This article has been accepted for publication in IEEE/ACM Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2021.3120630

Enhancement of Noisy Reverberant Speech Using Polynomial Matrix Eigenvalue Decomposition

Vincent W. Neo, *Student Member, IEEE,* Christine Evers, *Senior Member, IEEE* and Patrick A. Naylor, *Fellow, IEEE*

***Abstract*—Speech enhancement is important for applications such as telecommunications, hearing aids, automatic speech recognition and voice-controlled systems. Enhancement algorithms aim to reduce interfering noise and reverberation while minimizing any speech distortion. In this work for speech enhancement, we propose to use polynomial matrices to model the spatial, spectral and temporal correlations between the speech signals received by a microphone array and polynomial matrix eigenvalue decomposition (PEVD) to decorrelate in space, time and frequency simultaneously. We then propose a blind and unsupervised PEVD-based speech enhancement algorithm. Simulations and informal listening examples involving diverse reverberant and noisy environments have shown that our method can jointly suppress noise and reverberation, thereby achieving speech enhancement without introducing processing artefacts into the enhanced signal.**

***Index Terms*—Broadband multi-channel signal processing, noise reduction, polynomial matrix eigenvalue decomposition, speech dereverberation, speech enhancement.**

I. Introduction

THE enhancement of degraded speech signals remains important in many applications ranging from human-to-human communications in telecommunications and hearing aids [1]–[3] to human-to-machine interaction in automatic speech recognition, voice-controlled systems and robot audition [4], [5]. The main causes of degradation are additive background noise and reverberation due to multi-path reflections in enclosed spaces [6]. Consequently, the speech signal becomes temporally smeared and contaminated by interfering signals. Furthermore, in all these applications, prior information of the target speech or the acoustic environment is not available, motivating the need for a blind, or unsupervised, approach.

The speech enhancement task which includes noise reduction [7] and dereverberation [8] will be addressed in this paper. Existing enhancement algorithms may typically distort the speech signal and introduce processing artefacts [9]–[11]. Methods for controlling the trade-off between noise reduction against speech distortions are introduced in [12]–[14]. For instance, aggressive noise reduction might be preferred for human-to-machine applications while mobile phone and hearing aid users might prefer less speech distortion at the expense of higher residual noise.

Existing noise reduction techniques can be classified into single- and multi-channel techniques. Methods for single-channel noise reduction include spectral subtraction [15], [16], statistical-based and subspace-based approaches. Statistical methods are typically based on minimizing the mean-square error (MMSE) of the clean and estimated speech spectrum [12], the log-spectrum (log-MMSE) [17] or the single-channel Wiener filter [13], [18].

In subspace methods, noisy signals are decomposed into signal and noise subspaces and enhancement is achieved by recovering the speech signal from the signal subspace. The Karhunen-Loève transform (KLT) was used in [19] and an optimal solution was derived for white noise. This work was extended in [20] to cope with coloured noise by using a generalized eigenvalue decomposition (GEVD) to jointly diagonalize the speech and noise covariance matrices. Subspace-based methods have also been extended to multi-channel systems [21], [22] but they do not fully exploit spatial information to minimize speech distortion [7], as will be seen in Section VI.

Multi-microphone methods include beamformers [23]–[25] with optional post-filtering [26], [27], the optimal multi-channel Wiener filter (MWF) [13], [18], [28] and the multi-channel Kalman filter [29], [30]. While existing multi-channel approaches may potentially achieve noise reduction without speech distortion when the noise is spatially and temporally white, this remains a practical challenge under other noise conditions in real-world scenarios [7].

Speech dereverberation approaches can be classified into speech synthesis-based, reverberation cancellation and reverberation suppression methods. In synthesis-based methods, the linear prediction coding (LPC) residual is directly computed to generate an estimated clean speech signal [31], [32]. This approach, however, is usually limited to mildly reverberant signals in order to avoid introducing artefacts [33].

In one class of reverberation cancellation approaches, the acoustic channel is first estimated using, for example, blind system identification [34]–[37]. The estimated channel is then used in the design of inverse filter(s), such as [38] for a single-channel system and [39] for a multi-channel system based on the multi-channel inversion theorem (MINT). Because MINT is not robust to estimation errors, channel shortening techniques have been proposed [40]–[42]. In another class of cancellation methods, multi-channel linear prediction, such as the weighted prediction error (WPE) [43]–[45], is used to

This manuscript is first submitted on 29 Jan. 2021; revised 19 July 2021 and 27 September 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zheng-Hua Tan. This work is supported by the Engineering and Physical Sciences Research Council (EPSRC) grant no. EP/S035842/1 and the EPSRC Fellowship grant no. EP/P001017/1. (Corresponding author: Vincent W. Neo.)

Vincent W. Neo and Patrick A. Naylor are with the Department of Electrical and Electronic Engineering, Imperial College London, UK, (email: {vincent.neo09, p.naylor}@imperial.ac.uk).

Christine Evers is with the School of Electronic and Computer Science, University of Southampton, UK, (email: c.evers@soton.ac.uk).

directly estimate the dereverberated speech signal.

The reverberation suppression methods include single-channel spectral subtraction [46] and multi-channel spatial filtering techniques such as the MWF [13], [18], which may be used in conjunction with post-filtering techniques [27]. Because these methods use noise and/or transfer function estimators, the dereverberation effectiveness depends heavily on the performance of these estimators.

A large number of contributions are based on the assumption of narrowband signal models, e.g., [19]–[22]. Single-channel approaches exploit temporal correlations while multi-channel approaches capture spatio-temporal correlations. However, extensions to broadband signals using the short-time Fourier transform (STFT) ignores the correlations between frequency bands and cannot preserve phase coherence across bands [47]. Furthermore, the microphone outputs are temporally correlated because of the speech source signal and reverberation. Consequently, an eigenvalue decomposition (EVD) or KLT, which removes correlations at a single time lag, is inadequate in decorrelating the signals completely. Polynomial matrices can simultaneously capture the correlations in space, time and frequency and are, therefore, appropriate for modelling multi-channel broadband signals.

The processing of polynomial matrices has motivated the development of a family of polynomial matrix eigenvalue decomposition (PEVD) algorithms [48]–[50], which are based on the second-order sequential best rotation (SBR2) [51]. Unlike EVD or KLT, PEVD can achieve decorrelation over a suitably chosen range of time lags and is more suitable for broadband signals. It is also widely used in many multi-channel broadband signal processing applications such as blind source separation [52], source identification [53], localization [54], adaptive beamforming [55] and channel coding [56].

In this work, we introduce and extend PEVD to the field of speech signal processing and propose a blind and unsupervised PEVD-based speech enhancement algorithm. Preliminary studies separately focused on the task of noise reduction [57], [58] and speech dereverberation [59]. In contrast, in this paper, we present PEVD-based speech enhancement for both noise reduction and dereverberation. In addition, we further investigate the impact of the parameter settings of the proposed algorithm and provide a thorough performance evaluation in different acoustic conditions, benchmarked against state-of-the-art baseline approaches for speech enhancement. Therefore, in supplement to the earlier studies, the novel contributions of this paper are (i) the use of a polynomial matrix as a broadband, multi-channel signal model for speech, (ii) presentation of a novel PEVD-based speech enhancement algorithm, (iii) the investigation of the parameters of the proposed algorithm, (iv) a comprehensive evaluation and analysis of the proposed approach for realistic signals under wide-range of noise and reverberation conditions, and (v) an evaluation of our PEVD method against several comparative approaches.

II. Speech Enhancement Problem Formulation

The noisy and reverberant signal, $x_m(n)$, at the m-th microphone for discrete-time sample $n = 0, 1, \ldots, N$ is

$$x_m(n) = \mathbf{h}_m^T \mathbf{s}_0(n) + v_m(n) \tag{1}$$

where $\mathbf{s}_0(n) = \big[s_0(n), \ldots, s_0(n-J)\big]^T$, is the anechoic speech, $\mathbf{h}_m$ is the acoustic channel impulse response from the source to the m-th microphone, assumed to be stationary and modelled using a J-th order finite impulse response (FIR) filter, $v_m(n)$ represents the additive noise at the m-th microphone and $[\cdot]^T$ denotes the transpose operator. The noise signals are assumed to be zero-mean, not perfectly coherent with each other and uncorrelated with the source signal [7].

Using the reverberation model in [6], the early reflections represent closely spaced distinct echoes that perceptually reinforce the direct-path component and may improve speech intelligibility in certain conditions [60]. The late reflections comprise randomly distributed small amplitude components, which are commonly assumed uncorrelated with the direct-path and early reflections, and can be treated as an additive, uncorrelated noise component [61]. Consequently, (1) becomes

$$\begin{aligned} x_m(n) &= \mathbf{h}_{m,d}^T \mathbf{s}_0(n) + \mathbf{h}_{m,e}^T \mathbf{s}_0(n) + x_{m,l}(n) + v_m(n) \\ &= \tilde{s}_m(n) + \tilde{v}_m(n), \end{aligned} \tag{2}$$

where $\mathbf{h}_{m,d}$ and $\mathbf{h}_{m,e}$ are the impulse responses associated with the direct-path and early reflections, $x_{m,l}(n)$ is the late reverberant component, $\tilde{s}_m(n) = \mathbf{h}_{m,d}^T \mathbf{s}_0(n) + \mathbf{h}_{m,e}^T \mathbf{s}_0(n)$ and $\tilde{v}_m(n) = x_{m,l}(n) + v_m(n)$ are the desired speech and unwanted noise at the m-th microphone. The data vector from M microphones is $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$, with $\mathbf{s}(n)$, $\tilde{\mathbf{s}}(n)$, $\mathbf{v}(n)$ and $\tilde{\mathbf{v}}(n)$ similarly defined.

