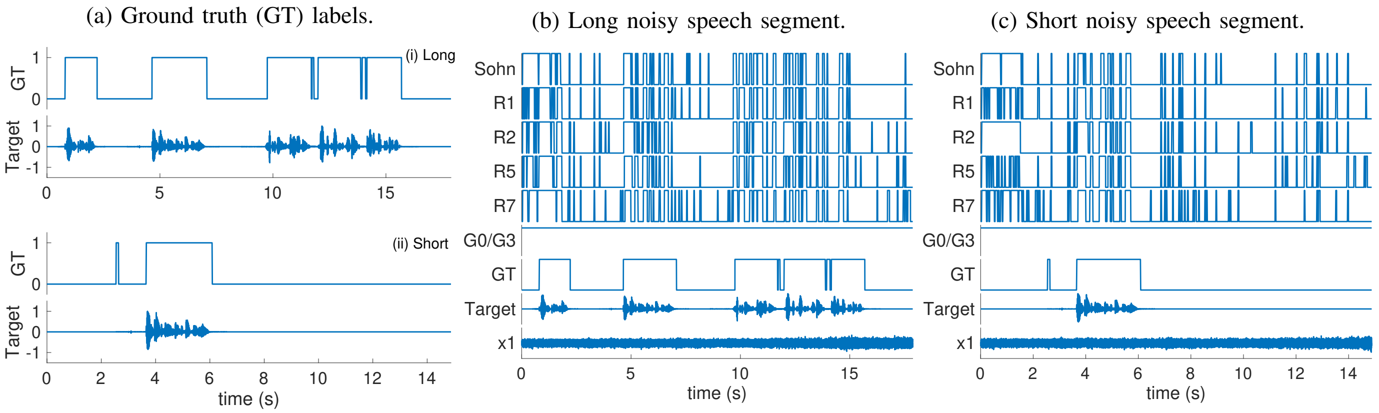
1) Experiment 1: Comparison of VAD on Destroyer Noise. The results are summarized in Table I. At 20 dB SIR, G0 and G3 outperform the other approaches. Slight improvement in F1 and BACC scores arising from an increase in TP is observed when we apply Sohn to the signals projected onto the lower-dimensional subspace (R1, R2, R3, R7) over the microphone signal.

As shown in Table I(b) at 10 dB SIR, the VAD outputs of G0 and G3 are consistently 1, resulting in very high TP and FP. This gives a F1 score of 0.866 but poor BACC score of 0.500 arising from zero negative labels. The proposed approach to perform Sohn [2] on the projected signals shows a slight improvement in F1 score over direct processing on the microphone signal.

At -30 dB SIR where the target signal is significantly weaker, Table I(c) highlights the more significant improvement in the proposed approach over the baseline Sohn. The subspace projection approach increases TP by up to 57 for R7, although this was traded against a drop in TN by 12.

At a high SIR of 20 dB, subspace leakage into the orthogonal complement subspace from the interferer-only subspace is less likely. Hence, R1, R2, R5 and R7 performed similarly. However, at low SIR, e.g. -30 dB, the interferer-only subspace is likely to have leaked into the complement subspace. This promotes high-rank, e.g. R7, so that the microphone signal can be projected into a 1-dimensional subspace where interferer-only components are mostly removed. Note that this small dimensional subspace projection will likely contain only a fraction of the target signal, and hence, the selection of the rank L represents a trade-off.

2) Experiment 2: Different Target Speech Durations. The target speech is corrupted by -20 dB SIR directional F16 cockpit noise. The VAD outputs are shown in Fig. 2(b) for the same long speech segment as Experiment 1. The target signal and the GT labels are shown along with the other VAD outputs. As described in the earlier experiment, the G3 VAD output is


(a) Ground truth (GT) labels.

(b) Long noisy speech segment.

(c) Short noisy speech segment.

Fig. 2: Comparison of VAD binary outputs $m_1(n)$ using Sohn VAD [2] on the microphone signal $x_1(n)$ (Sohn), proposed approach by applying [2] on projected signal $y_1(n)$ using different estimated ranks (R1-R7), WebRTC using modes 0 and 3 (G0, G3) [13]. The plots show (a) the ground truth (GT) labels for (i) long and (ii) short target signal component in $x_1(n)$; (b) long noisy and (c) short noisy segments of speech corrupted by -20 dB SIR F16 cockpit noise from Noisex database [31].

TABLE I: Confusion matrix and scores for VAD output on target speech in directional destroyer noise at various SIR.

