MVDR beamforming for underdetermined sound source separation using iterative PSD estimation in beamspace

Jin Xuan Teh and Yusuke Hioka
Acoustics and Vibration Research Centre, University of Auckland, Auckland, New Zealand
E-mail: jteh840@aucklanduni.ac.nz

***Abstract*—We present an iterative beamforming framework that combines minimum-variance distortionless response (MVDR) beamforming with power spectral density (PSD) estimation in beamspace. This framework leverages the sparsity of speech in the time-frequency (TF) domain to improve the performance of MVDR beamforming for underdetermined sound source separation. Initially, PSD of interferences are estimated in beamspace using multiple fixed beamformers. These estimates are used to compute MVDR weights for each TF bin, attenuating the dominant interferers in each bin. In each iteration, the enhanced spatial selectivity of the MVDR outputs refines the interference PSD estimates, which in turn improve the subsequent weight computations. This iterative process preserves the distortionless response of MVDR beamforming while eliminating the need for nonlinear post-filtering and its associated artefacts. Simulations and real-world experiments conducted across diverse acoustic environments demonstrate substantial improvements in signal-to-interference ratio (SIR) and short-time objective intelligibility (STOI) compared to conventional MVDR. Notably, under anechoic conditions, the proposed method approaches the SIR improvement of a time-frequency MVDR beamformer constructed with oracle source PSD.**

***Index Terms*—Sound source separation, underdetermined scenarios, MVDR beamforming, power spectral density estimation, microphone arrays.**

I. Introduction

Hands-free audio applications, spanning teleconferencing systems, and hearing aids, require effective sound source separation techniques to ensure excellent speech quality and intelligibility. In scenarios with multiple competing talkers, microphones often suffer from low signal-to-interference ratios (SIRs), especially in underdetermined scenarios (where the number of sound sources N exceeds the number of microphones M). To address these challenges, a variety of approaches have been developed. Early approaches include time-frequency (TF) masking based on the W-disjoint orthogonality assumption, such as the DUET algorithm [1]–[3], multichannel Wiener filtering (MWF) for speech enhancement [4], and statistical latent variable models for blind source separation [5]. Beyond these classical methods, Hioka *et al*. [6] proposed PSD estimation in beamspace, a method that employs source PSD estimated using a bank of fixed beamformers with varied directivity patterns, which are then used to design nonlinear post-filters. More recently, deep neural networks have been trained to predict TF masks as nonlinear post-filters achieving strong interference reduction [7] but at the cost of extensive supervised training data. However, despite their strong interference suppression in underdetermined scenarios, these techniques frequently introduce target-signal distortions [8] that degrade the performance of downstream applications such as automatic speech recognition [9].

In contrast to these nonlinear methods, beamforming techniques apply a linear spatial filter, to enhance sound from the desired direction while attenuating others, without introducing nonlinear artefacts [10]. Notably, the minimum-variance distortionless response (MVDR) beamformer [11] is particularly attractive because it minimises output power under the constraint of a distortionless response for the target signal. A major limitation a MVDR beamformer is its degraded performance in underdetermined scenarios. When $N > M$, the microphone array cannot form enough independent spatial nulls to suppress all $N-1$ interferers with only M microphones, leaving residual interference. Furthermore, MVDR weight computation depends on the mixture's spatial covariance, which itself requires accurate estimates of the interference PSDs; any error in those PSD estimates leads to suboptimal beamformer weights [12].

In related work, Kubo *et al.* introduced a mask-based MVDR with time-varying target and noise spatial covariance matrix estimated from TF masks, but this approach relies on accurate masks and is vulnerable to permutation errors across frequency [13]. Yamaoka *et al.* proposed TF-bin-wise switching/combination of multiple distortionless beamformers; its performance is sensitive to the number of beamformers employed and to how the beamformer set is initialised [14].

In this paper, we propose an iterative framework that improves the performance of MVDR beamforming in underdetermined scenarios. The proposed method leverages speech sparsity in the TF domain by integrating PSD estimation in beamspace [6] with MVDR beamforming through a closed-loop refinement process. The iteration is initialised with PSD estimates obtained via the estimation method of [6], where a bank of fixed beamformers – each steered to one of the estimated directions of arrival (DOA) of the sound signals – extracts individual source PSDs from the mixture. These initial estimates are used to compute MVDR beamformer weights for each TF bin. Unlike the original approach in [6], which uses PSD estimates only to design a postfilter and treats beamforming and PSD estimation as separate stages, the proposed framework exploits the improved spatial selectivity of each beamformer output to refine PSD estimates in the following iteration. This feedback mechanism dynamically

979-8-3315-7206-8/25/$31.00

adjusts the beamformer's spatial nulls to the interferences observed in each TF bin based on the progressively refined interference PSDs. The proposed method requires the same prior as a conventional MVDR (requiring only source DOAs) and achieves superior sound separation performance in underdetermined scenarios without distorting the target signal.