A. Noise Reduction in Anechoic Environment

Without reverberation, each acoustic channel only comprises the direct-path propagation, modelled using a delay. Typically, the first channel, or the channel with the shortest propagation time from the source, can be taken as the reference. Consequently, (1) simplifies to

$$x_m(n) = s_m(n) + v_m(n), \tag{3}$$

where $s_m(n)$ is an attenuated and delayed version of $s_0(n)$ at the m-th microphone. The goal of noise reduction is to recover $\mathbf{s}(n)$ from $\mathbf{x}(n)$ while keeping $\mathbf{v}(n)$ suppressed.

B. Speech Dereverberation in Noiseless Environment

In the absence of additive noise, (1) simplifies to

$$x_m(n) = \mathbf{h}_m^T \mathbf{s}_0(n), \tag{4}$$

since $v_m(n) = 0$. While retaining the early reflections is not normally perceived to be perceptually harmful, in this paper, the evaluation of dereverberation targets the anechoic speech, $\mathbf{s}_0(n)$. Furthermore, if noise is present in the environment, residual additive noise after processing may remain but is not considered in the dereverberation evaluation.

III. Polynomial Matrix Decomposition

A. Motivation for Polynomial Matrices

Broadband signals received by microphone arrays exhibit spatial, spectral and temporal correlations. The space-time covariance matrix is defined as [51]

$$\mathbf{R}_{\mathbf{xx}}(\tau) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^H(n-\tau)\}, \tag{5}$$

where $\mathbb{E}\{\cdot\}$ and $[\cdot]^H$ are the expectation and Hermitian transpose, respectively. The $(p,q)^{\text{th}}$ element of $\mathbf{R}_{\mathbf{xx}}(\tau)$, $r_{p,q}(\tau) = \mathbb{E}\{x_p(n)x_q^*(n-\tau)\}$, is computed using the correlation function between the received signal at the p-th and q-th microphone and is parameterized by the temporal lag, τ, and $[\cdot]^*$ is the complex conjugate operator.

If the source, $s_0(n)$, is a narrowband signal at frequency f_0 propagating in an anechoic environment, $x_m(n)$ is related to the reference signal at the first microphone, $x_1(n)$, by a constant phase shift, $\phi_m = 2\pi f_0 \tau_m$. The data vector becomes $\mathbf{x}(n) = [x_1(n), x_1(n)e^{-j\phi_2}, \ldots, x_1(n)e^{-j\phi_M}]^T$, such that the correlation between the $(p,q)^{\text{th}}$ sensor pair is

$$\begin{aligned} r_{p,q}(\tau) &= \mathbb{E}\{x_1(n)e^{-j\phi_p}[x_1(n-\tau)e^{-j\phi_q}]^*\} \\ &= \mathbb{E}\{x_1(n)x_1(n)\}e^{-j\phi_{p,q}+j2\pi f_0\tau}, \end{aligned} \tag{6}$$

where $\phi_{p,q} = \phi_p - \phi_q$ is the phase difference that depends only on the array geometry and is associated with the time difference of arrival (TDoA). Since the expectation term in (6) is independent of τ and the array geometry is fixed, $\phi_{p,q}$ remains constant across all lags and (6) can be computed at any τ for decorrelation using an EVD. Classical subspace-based approaches for the enhancement of narrowband signals compute the instantaneous spatial covariance matrix by evaluating (5) at $\tau = 0$, expressed as

$$\mathbf{R}_{\mathbf{xx}}(0) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}. \tag{7}$$

However, for a broadband source signal like speech, the expectation term in (6) no longer holds since the source correlation now depends on τ. Therefore, correlations across different sensors and temporal lags need to be considered. Accordingly, concatenating the covariance matrix, $\mathbf{R}_{\mathbf{xx}}(\tau)$, for all values of $\tau \in \{-N, \ldots, N\}$, results in a tensor of dimension, $M \times M \times (2N+1)$.

Instead of processing signals in the STFT domain, the z-transform which captures and preserves the correlations of the received signals in space, time and frequency is used. The z-transform of (5) is a para-Hermitian polynomial matrix [51]

$$\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) = \sum_{\tau=-\infty}^{\infty} \mathbf{R}_{\mathbf{xx}}(\tau)z^{-\tau}. \tag{8}$$

The polynomial matrix is a matrix with polynomial elements, or equivalently, a polynomial with matrix coefficients. In the former, each element in the polynomial matrix represents the correlation function in z between a specific microphone pair. In the latter, the polynomial matrix shows how the spatial correlation between all sensor pairs changes with z.

B. Family of PEVD Algorithms

The PEVD of a para-Hermitian polynomial matrix is [51]

$$\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) \approx \boldsymbol{\mathcal{U}}^P(z)\boldsymbol{\Lambda}(z)\boldsymbol{\mathcal{U}}(z), \tag{9}$$

where the rows of $\boldsymbol{\mathcal{U}}(z)$ are the eigenvectors and the diagonal polynomial matrix, $\boldsymbol{\Lambda}(z)$ contains the eigenvalues, and $[\cdot]^P$ is the para-Hermitian operator such that $\boldsymbol{\mathcal{U}}^P(z) = \boldsymbol{\mathcal{U}}^H(1/z^*)$. The prefix 'para' indicates a time-reversal. The approximation in (9) arises because an exact PEVD is only possible with infinite polynomial order, as described in [51]. Although exact diagonalization of $\boldsymbol{\Lambda}(z)$ is unachievable using only FIR filters, diagonalization can be attained to a good approximation for sufficiently large order of filters or polynomial elements [62].

The PEVD can be computed using an iterative algorithm [48]–[51] based on similarity transforms involving L para-unitary polynomial matrices [63], $\boldsymbol{\mathcal{U}}(z) = \boldsymbol{\mathcal{U}}_L(z) \ \ldots \ \boldsymbol{\mathcal{U}}_1(z)$. At the ℓ-th iteration, the algorithm first searches for the largest off-diagonal element exceeding a predefined threshold, δ. $\boldsymbol{\mathcal{U}}_\ell(z)$ is then constructed using delay polynomial matrices and unitary matrices, which are designed to zero out the off-diagonal elements on the matrix coefficient of z^0, and applied to the entire polynomial matrix. To keep the polynomial order compact, a fraction μ, of the total Frobenius-norm squared, is truncated as detailed in [51]. After L iterations, $\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z)$ is approximately diagonalized according to [64]

$$\boldsymbol{\Lambda}(z) \approx \boldsymbol{\mathcal{U}}(z)\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z)\boldsymbol{\mathcal{U}}^P(z) = \boldsymbol{\mathcal{U}}(z)\mathbb{E}\{\mathbf{x}(z)\mathbf{x}^P(z)\}\boldsymbol{\mathcal{U}}^P(z), \tag{10}$$

where $\mathbf{x}(z)$ is the z-transform of $\mathbf{x}(n)$ based on (8).

The zeroing unitary matrix computed at each iteration can take the form of a Givens rotation in SBR2 [51], a Householder-like optimization procedure as in [49], a combination of Householder reflection and Givens rotation matrices in [50] or an eigenvector matrix in the sequential matrix diagonalization (SMD) algorithm [48].

IV. PEVD-based Speech Enhancement

The z-transform of the space-time covariance matrix of the microphone signals obtained by applying (2) to (5) is

$$\begin{aligned} \boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) &= \mathbf{H}_d^T\boldsymbol{\mathcal{R}}_{\mathbf{ss}}(z)\mathbf{H}_d + \mathbf{H}_e^T\boldsymbol{\mathcal{R}}_{\mathbf{ss}}(z)\mathbf{H}_e \\ &\quad + \boldsymbol{\mathcal{R}}_{\boldsymbol{\mathcal{U}}}(z) + \boldsymbol{\mathcal{R}}_{\mathbf{vv}}(z) \\ &= \boldsymbol{\mathcal{R}}_{\tilde{\mathbf{s}}\tilde{\mathbf{s}}}(z) + \boldsymbol{\mathcal{R}}_{\tilde{\mathbf{v}}\tilde{\mathbf{v}}}(z), \end{aligned} \tag{11}$$

where $\boldsymbol{\mathcal{R}}_{\mathbf{ss}}(z)$, $\boldsymbol{\mathcal{R}}_{\mathbf{vv}}(z)$ and $\boldsymbol{\mathcal{R}}_{\boldsymbol{\mathcal{U}}}(z)$ are respectively, the space-time covariance polynomial matrices of the anechoic speech, noise and late reverberation modelled as a spatially diffuse field [61], and $\mathbf{H}_d = [\mathbf{h}_{1,d}, \ldots, \mathbf{h}_{M,d}]$, with $\mathbf{H}_e$ similarly defined. $\boldsymbol{\mathcal{R}}_{\tilde{\mathbf{s}}\tilde{\mathbf{s}}}(z)$, which is obtained by grouping the direct-path and early reflections, is uncorrelated with $\boldsymbol{\mathcal{R}}_{\tilde{\mathbf{v}}\tilde{\mathbf{v}}}(z)$, that includes the late reflections and additive noise components. Assuming stationarity within each processing frame, (5) is estimated using $T+1$ samples per frame according to

$$\hat{\mathbf{R}}_{\mathbf{xx}}(\tau) \approx \frac{1}{T+1}\sum_{n=0}^{T}\mathbf{x}(n)\mathbf{x}^H(n-\tau). \tag{12}$$