(a) SIR = 20 dB

Method	TP	TN	FP	FN	F1	BACC
Sohn	283	175	104	32	0.806	0.763
R1	287	174	105	28	0.812	0.767
R2	286	175	104	29	0.811	0.768
R5	287	173	106	28	0.811	0.766
R7	291	171	108	24	0.815	0.768
G0	311	249	30	4	0.948	0.940
G3	293	273	6	22	**0.954**	**0.954**

(b) SIR = 10 dB

Method	TP	TN	FP	FN	F1	BACC
Sohn	271	238	41	44	0.864	**0.857**
R1	275	233	46	40	0.865	0.854
R2	275	224	55	40	0.853	0.838
R5	280	227	52	35	**0.866**	0.851
R7	277	231	48	38	**0.866**	0.854
G0	315	0	279	0	0.693	0.500
G3	315	0	279	0	0.693	0.500

(c) SIR= -30 dB

Method	TP	TN	FP	FN	F1	BACC
Sohn	94	226	53	221	0.407	0.628
R1	111	227	52	204	0.464	0.651
R2	102	235	44	213	0.443	0.642
R5	138	226	53	177	0.545	0.688
R7	151	214	65	164	0.569	**0.699**
G0	315	0	279	0	**0.693**	0.500
G3	315	0	279	0	**0.693**	0.500

TABLE II: Confusion matrix and scores for VAD output on long target speech in directional F16 noise at -20 dB SIR.

Method	TP	TN	FP	FN	F1	BACC
Sohn	130	241	38	185	0.538	0.638
R1	136	249	30	179	0.565	0.662
R2	158	244	35	157	0.622	**0.688**
R5	148	247	32	167	0.598	0.678
R7	136	224	55	179	0.538	0.617
G0	315	0	279	0	**0.693**	0.500
G3	315	0	279	0	**0.693**	0.500

TABLE III: Confusion matrix and scores for VAD output on short target speech in directional F16 noise at -20 dB SIR.

Method	TP	TN	FP	FN	F1	BACC
Sohn	28	334	76	56	0.298	0.574
R1	30	342	68	54	0.330	0.596
R2	45	349	61	39	**0.474**	**0.693**
R5	32	344	66	52	0.352	0.610
R7	32	325	85	52	0.318	0.587
G0	84	0	410	0	0.291	0.500
G3	84	0	410	0	0.291	0.500

always 1, which implies that it always predicts the presence of speech. This results in a high TP and, subsequently, good F1 score but is penalized by the poor BACC score arising from high FP, as shown in Table II.

When the target speech segment is short, as shown in Fig. 2(c), the G0 and G3 VAD outputs are also always 1. However, this time, the FP tremendously increases to 410 and this severely affects the F1 score. The proposed approach demonstrates that pre-processing the microphone with the subspace projection almost always improves the F1 and BACC scores. In this case, R2 provides an improvement over [2] in F1 and BACC scores by 0.176 and 0.119, respectively.

V. Conclusion

In this work, a polynomial subspace projection approach has been proposed as a pre-processor to improve VAD performance. We have shown that performing this multi-channel pre-processor prior to applying the single-channel Sohn VAD algorithm [2] almost always improves the F1 and balanced accuracy (BACC) scores even in adverse environments, e.g., -30 dB SIR. This improvement over the baseline of applying VAD to the microphone signal is less significant at high SIRs and more significant at low SIRs. Note that it is particularly in the low SIR regime, i.e., for weak speaker signals, where we set out to boost VAD performance. We have also shown that the rank estimate of the interferer-only subspace directly impacts the orthogonal complement subspace used for the projection and, subsequently, the VAD performance. Informal listening examples are available [34].

References

[1] 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.

[2] 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.

[3] M. W. Mak and H. B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluation," *Comput. Speech and Language*, vol. 28, no. 1, pp. 295–313, 2014.

[4] 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.

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

[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] 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.

[11] 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.

[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] 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.

[15] 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.

[16] 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.

[17] V. W. Neo, C. Evers, and P. A. Naylor, "Speech enhancement using polynomial eigenvalue decomposition," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, Oct. 2019.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[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] 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.

[25] 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.

[26] J. Corr, K. Thompson, S. Weiss, J. G. McWhirter, and I. K. Proudler, "Causality-constrained multiple shift sequential matrix diagonalisation for parahermitian matrices," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2014, pp. 1277–1281.

[27] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer-Verlag, 2010.

[28] 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.

[29] 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.

[30] 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.

[31] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 3, no. 3, pp. 247–251, Jul. 1993.

[32] 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.

[33] 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.

[34] V. W. Neo, "VAD exploiting a polynomial subspace projection approach," Apr. 2022. [Online]. Available: https://vwn09.github.io/research/pevd-vad

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