II. Preparation

A. Problem Setup and Signal Model

We consider a scenario with N sound sources located in the far-field and their sound waves impinging on an array of M microphones. We assume that the contribution from diffuse noise is negligible. The signal captured by the m-th microphone in the short-time Fourier transform (STFT) domain at time frame index t and frequency bin f is modelled as:

$$\boldsymbol{x}(t,f) = \sum_{n=1}^{N} \boldsymbol{h}_n(f)\, s_n(t,f), \tag{1}$$

where $s_n(t,f)$ is the STFT of the n-th source signal.

$$\boldsymbol{h}_n(f) = \big[h_{1,n}(f), \ldots, h_{M,n}(f)\big]^T \tag{2}$$

is the $M \times 1$ steering vector [15] for the n-th source where $h_{m,n}(f)$ is the transfer function from the source n to the m-th microphone, and $(\cdot)^T$ denotes transpose.

We assume that, within each frame, the source signals are mutually uncorrelated, zero-mean stationary processes. The spatial covariance matrix of the observed microphone signals can be expressed as

$$\begin{aligned} \boldsymbol{\Phi}_{xx}(f) &= \mathbb{E}\big[\boldsymbol{x}(t,f)\,\boldsymbol{x}^H(t,f)\big] \\ &= \sum_{n=1}^{N} \phi_{s_n}(f)\,\boldsymbol{h}_n(f)\,\boldsymbol{h}_n^H(f) \\ &= \boldsymbol{H}(f)\,\boldsymbol{\Phi}_{ss}(f)\,\boldsymbol{H}^H(f), \end{aligned} \tag{3}$$

where $(\cdot)^H$ denotes the Hermitian transpose and $\mathbb{E}[\cdot]$ denotes expectation. $\boldsymbol{\Phi}_{ss}(f) = \mathrm{diag}\big(\phi_{s_1}(f), \ldots, \phi_{s_N}(f)\big)$ is the diagonal matrix of source PSDs, $\boldsymbol{H}(f) = [\boldsymbol{h}_1(f), \ldots, \boldsymbol{h}_N(f)]$ is the $M \times N$ steering (mixing) matrix, and the PSD of the n-th source is denoted by

$$\phi_{s_n}(f) \;=\; \mathbb{E}\big[|s_n(t,f)|^2\big], \tag{4}$$

Our objective is to recover each $s_n(t,f)$ from $\boldsymbol{x}(t,f)$, given that the number of sources N with their corresponding DOAs are known.

B. MVDR Beamforming

A MVDR beamformer pointed towards the n-th source is applied by:

$$y_n(t,f) = \boldsymbol{w}_n^H(f)\,\boldsymbol{x}(t,f), \tag{5}$$

where $y_n(t,f)$ is the beamformer's output and $\boldsymbol{w}_n(f)$ is the beamformer's filter weights which are obtained by solving the constrained optimisation problem [11]:

$$\begin{aligned} \min_{\boldsymbol{w}_n(f)} \quad & \mathbb{E}\big[|y_n(t,f)|^2\big] \;=\; \boldsymbol{w}_n^H(f)\,\boldsymbol{\Phi}_{xx}(f)\,\boldsymbol{w}_n(f) \\ \text{s.t.} \quad & \boldsymbol{w}_n^H(f)\,\boldsymbol{h}_n(f) = 1. \end{aligned} \tag{6}$$

This constraint enforces a distortionless response in the target direction, while (in the Capon form [11]) minimising the total output power with respect to the overall covariance. The solution to (6) is given by

$$\boldsymbol{w}_n^{\mathrm{Capon}}(f) = \frac{\boldsymbol{\Phi}_{xx}^{-1}(f)\,\boldsymbol{h}_n(f)}{\boldsymbol{h}_n^H(f)\,\boldsymbol{\Phi}_{xx}^{-1}(f)\,\boldsymbol{h}_n(f)}. \tag{7}$$