Furthermore, (8) can be approximated using

$$\hat{\boldsymbol{\mathcal{R}}}_{\mathbf{xx}}(z) \approx \sum_{\tau=-W}^{W}\mathbf{R}_{\mathbf{xx}}(\tau)z^{-\tau}, \tag{13}$$

where W is the truncation window which reflects the extent of temporal correlation of the speech signals. Hence, in addition to the PEVD parameters, the frame size, T, and window length, W, can affect the performance of the proposed

algorithm as investigated in Section VI-A. Since noise and speech are assumed uncorrelated, the PEVD gives [58]

$$\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) \approx \left[\boldsymbol{\mathcal{U}}^P_{\tilde{s}}(z) \mid \boldsymbol{\mathcal{U}}^P_{\tilde{v}}(z)\right] \left[\begin{array}{c|c} \boldsymbol{\Lambda}_{\tilde{s}}(z) & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\Lambda}_{\tilde{v}}(z) \end{array}\right] \left[\begin{array}{c} \boldsymbol{\mathcal{U}}_{\tilde{s}}(z) \\ \hline \boldsymbol{\mathcal{U}}_{\tilde{v}}(z) \end{array}\right], \tag{14}$$

where $\{.\}_{\tilde{s}}$ and $\{.\}_{\tilde{v}}$ are associated with the signal-plus-noise (or simply signal) and noise-only (or simply noise) subspaces.

The eigenvector polynomial matrix, $\boldsymbol{\mathcal{U}}(z)$, can be applied as a filterbank for $\mathbf{x}(z)$ so that the outputs, $\mathbf{y}(z) = \boldsymbol{\mathcal{U}}(z)\mathbf{x}(z)$, are strongly decorrelated [48], according to

$$\mathbb{E}\{\mathbf{y}(z)\mathbf{y}^P(z)\} = \mathbb{E}\{\boldsymbol{\mathcal{U}}(z)\mathbf{x}(z)\mathbf{x}^P(z)\boldsymbol{\mathcal{U}}^P(z)\} \approx \boldsymbol{\Lambda}(z). \tag{15}$$

Unlike some speech enhancement approaches, the proposed method does not rely on noise estimation since the strong decorrelation property of PEVD implicitly orthogonalizes the subspaces across all time lags in the range of W. Furthermore, $\boldsymbol{\mathcal{U}}(z)$ is lossless or para-unitary by construction and has an all-pass filter frequency response [63]. This implies that $\boldsymbol{\mathcal{U}}(z)$ can only redistribute spectral power among channels and not change the total (over all subspaces) signal and noise power [51]. The PEVD algorithms in [51] also tend to sort the strongly decorrelated outputs in descending order of signal energy because of the spectral majorization property [48]. The signal subspace comprises mostly speech components, originally distributed over all microphones but now summed coherently. In contrast, the noise subspace is dominated by ambient noise and late reflections in the reverberant channels. Consequently, speech enhancement is achieved by combining components in the signal subspace, defined as the first channel with the largest total spectral power, and nulling components in the noise subspace, e.g., $\boldsymbol{\mathcal{U}}_{\tilde{v}}(z) = 0$. The PEVD-based speech enhancement algorithm is summarized in Algorithm 1.

A. Noise Reduction in Anechoic Environment Case

In an anechoic environment, (3) simplifies (11) to

$$\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) = \mathbf{H}_d^T \boldsymbol{\mathcal{R}}_{\mathbf{ss}}(z)\mathbf{H}_d + \boldsymbol{\mathcal{R}}_{\mathbf{vv}}(z), \tag{16}$$

where the $(p,q)^{\text{th}}$ element can be written as

$$\begin{aligned} r_{p,q}(\tau) &= s_p(n)s_q(n-\tau) + v_p(n)v_q(n-\tau) \\ &= s_0(n-\tau_p)s_0(n-\tau_q-\tau) + v_p(n)v_q(n-\tau), \end{aligned} \tag{17}$$

which implies that the difference in lags between every channel pair can only be captured by PEVD in the range of W. Then, PEVD decomposes the signal information into subspaces associated with the anechoic speech and the noise signals based on (14).

B. Speech Dereverberation in Noiseless Environment Case

When the environment is noiseless, (4) simplifies (11) to

$$\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) = \mathbf{H}_d^T \boldsymbol{\mathcal{R}}_{\mathbf{ss}}(z)\mathbf{H}_d + \mathbf{H}_e^T \boldsymbol{\mathcal{R}}_{\mathbf{ss}}(z)\mathbf{H}_e + \boldsymbol{\mathcal{R}}_{\boldsymbol{ll}}(z) \tag{18}$$

Comparing with (11), the first two terms are part of the desired speech subspace. They are uncorrelated and orthogonal with the last term, which contains only the late components and forms the noise subspace. The window size, W, should be large enough to include the direct-path and early reflections in order to achieve the desired enhancement.

Algorithm 1 PEVD-based speech enhancement [58].

Inputs: $\mathbf{x}(n) \in \mathbb{R}^M, n \in \{0, \ldots, T\}, W, \delta, \mu, L$.
$\mathbf{R}_{\mathbf{xx}}(\tau) \leftarrow E\{\mathbf{x}(n)\mathbf{x}^T(n-\tau)\}$ *// see* (5)
$\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z) \leftarrow \mathcal{Z}\{\mathbf{R}_{\mathbf{xx}}(\tau)\}$ *// see* (8)
$\boldsymbol{\mathcal{U}}(z), \boldsymbol{\Lambda}(z) \leftarrow$ PEVD $\{\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z), \delta, \mu, L\}$ *// any [48]–[51]*
$\mathbf{x}(z) \leftarrow \mathcal{Z}\{\mathbf{x}(n)\}$ *// see* (8)
$\mathbf{y}(z) \leftarrow \boldsymbol{\mathcal{U}}(z)\mathbf{x}(z)$ *// speech enhancement*
$\mathbf{y}_1(z) \leftarrow \mathbf{y}(z)$ *// enhanced signal in the first channel*
return $\mathbf{y}_1(z)$.

V. Experimental Setup

A. Acoustic Environments and Setup

Anechoic speech signals, which are sampled at 16 kHz, are taken from TIMIT corpus [65]. Room impulse response measurements and noise recordings are taken from the complete ACE corpus [66]. The T_{60} of the rooms in the ACE corpus range from 0.332 s to 1.22 s. The 3-channel 'mobile' array was used in most experiments. The 8-channel linear array, with microphones spaced by 60 mm, was used for the study on the use of a different number of microphones. In the direct-path-only experiments, the propagation delays were drawn from the discrete uniform distribution, $U(1, 1000)$, and ordered such that $\tau_1 < \tau_2 < \tau_3$.

Babble, car and factory noise from the Noisex database [67], as well as restaurant, residential traffic and city street noise from the International Sound Effects (SoundFx) library [68], and white noise, were used.

The noise recordings in ACE were used directly. With the Noisex and SoundFx noise signals, diffuse noise signals were produced using [69]. In each trial, sentences from a randomly selected speaker were concatenated to have 8 to 10 s duration. The anechoic speech signals were then convolved with the impulse responses at each microphone channel before being corrupted by additive noise using [70], implemented in [71]. The signal to noise ratio (SNR) ranged from -10 dB to 20 dB. For each Monte-Carlo simulation, 50 trials were conducted.

B. Speech Enhancement Comparative Algorithms

The proposed PEVD method was compared against two versions of the MWF, two subspace approaches, the generalized weighted prediction error (GWPE), and two joint approaches that can suppress both noise and reverberation. Both MWFs are based on the concatenation of a minimum variance distortionless response (MVDR) beamformer followed by a single-channel Wiener filter [27]. The first is a practical MWF which uses a relative transfer function (RTF)-based speech estimator and a noise estimator based on the parameters used in [29]. The second is the Oracle MWF (OMWF) which provides an ideal performance upper-bound since it uses complete prior knowledge of the clean speech signal, based on [13] where the filter length is 80. The PEVD subspace approach was also compared against the subspace method for coloured noise (COLSUB) which uses a GEVD [20] and the multi-channel subspace (MCSUB) method [21]. The GWPE dereverberation algorithm [43], and integrated methods for noise reduction and dereverberation such as the weighted power minimization

This article has been accepted for publication in IEEE/ACM Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2021.3120630

distortionless response (WPD) [45] and the integrated sidelobe cancellation and linear prediction (ISCLP) Kalman filter [30], were also included as benchmark approaches. The published code was used for ISCLP while the published parameters and ground truth direction of arrivals (DoAs) from ACE were used to compute the steering vectors for WPD [45] to avoid signal direction mismatch errors.

For all experiments, the PEVD parameters, chosen following [57]–[59], were $\delta = \sqrt{N_1/3} \times 10^{-2}$ where N_1 is the square of the trace-norm of $\mathbf{R_{xx}}(0)$, $\mu = 10^{-3}$ and $L = 500$. In all experiments, except for those investigating the effects of varying T and W, $T = W = 1600$ samples were used. With this parameter selection, correlations within 100 ms, which were assumed to include the direct-path and early reflection components, were captured and used by the algorithm. The source code for our method and experiments is available [72].

C. Evaluation Measures

For the noise reduction evaluation, the segmental signal to noise ratio (SegSNR) and frequency-weighted SegSNR (FwSegSNR) [73] are used. To measure the speech quality and intelligibility and to account for processing artefacts, short-time objective intelligibility (STOI) [74] and perceptual evaluation of speech quality (PESQ) [75] are used. A key measure of dereverberation is the direct-to-reverberant ratio (DRR) but the modified impulse responses after processing are generally unavailable. Instead, the normalized signal-to-reverberant ratio (NSRR), which is a signal-based measure shown to be equivalent to DRR under certain conditions [76], and the Bark spectral distortion (BSD) [6], are used.

These measures are computed for the signals before and after enhancement using the proposed and benchmark algorithms and the improvement Δ is reported. Positive Δ values show improvements in all measures except ΔBSD, for which a negative value indicates a reduction in spectral distortions.

VI. Results and Discussions

Monte-Carlo simulations have been conducted to understand 1) the impact of varying the parameters of the proposed PEVD-based approach and; 2) the effectiveness of the proposed approach in comparison to other methods for speech enhancement, including the specific cases of noise reduction in anechoic environments and dereverberation in noiseless environments. Listening examples are also available [72].

A. Parameter Selection for the PEVD-based Algorithm

These parameter values have also been confirmed to exhibit similar trends for different scenarios including noise types and rooms. However, due to space limitations, we present in this paper only the results for white noise in Lecture Room 2.

1) Frame Size and Window Length: With a fixed frame size, $T = 1600$ samples, the window length W was varied from 0 to 1600 samples. When $W = 0$, PEVD is equivalent to an EVD applied to the instantaneous spatial covariance matrix in (7), and this can only decorrelate the microphone signals at a single time lag. As W increases, correlations between microphones at other lags are computed and used in the PEVD

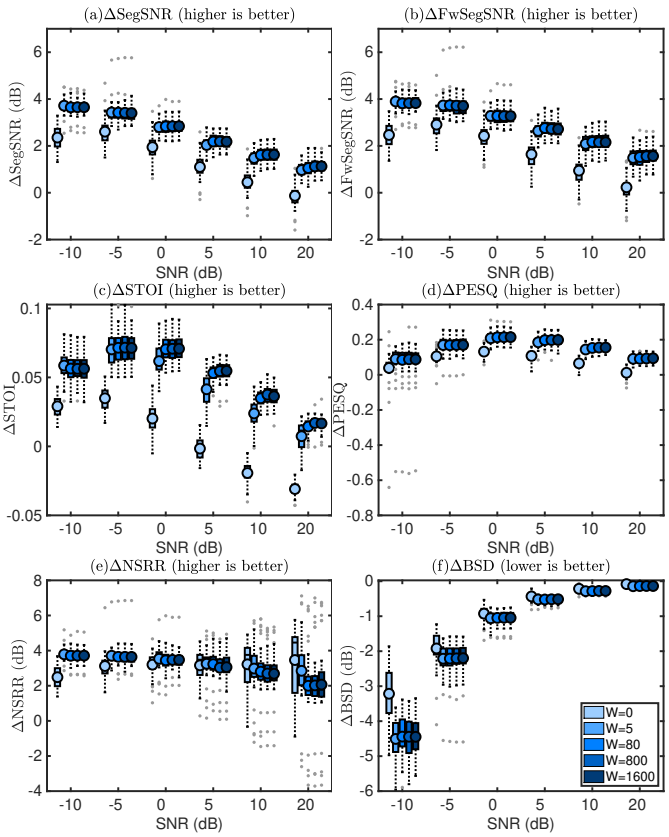

Fig. 1. Speech enhancement for white noise in Lecture Room 2 using different window length W for the approximation of the z-transform.

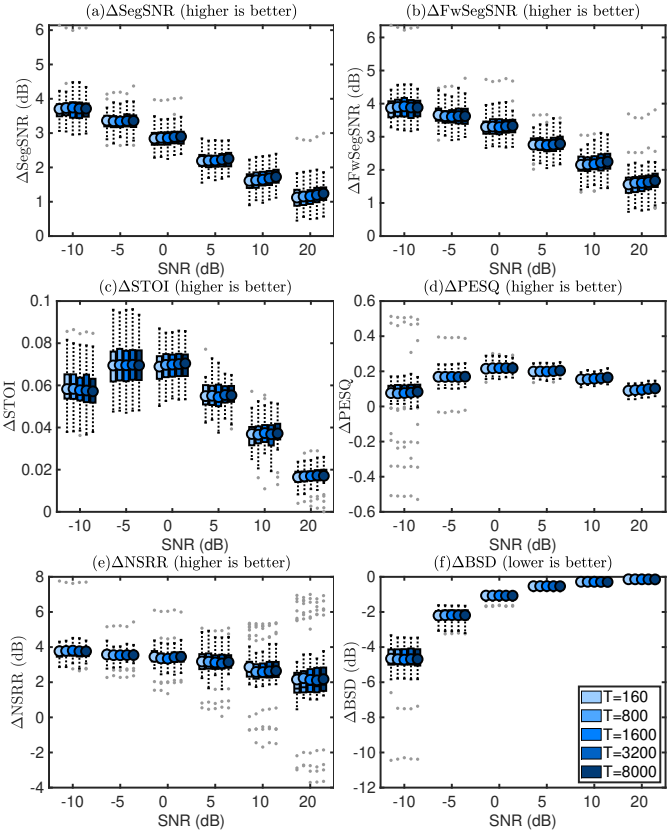

Fig. 2. Speech enhancement for white noise in Lecture Room 2 using different frame length T for the computation of the space-time covariance matrix.

This article has been accepted for publication in IEEE/ACM Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2021.3120630

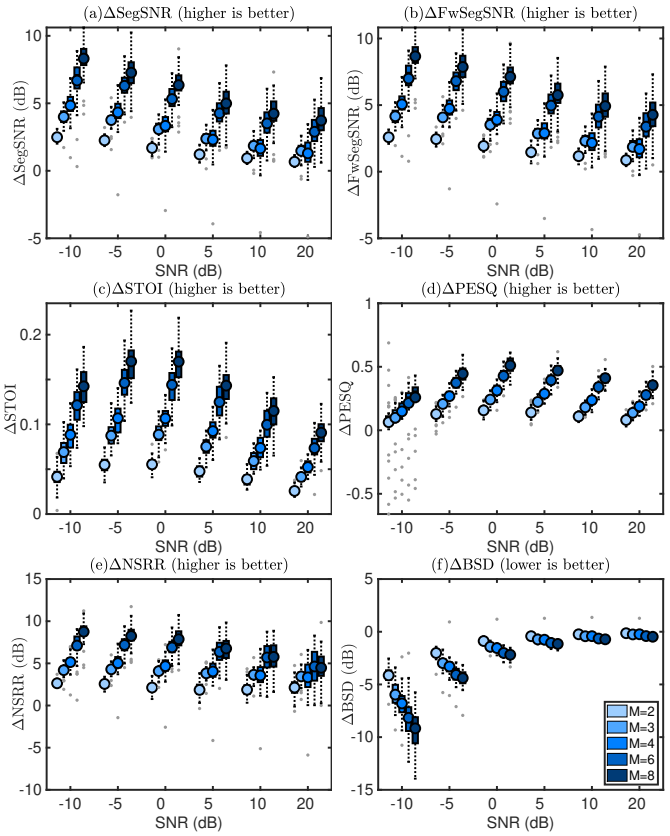

Fig. 3. Speech enhancement results for white noise in Lecture Room 2 using different numbers of microphones M.

algorithm. In the presence of white noise in Lecture Room 2, Fig. 1 highlights the limitation of using only the instantaneous lag as indicated by the lowest scores in all measures while the scores substantially improve as W increases.

The marginal improvement in all measures across all SNRs with frame size T is shown in Fig. 2. Similar results are observed for intermediate values of T. This is expected since larger frame sizes will provide a better estimate of the second-order statistics for the PEVD algorithm, assuming stationarity.

2) Number of Microphones: The impact of changing the number of microphones in the range 2 to 8 was investigated using the 8-channel linear array. For white noise in Lecture Room 2, Fig. 3 shows that PEVD can reduce noise and reverberation without sacrificing speech intelligibility and quality even with 2 microphones as indicated by the improvement in all metrics. Generally, PEVD performs better with the number of channels as expected, with very similar results obtained for intermediate values of M.