An alternative formulation minimises only the interference (and noise) power is known as MVDR beamformer [12]. Instead of using the total covariance $\boldsymbol{\Phi}_{xx}(f)$ to suppress all undesired components, it explicitly targets competing interference for maximal suppression:

$$\begin{aligned} \min_{\boldsymbol{w}_n(f)} \quad & \boldsymbol{w}_n^H(f)\,\boldsymbol{\Phi}_{\mathrm{int},n}(f)\,\boldsymbol{w}_n(f) \\ \text{s.t.} \quad & \boldsymbol{w}_n^H(f)\,\boldsymbol{h}_n(f) = 1, \end{aligned} \tag{8}$$

where the interference covariance matrix is

$$\boldsymbol{\Phi}_{\mathrm{int},n}(f) = \sum_{\substack{n'=1 \\ n' \neq n}}^{N} \phi_{s_{n'}}(f)\,\boldsymbol{h}_{n'}(f)\,\boldsymbol{h}_{n'}^H(f). \tag{9}$$

The corresponding solution is

$$\boldsymbol{w}_n^{\mathrm{MVDR}}(f) = \frac{\boldsymbol{\Phi}_{\mathrm{int},n}^{-1}(f)\,\boldsymbol{h}_n(f)}{\boldsymbol{h}_n^H(f)\,\boldsymbol{\Phi}_{\mathrm{int},n}^{-1}(f)\,\boldsymbol{h}_n(f)}. \tag{10}$$

In practice, $\boldsymbol{\Phi}_{xx}(f)$ can be estimated from the observed data (e.g., via sample covariance), however, the individual source PSDs $\phi_{s_{n'}}(f)$ required to construct $\boldsymbol{\Phi}_{\mathrm{int},n}(f)$ are typically unavailable. Consequently, the interference-focused MVDR formulation in (8) cannot be applied directly. These challenges motivate the development of PSD estimation techniques for interference.

C. PSD estimation in beamspace

In the PSD estimation in beamspace [6] we apply a bank of L fixed beamformers to the microphone array inputs, which provides beamformer outputs $y_\ell(f)$ for $\ell = 1, \ldots, L$. Each beamformer is designed with a distinct spatial response so that its output power represents a weighted sum of the contributions from all sources with varying weights. When diffuse noise is negligible, the beamformer output PSDs can be approximated as:

$$\begin{aligned} \phi_{y_\ell}(t,f) &= \mathbb{E}\big[\,|y_\ell(t,f)|^2\big] \\ &= \sum_{n=1}^{N} d_{\ell,n}(f)\,\phi_{s_n}(t,f), \end{aligned} \tag{11}$$

where

$$d_{\ell,n}(f) = \big|\boldsymbol{w}_\ell^{\mathrm{fixed}\,H}(f)\,\boldsymbol{h}_n(f)\big|^2$$

is the directivity gain of the ℓ-th beamformer towards the n-th source. By stacking the PSDs of the beamformer outputs and sources into vectors:

$$\boldsymbol{\phi}_y(t,f) = \begin{bmatrix} \phi_{y_1}(t,f) \\ \vdots \\ \phi_{y_L}(t,f) \end{bmatrix}, \quad \boldsymbol{\phi}_s(t,f) = \begin{bmatrix} \phi_{s_1}(t,f) \\ \vdots \\ \phi_{s_N}(t,f) \end{bmatrix},$$

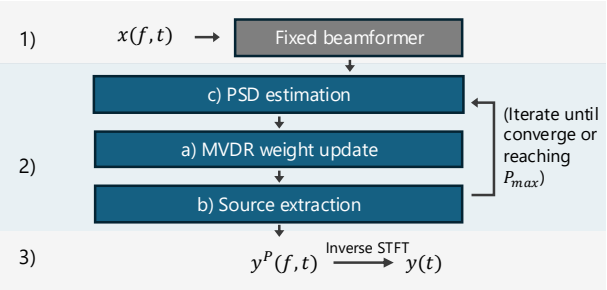

Fig. 1: Framework of the proposed method.

and organising the directivity gains into an $L \times N$ matrix

$$\boldsymbol{D}(f) = \begin{bmatrix} |d_{1,1}(f)|^2 & \cdots & |d_{1,N}(f)|^2 \\ \vdots & \ddots & \vdots \\ |d_{L,1}(f)|^2 & \cdots & |d_{L,N}(f)|^2 \end{bmatrix}, \tag{12}$$

we obtain the linear simultaneous equation:

$$\boldsymbol{\phi}_y(t,f) = \boldsymbol{D}(f)\,\boldsymbol{\phi}_s(t,f). \tag{13}$$

Given the beamformer output PSDs, $\boldsymbol{\phi}_y(t,f)$, and the known directivity matrix $\boldsymbol{D}(f)$ (determined by the fixed beamformer design and array geometry), the unknown source PSDs, $\boldsymbol{\phi}_s(t,f)$, can be estimated by solving (13) via least squares method:

$$\hat{\boldsymbol{\phi}}_s(t,f) = \boldsymbol{D}^{-1}(f)\,\boldsymbol{\phi}_y(t,f), \tag{14}$$

where $\hat{}$ denotes an estimate. A well-conditioned $\boldsymbol{D}(f)$ ensures stable inversion and precise PSD estimates, while an ill-conditioned $\boldsymbol{D}(f)$ amplifies noise and yields large estimation errors [6]; hence it is crucial to design the fixed beamformer bank that has diverse directivity patterns to minimise the condition number of $\boldsymbol{D}(f)$.