B. Comparison of Speech Enhancement Algorithms

1) Noise Reduction of Anechoic Speech in Noise: The illustrative example based on a clean speech corrupted by 0 dB babble noise in an anechoic environment is shown in Fig. 4. Overall, the energy of the babble noise has been significantly reduced after the PEVD-based enhancement, as reflected by an improvement in ΔSegSNR and ΔFwSegSNR, respectively in Table I. Although COLSUB [20], which uses a GEVD, improved in both ΔSegSNR and ΔFwSegSNR, the structures of the babble noise and speech signals are lost for example

TABLE I
NOISE REDUCTION PERFORMANCE OF AN ANECHOIC SPEECH IN 0 DB DIFFUSE BABBLE NOISE EXAMPLE, WITH SCORES (SEGSNR=-8.93 DB, FWSEGSNR=-6.86 DB, STOI=0.674 AND PESQ=1.63).

Algorithm	ΔSegSNR	ΔFwSegSNR	ΔSTOI	ΔPESQ
MWF	-2.37 dB	-2.91 dB	-0.122	-0.22
OMWF	**5.37 dB**	**4.31** dB	**0.104**	**0.41**
COLSUB	5.23 dB	3.92 dB	0.024	0.14
MCSUB	1.96 dB	-0.60 dB	0.008	0.17
PEVD	4.36 dB	4.22 dB	0.097	0.34
GWPE	-0.04 dB	-0.23 dB	-0.009	-0.01
WPD	0.60 dB	-0.03 dB	-0.004	0.02
ISCLP	-7.65 dB	-7.99 dB	-0.167	-0.33

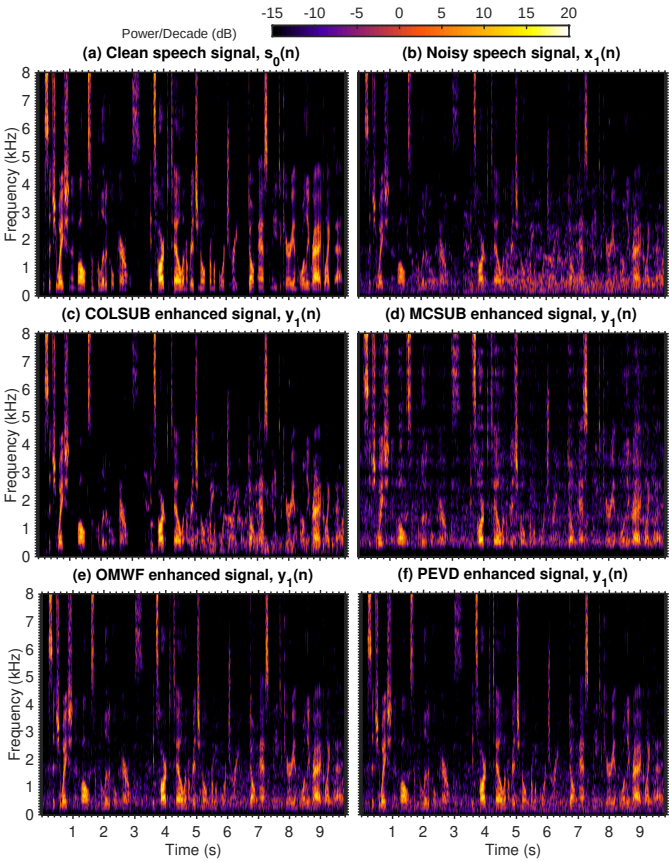

Fig. 4. Normalized spectrograms of anechoic speech in 0 dB babble noise example and the processed signals.

at 3 s and processing artefacts are observed in the listening examples [72]. These artefacts may have led to lower ΔPESQ and ΔSTOI for COLSUB compared to PEVD.

For this example, Table I shows that OMWF performed best in all measures as expected because it uses prior knowledge of the clean speech signal. Despite being a completely blind approach without using any noise estimator, PEVD is the best performing algorithm after OMWF in terms of ΔFwSegSNR, ΔSTOI and ΔPESQ. This example demonstrates that the PEVD approach can perform comparably to an oracle algorithm. Enhancement using MCSUB or WPD offers some improvement in some measures while MWF, GWPE or ISCLP worsens the scores across all measures.

The $\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z)$ for the above example used in PEVD is presented in Fig. 5, with the $(p,q)^{\text{th}}$ subplot representing the

This article has been accepted for publication in IEEE/ACM Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2021.3120630

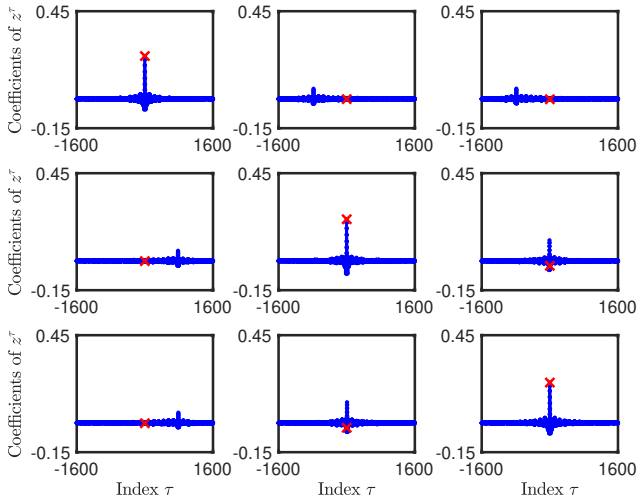

Fig. 5. Space-time covariance, $\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z)$, of the noisy speech example in Fig. 4 with the instantaneous covariance matrix marked by red cross signs. The $(p,q)^{\text{th}}$ subplot represents the z-transform of the element $r_{p,q}(\tau)$.

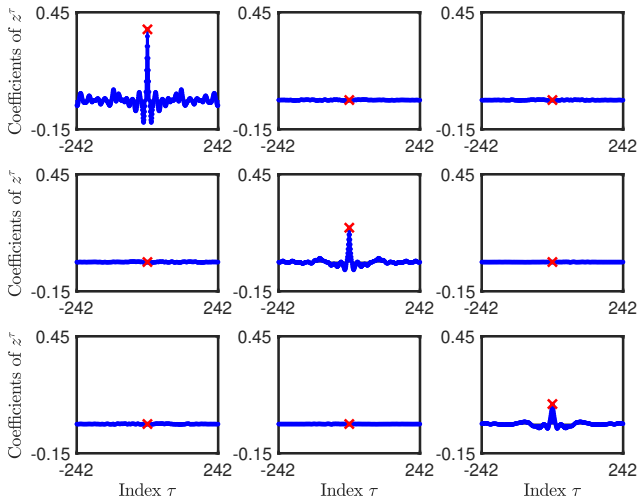

Fig. 6. $\boldsymbol{\Lambda}(z)$ polynomial matrix of the noisy speech example in Fig. 5 with each subplot representing the z-transform of each element.

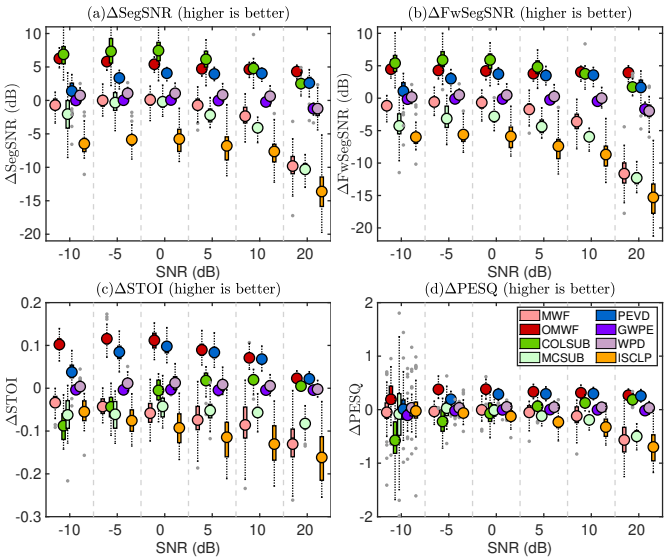

Fig. 7. Noise reduction results for Noisex babble noise.

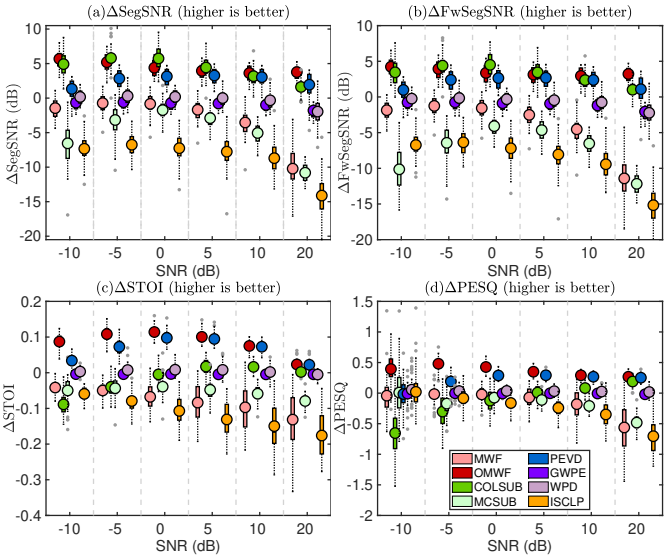

Fig. 8. Noise reduction results for SoundFx restaurant noise.

z-transform of $r_{p,q}(\tau)$, p and q correspond, respectively, to the row and column indices of Fig. 5. An EVD, which is applied to the instantaneous covariance matrix corresponding to the coefficient of z^0, can only decorrelate the signals at a single time lag [19], [21]. This is inadequate for this example because the cross-correlations on the off-diagonals have peaks at other time lags. Instead, the PEVD can impose decorrelation over a range of time lags as shown in Fig. 6, where every off-diagonal element has a magnitude less than 0.58% of the trace-norm of $\mathbf{R}_{\mathbf{xx}}(0)$. Because of trimming, the polynomial order of $\boldsymbol{\Lambda}(z)$ in Fig. 6 is much smaller than that of $\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z)$ in Fig. 5.

The results of Monte-Carlo simulations involving 50 trials of anechoic speech with -10 dB to 20 dB babble noise are plotted in Fig. 7. OMWF is the best performing algorithm which can consistently provide improvement in all metrics over the entire range of SNRs. While COLSUB performs up to 2 dB better than OMWF in ΔSegSNR and ΔFwSegSNR at lower SNRs, the ΔSTOI and ΔPESQ are negative. After OMWF, PEVD is the best performing algorithm in ΔSTOI and ΔPESQ and one of the best in ΔSegSNR and ΔFwSegSNR. The noise reduction performances of MWF, MCSUB, GWPE, WPD and ISCLP are limited in all measures across all SNRs.

The proposed and baseline approaches were tested for an extensive range of noise types with results given in Fig. 8, for the example of restaurant noise. Noise reduction for COLSUB is achieved at the cost of reduced speech intelligibility and quality. The OMWF, which uses oracle knowledge of the clean speech, performs best across all measures and can simultaneously reduce noise and improve speech intelligibility and quality. PEVD comes close as second best, despite being completely blind, just like the other approaches. Most approaches do not improve and may even reduce the measures.

2) Speech Dereverberation in the Absence of Noise: An illustrative result for a single reverberant speech example in Lecture Room 2 without background noise is shown in Fig. 9. The spectrograms show qualitatively that both GWPE and PEVD can suppress reverberation while retaining the overall speech structure. However, GWPE seems to have applied

This article has been accepted for publication in IEEE/ACM Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2021.3120630

TABLE II
DEREVERBERATION PERFORMANCE OF A NOISELESS, REVERBERANT EXAMPLE IN LECTURE ROOM 2, WITH MEASURED SCORES (NSRR=-6.03 DB, BSD=0.38 DB, STOI=0.810, PESQ=2.01).

Algorithm	ΔNSRR	ΔBSD	ΔSTOI	ΔPESQ
MWF	-1.06 dB	0.27 dB	-0.057	-0.26
OMWF	0.10 dB	0.04 dB	0.009	0.16
COLSUB	0.00 dB	0.00 dB	0.000	0.00
MCSUB	-3.20 dB	0.28 dB	-0.028	0.01
PEVD	**5.79 dB**	-0.12 dB	0.018	0.12
GWPE	0.68 dB	**-0.25 dB**	**0.091**	**0.70**
WPD	4.02 dB	-0.23 dB	0.071	0.40
ISCLP	0.29 dB	-0.22 dB	0.040	0.41

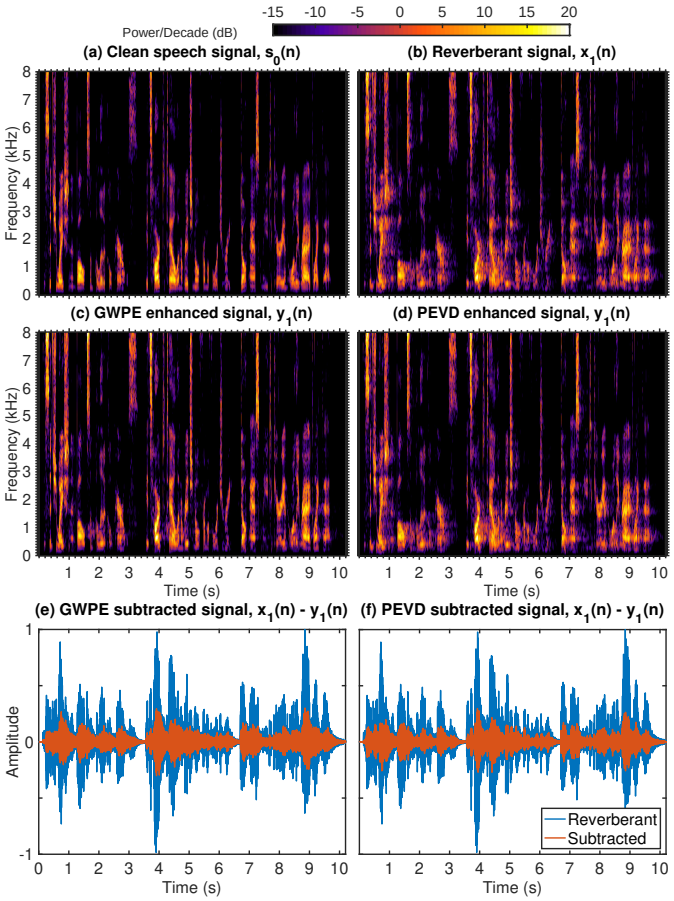

Fig. 9. Normalized spectrograms of reverberant speech example and the improvement, $x_1(n) - y_1(n)$, is scaled and time aligned with $x_1(n)$.

a more aggressive suppression which results in a cleaner spectrogram, as supported by the lowest ΔBSD in Table II, indicating lower spectral distortions than PEVD. PEVD outperforms all algorithms in ΔNSRR and WPD, which uses the steering vector computed from the ground truth DoA, ranks second. Based on ΔSTOI and ΔPESQ, GWPE performed best, followed by WPD, ISCLP, PEVD and OMWF. The other algorithms do not provide significant improvements.

To understand the processing involved, the difference between the reverberant and processed signals are plotted in red along with the reverberant signal in blue. The PEVD improvement in Fig. 9f follows the temporal changes of the reverberant signal more closely while the GWPE improvement has higher amplitudes for example at roughly 1 s and 4 s,

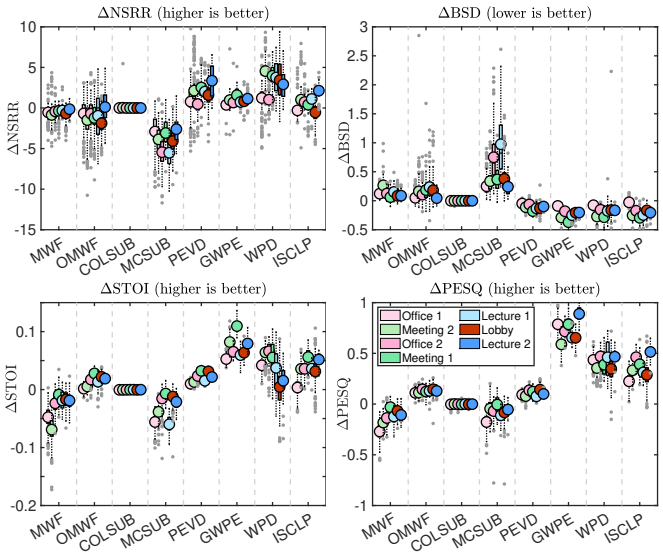

Fig. 10. Speech dereverberation results in noiseless environments.

which suggests more aggressive processing. Listening examples for GWPE indicate the removal of most of the early but not the direct-path and some late reverberant components, as also observed in [43]. Listening examples for PEVD, on the other hand, indicate that the direct-path and some early reflection components are retained in the enhanced signal in the first channel, as expected for reasons given in Section IV-B. The late reverberant components, which are absent in the enhanced signal, are observed in the second and third channels because of orthogonality [72].

Across all rooms, due to the removal of late reflections, PEVD ranks second in ΔNSRR after WPD as shown in Fig. 10. Moreover, GWPE processes the signals aggressively as discussed above and therefore leads to the best improvement in ΔBSD, ΔSTOI and ΔPESQ. The overall performance of WPD and ISCLP are comparable and is followed by PEVD. In this noise-free scenario, speech dereverberation using OMWF is equivalent to a Wiener deconvolution. Averaged across all rooms, ΔNSRR and ΔBSD for OMWF are worsened by -0.94 dB and 0.14 dB, while ΔSTOI and ΔPESQ are increased by 0.015 and 0.13 respectively. This is expected because the FIR filter of 80 taps is insufficient to invert with high accuracy the room impulse responses which are thousands of samples long. COLSUB also offers no improvement while MWF and MCSUB tend to worsen all metrics.