III. Proposed Method

The core novelty of the proposed method is the integration of PSD estimation in beamspace and MVDR beamforming in a feedback loop that exploits the sparsity of speech in the TF domain. Under the W-disjoint orthogonality assumption [1], [2], only a subset of the N sources is active in each TF bin. Since a conventional MVDR beamformer can control up to $M-1$ spatial nulls, it can effectively cancel all interferers in any bin with the number of active sources not exceeding M. Given that each source's DOA is known a priori, we first construct a well-conditioned directivity matrix $\boldsymbol{D}(f)$ using fixed beamformers steered to each DOA, and obtain initial PSD estimates of sources. These PSD estimates guide the computation of MVDR weights to attenuate the dominant interferer, yielding beamformer outputs with improved spatial selectivity. The refined MVDR outputs are then used to update the PSD estimates, which further improves the conditioning of $\boldsymbol{D}(f)$ in subsequent iterations. By alternately updating PSD estimates and recomputing MVDR weights, the algorithm achieves progressively more accurate PSD estimation and improved beamformer performance. Exploiting sparsity in the TF domain, it effectively suppresses interferers even when $N > M$ in the time domain, while preserving a distortionless response for the target signal.

Our algorithm realises a closed-loop refinement process, whereby each component iteratively enhances the other. As shown in Fig. 1, processing proceeds in three stages:

1) **Initialisation.** A bank of $L = N$ fixed beamformers $\{\boldsymbol{w}_\ell^{\text{fixed}}(f)\}$ is steered to the N source DOAs. Applying these beamformer to the mixture $\boldsymbol{x}(t,f)$ yields initial PSD observations $\boldsymbol{\phi}_y^{(0)}(t,f)$, from which the source PSDs $\hat{\boldsymbol{\phi}}_s^{(0)}(t,f)$ are recovered by solving the simultaneous equation (see (14)).
2) **Iterative PSD estimation and MVDR update.** For iteration $p = 1, 2, \ldots, P_{\max}$, perform the following steps for each source $n = 1, \ldots, N$:
   a) **Interference covariance estimation and MVDR update:** Estimate the interference covariance matrix
   $$\hat{\boldsymbol{\Phi}}_{\text{int},n}^{(p)}(t,f) = \sum_{\substack{n'=1 \\ n' \neq n}}^{N} \hat{\phi}_{s_{n'}}^{(p-1)}(t,f)\,\boldsymbol{h}_{n'}(f)\,\boldsymbol{h}_{n'}^H(f). \tag{15}$$
   Then update the MVDR beamformer weights:
   $$\boldsymbol{w}_n^{(p)}(t,f) = \frac{\hat{\boldsymbol{\Phi}}_{\text{int},n}^{(p)}(t,f)^{-1}\,\boldsymbol{h}_n(f)}{\boldsymbol{h}_n^H(f)\,\hat{\boldsymbol{\Phi}}_{\text{int},n}^{(p)}(t,f)^{-1}\,\boldsymbol{h}_n(f)}. \tag{16}$$
   b) **Source extraction:** Apply the updated MVDR beamformers to the n-th source:
   $$y_n^{(p)}(t,f) = \boldsymbol{w}_n^{(p)H}(t,f)\,\boldsymbol{x}(t,f). \tag{17}$$
   c) **PSD estimation and smoothing:** Estimate the PSD of each separated signals:
   $$\tilde{\phi}_{s_n}^{(p)}(t,f) = \mathbb{E}\big[\,|y_n^{(p)}(t,f)|^2\big], \tag{18}$$
   and update with smoothing factor $0 < \alpha < 1$:
   $$\hat{\phi}_{s_n}^{(p)}(t,f) = \alpha\,\hat{\phi}_{s_n}^{(p-1)}(t,f) + (1-\alpha)\,\tilde{\phi}_{s_n}^{(p)}(t,f). \tag{19}$$
   d) **Convergence check:** Compute the relative change in $\hat{\boldsymbol{\phi}}_n^{(p)}(t,f)$ for all n. Terminate the iterations when the relative change is below a set threshold, chosen heuristically based on validation experiments to balance convergence speed with PSD-estimation accuracy.
3) **Reconstruction.** After convergence (or upon reaching $P_{\max}$), the time-domain estimate $y_n(t)$ is obtained via inverse STFT (ISTFT) of $y_n^{(P)}(t,f)$ in (17).