3) Speech Enhancement of Reverberant Speech in Noise: Results for reverberant speech corrupted by ACE babble noise in the strongly reverberant Lecture Room 2 is shown in Fig. 11. For SNR$\leq$10 dB, a trade-off between noise reduction and speech intelligibility and quality, is observed for COLSUB which shows negative ΔSTOI and ΔPESQ. On the other hand, OMWF, which is designed to minimize speech distortion using knowledge of the clean speech, performs the best in ΔSTOI and ΔPESQ but not in ΔSegSNR and ΔFwSegSNR. This also reflects the fact that speech intelligibility may not necessarily be affected by noise levels, up to some limit, compared to speech. Using prior knowledge of DoA to compute the steering

This article has been accepted for publication in IEEE/ACM Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2021.3120630

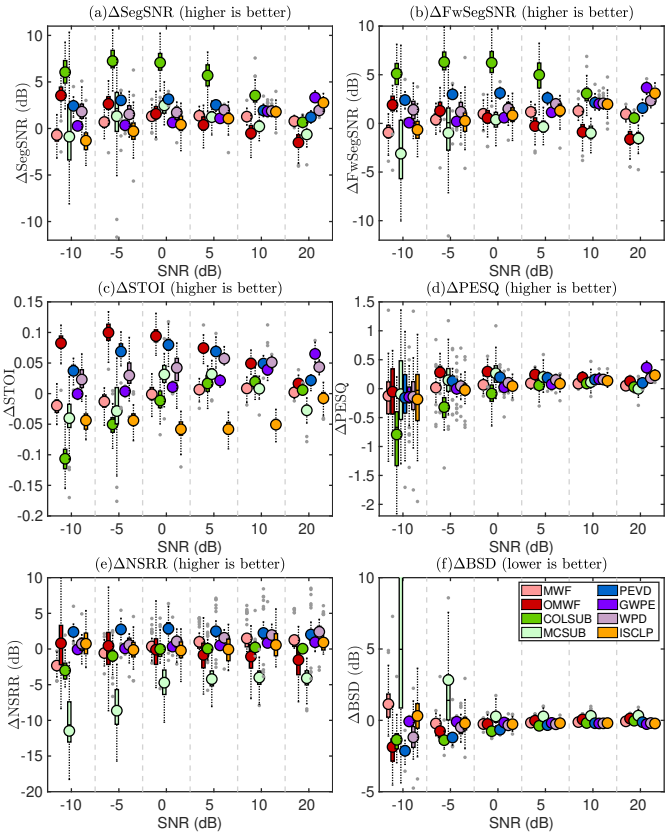

Fig. 11. Speech enhancement results for babble noise in Lecture Room 2.

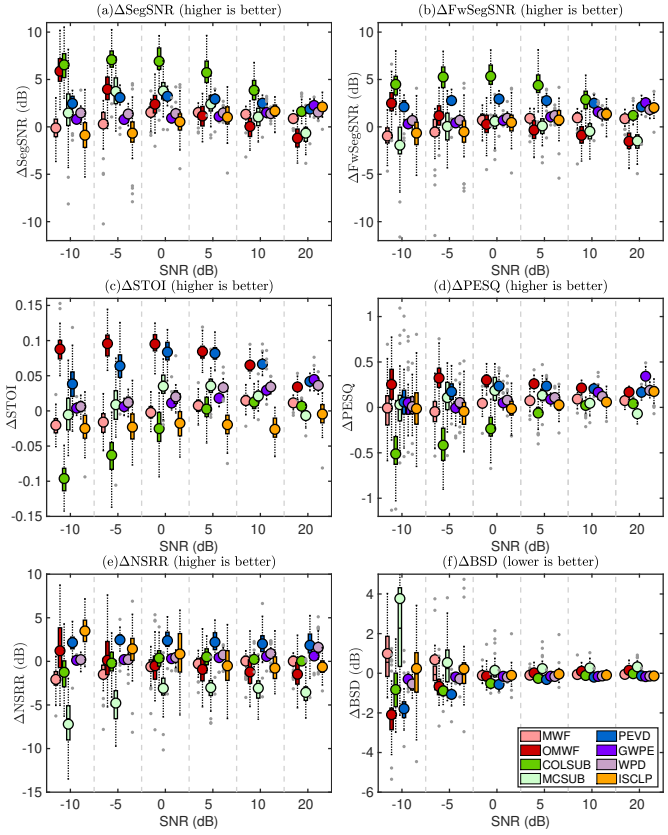

Fig. 12. Speech enhancement results for fan noise in Office 1.

vector, WPD provides further improvement in all measures over GWPE. At 20 dB SNR, algorithms targeting reverberation such as GWPE and joint approaches like WPD and ISCLP, perform better than the noise reduction approaches. Fig. 11 also shows that PEVD performs best in ΔNSRR and ΔBSD, ranks second in ΔSegSNR, ΔFwSegSNR and ΔSTOI, and one of the best in ΔPESQ for all SNRs.

Results for the ACE fan noise in the scenario for Office 1 in Fig. 12 also show that noise reduction algorithms perform better at low SNRs, i.e. below 5 dB, while dereverberation algorithms perform better at high SNRs, for example at 20 dB. Across all noise and reverberation conditions, PEVD can consistently suppress both noise and reverberation and improve both speech intelligibility and quality. In addition, PEVD can offer further improvement by up to 5 dB in ΔFwSegSNR, 0.1 in ΔSTOI and -2 dB in ΔBSD over state-of-the-art joint approaches, WPD and ISCLP. Furthermore, listening examples in [72] provide supporting evidence that our PEVD-based approach does not introduce any processing artefacts into the enhanced signal.

Comprehensive testing over the complete ACE corpus has been performed with summary results given in Table III, for the example of the most and least reverberant rooms in 0 dB babble noise. Results indicate that PEVD can consistently provide the best dereverberation and one of the best noise reduction performances. Only OMWF outperforms PEVD in ΔSTOI. However, OMWF requires oracle prior information, which is unavailable in practice. In contrast, the proposed PEVD approach is blind, using only the microphone signals. The source code and listening examples corresponding to the results are available in the supplementary files of this manuscript as well as [72].

VII. Conclusion

In this paper, multi-channel broadband speech signals were modelled using polynomial matrices in order to capture the spatial, spectral as well as temporal correlations between microphones. It was shown that the proposed PEVD approach can decorrelate the microphone signals in space, time and frequency simultaneously. A novel speech enhancement algorithm based on the PEVD was proposed. The proposed algorithm achieves significant noise reduction and dereverberation using a weighted combination of the signal subspace. Comparative simulations under diverse acoustic conditions have indicated that the proposed method consistently improves noise reduction metrics, speech intelligibility and quality scores, and dereverberation measures. Despite being a blind and unsupervised algorithm, the approach does not rely on any noise estimator and does not introduce any processing artefacts into the enhanced signal as observed in the listening examples at [72].

VIII. Acknowledgment

The authors are grateful to the anonymous reviewers for making suggestions that resulted in an improved manuscript.

TABLE III
ENHANCEMENT PERFORMANCE EVALUATED ON THE ACE CORPUS FOR 0 DB BABBLE NOISE, MEASURED USING (MEAN±STANDARD DEVIATION). ΔFWSEGSNR, ΔNSRR AND ΔBSD ARE MEASURED IN DB. THE PROPOSED PEVD APPROACH CONSISTENTLY PERFORMS AMONG THE BEST.

Room (T_{60})	Office 1 (0.332 s)				Lecture Room 2 (1.22 s)			
Algorithm	ΔFwSegSNR	ΔSTOI	ΔNSRR	ΔBSD	ΔFwSegSNR	ΔSTOI	ΔNSRR	ΔBSD
MWF	-0.96 ± 1.30	-0.020 ± 0.015	-2.07 ± 1.30	0.99 ± 1.70	-0.95 ± 1.10	-0.020 ± 0.015	-2.32 ± 0.99	1.14 ± 1.30
OMWF	2.49 ± 1.90	**0.088 ± 0.022**	1.21 ± 3.30	**-2.09 ± 1.40**	1.90 ± 1.50	**0.083 ± 0.018**	0.80 ± 4.10	-1.86 ± 1.70
COLSUB	**4.49 ± 1.50**	-0.096 ± 0.024	-1.25 ± 2.20	-0.82 ± 1.00	**5.12 ± 1.60**	-0.107 ± 0.025	-3.00 ± 1.50	-1.36 ± 1.00
MCSUB	-1.93 ± 2.70	-0.005 ± 0.029	-7.21 ± 3.10	3.77 ± 6.50	-3.11 ± 6.30	-0.040 ± 0.038	-11.48 ± 6.10	17.80 ± 44.00
PEVD	2.11 ± 1.00	0.038 ± 0.030	2.15 ± 1.20	-1.79 ± 0.81	2.40 ± 0.60	0.037 ± 0.013	**2.39 ± 0.80**	**-2.13 ± 0.66**
GWPE	0.32 ± 0.18	0.004 ± 0.005	0.12 ± 0.19	-0.30 ± 0.17	0.08 ± 0.11	-0.001 ± 0.004	-0.06 ± 0.16	-0.08 ± 0.11
WPD	0.67 ± 0.41	0.006 ± 0.008	0.20 ± 0.50	-0.54 ± 0.36	1.40 ± 1.30	0.023 ± 0.019	0.73 ± 1.30	-1.19 ± 1.20
ISCLP	-0.65 ± 1.90	-0.025 ± 0.024	**3.49± 1.70**	0.24 ± 2.00	-0.66 ± 1.40	-0.044 ± 0.016	0.74 ± 2.00	0.31 ± 1.50

REFERENCES

[1] B. Widrow and F.-L. Luo, "Microphone arrays for hearing aids: An overview," *Speech Commun.*, vol. 39, pp. 139–146, 2003.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. Ray Liu, Eds. John Wiley & Sons, Inc., 2008.

[3] A. Kuklasinski and J. Jensen, "Multichannel Wiener filters in binaural and bilateral hearing aids—Speech intelligibility improvement and robustness to DoA errors," *J. Audio Eng. Soc. (AES)*, vol. 65, no. 1/2, pp. 8–16, 2017.

[4] C. Evers, Y. Dorfan, S. Gannot, and P. A. Naylor, "Source tracking using moving microphone arrays for robot audition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2017.

[5] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018.

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

[7] Y. Huang, J. Benesty, and J. Chen, "Analysis and comparison of multichannel noise reduction methods in a common framework," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 5, pp. 957–968, Jul. 2008.

[8] E. A. P. Habets and P. A. Naylor, "Dereverberation," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. John Wiley & Sons, Inc., 2018, pp. 317–343.

[9] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement systems," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2009, pp. 4409–4412.

[10] M. Torcoli, "An improved measure of musical noise based on spectral kurtosis," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*, 2019, pp. 90–94.

[11] K. Tan and D. Wang, "Improving robustness of deep learning based monaural speech enhancement against processing artifacts," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 6914–6918.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[13] 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.

[14] C. Uhle, M. Torcoli, and J. Paulus, "Controlling the perceived sound quality for dialogue enhancement with deep learning," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 51–55.

[15] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[16] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[18] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[19] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[20] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," *IEEE Signal Process. Lett.*, vol. 9, no. 7, pp. 204–206, Jul. 2002.

[21] F. Jabloun and B. Champagne, "A multi-microphone signal subspace approach for speech enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2001, pp. 205–208.

[22] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 497–507, 2000.

[23] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[24] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[25] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, ser. Springer Topics in Signal Processing, 2017.

[26] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Orlando, Florida, USA, May 2002, pp. 901–904.

[27] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 39–60.

[28] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum Likelihood PSD Estimation for Speech Enhancement in Reverberation and Noise," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, Sep. 2016.

[29] W. Xue, A. H. Moore, M. Brookes, and P. A. Naylor, "Modulation-Domain Multichannel Kalman Filtering for Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1833–1847, Oct. 2018.

[30] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 740–754, Jan. 2020.

[31] J. B. Allen, "Synthesis of pure speech from a reverberant signal," U.S. Patent 3 786 188, 1974.

[32] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *Proc. Int. on Workshop Acoust. Echo and Noise Control (IWAENC)*, 2003, pp. 99–102.

[33] Y. Huang, J. Benesty, and J. Chen, "Dereverberation," in *Springer Handbook of Speech Processing*. Springer-Verlag, 2008, pp. 929–943.

[34] Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Process.*, vol. 82, pp. 1127–1138, Aug. 2002.

[35] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[36] L. Tong and S. Perreau, "Multichannel blind identification: From subspace to maximum likelihood methods," *Proc. IEEE*, vol. 86, no. 10, pp. 1951–1968, 1998.

[37] A. W. Khong, P. A. Naylor, and J. Benesty, "A low delay and fast converging improved proportionate algorithm for sparse system identifi-

cation," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2007, no. 1, pp. 1–8, Apr. 2007.

[38] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 165–169, Jul. 1979.

[39] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[40] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. Int. on Workshop Acoust. Echo and Noise Control (IWAENC)*, Aug. 2010.

[41] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[42] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 9, pp. 1379–1390, Sep. 2014.

[43] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[44] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, and T. Nakatani, "Strategies for distant speech recognitionin reverberant environments," *EURASIP J. on Advances in Signal Process.*, vol. 2015, no. 1, pp. 1–15, 2015.

[45] T. Nakatani and K. Kinoshita, "A unified convolutional beamformer for simultaneous denoising and dereverberation," *IEEE Signal Process. Lett.*, vol. 26, no. 6, pp. 903–907, Jun. 2019.

[46] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[47] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 240–254, May 2000.

[48] 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.

[49] S. Redif, S. Weiss, and J. G. McWhirter, "An approximate polynomial matrix eigenvalue decomposition algorithm for para-hermitian matrices," in *Proc. Int. Symp. on Signal Process. and Inform. Technol. (ISSPIT)*, 2011, pp. 421–425.

[50] 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.

[51] 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.

[52] 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.

[53] S. Weiss, N. J. Goddard, S. Somasundaram, I. K. Proudler, and P. A. Naylor, "Identification of broadband source-array responses from sensor second order statistics," in *Sensor Signal Process. for Defence Conf. (SSPD)*, 2017.

[54] S. Weiss, M. Alrmah, S. Lambotharan, J. G. McWhirter, and M. Kaveh, "Broadband angle of arrival estimation methods in a polynomial matrix decomposition framework," in *Proc. IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Process. (CAMSAP)*, 2013, pp. 109–112.

[55] 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.

[56] S. Weiss, S. Redif, T. Cooper, C. Liu, P. D. Baxter, and J. G. McWhirter, "Paraunitary oversampled filter bank design for channel coding," *EURASIP J. on Advances in Signal Process.*, vol. 2006, no. 1, Mar. 2006.

[57] 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)*, 2019, pp. 125–129.

[58] V. W. Neo, C. Evers, and P. A. Naylor, "PEVD-based speech enhancement in reverberant environments," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 186–190.

[59] V. W. Neo, C. Evers, and P. A. Naylor, "Speech dereverberation performance of a polynomial-EVD subspace approach," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2020, pp. 221–225.

[60] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244, Jun. 2003.

[61] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1106–1118, Jun. 2018.

[62] 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.

[63] P. P. Vaidyanathan, *Multirate Systems and Filters Banks*. New Jersey, USA: Prentice Hall, 1993.

[64] P. D. Baxter and J. G. McWhirter, "Blind signal separation of convolutive mixtures," in *Proc. Asilomar Conf. on Signals, Syst. & Comput.*, 2003, pp. 124–128.

[65] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium (LDC), Philadelphia, USA, Corpus, 1993.

[66] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[67] 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.

[68] S. Ideas, "International Sound Effects Library," Richmond Hill, Ont, 1999.

[69] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[70] "Objective measurement of active speech level," Int. Telecommun. Union (ITU-T), Recommendation, Mar. 1993.

[71] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[72] V. W. Neo, C. Evers, and P. A. Naylor, "Enhancement of noisy reverberant speech using PEVD," Oct. 2020. [Online]. Available: https://vwn09.github.io/pevd-enhance/

[73] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2006, pp. 1447–1450.

[74] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[75] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Int. Telecommun. Union (ITU-T), Recommendation P.862, Nov. 2003.

[76] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets, "Signal-based performance evaluation of dereverberation algorithms," *J. of Elect. and Comput. Eng.*, vol. 2010, pp. 1–5, 2010.

Vincent W. Neo (Student Member, IEEE) is a PhD candidate at Imperial College London funded through a scholarship provided by the Defence Science and Technology Agency (DSTA), Singapore. He has worked at DSTA and Nuance Communications in various engineering roles. He received the MEng degree in Electrical and Electronic Engineering from Imperial College London, UK, in 2014. His research focuses on polynomial matrix techniques with applications to speech, audio and acoustic signal processing.

This article has been accepted for publication in IEEE/ACM Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2021.3120630

Christine Evers (Senior Member, IEEE) is a lecturer in the School of Electronics and Computer Science at the University of Southampton. She was the recipient of an EPSRC Fellowship, hosted at Imperial College London, between 2017-2019. She worked as a research associate at Imperial College London between 2014-2017; as a senior systems engineer at Selex Electronic Systems between 2010-2014; and as a research fellow at the University of Edinburgh between 2009-2010. She received her PhD from the University of Edinburgh in 2010; her MSc degree in Signal Processing and Communications from the University of Edinburgh in 2006; and her BSc degree in Electrical Engineering and Computer Science from Jacobs University, Germany, in 2005. Her research focuses on Bayesian learning for machine listening. She is currently member of the IEEE SPS Technical Committee on Audio and Acoustic Signal Processing and serves as an associate editor for IEEE/ACM Transactions on Audio, Speech, and Language Processing, and the EURASIP Journal on Audio, Speech, and Music Processing.

Patrick A. Naylor (M'89, SM'07, F'20) is Professor of Speech and Acoustic Signal Processing at Imperial College London. He received the BEng degree in Electronic and Electrical Engineering from the University of Sheffield, UK, and the PhD degree from Imperial College London, UK. His research interests are in speech, audio and acoustic signal processing. His current research addresses microphone array signal processing, speaker diarization, and multichannel speech enhancement for applications including binaural hearing aids and augmented reality. He has also worked on speech dereverberation including blind multichannel system identification and equalization, acoustic echo control, non-intrusive speech quality estimation, and speech production modelling with a focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several collaborative links with industry. He is currently President of the European Association for Signal Processing (EURASIP) and a member of the Board of Governors of the IEEE Signal Processing Society.