IV. Experimental Setup

We evaluate the proposed method under two conditions:

- **Simulated anechoic conditions:** An M-channel uniform linear array (ULA) with 4 cm inter-element spacing was used. The target and $N-1$ interferers were placed uniformly on a circle of radius 1 m around the array. A spatially diffuse noise field constructed by superposing uncorrelated noise from multiple directions, was added at a signal-to-noise ratio (SNR) [12] of 20 dB relative to the source signals.
- **Real-world reverberant experiments:** Recorded room impulse responses (RIRs) from an 8-channel ULA (4 cm spacing) were derived from the database of Hadad *et al.* [16]. These RIRs cover reverberation times $T_{60} \in \{160, 360, 610\}$ ms. To vary the microphone count, we selected M channels at the centre of the 8-channel ULA. By adjusting M, N,

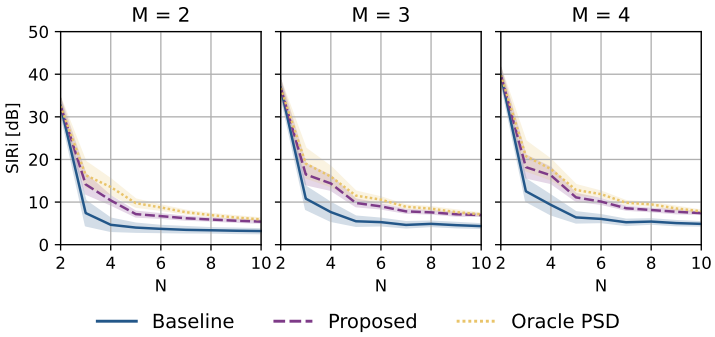

Fig. 2: Simulated SIRi in anechoic conditions; shaded areas show 95% confidence intervals. (CIs).

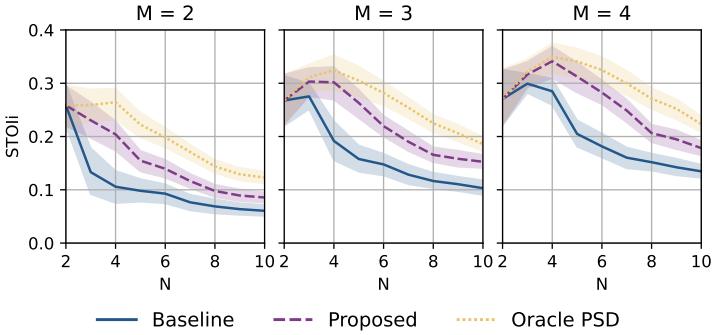

Fig. 3: Simulated STOIi in anechoic conditions; shaded areas show 95% CIs.

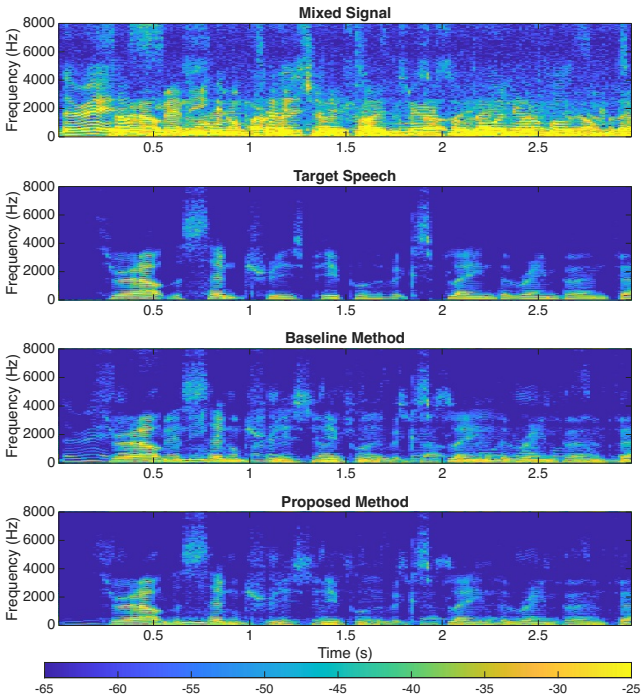

Fig. 4: Spectrogram of the mixed signal, target speech, baseline MVDR output, and proposed method output for the simulation with $M = 3$ and $N = 5$. The colour bar indicates magnitude in dB.

and T_{60}, we assessed robustness across a range of realistic acoustic scenarios.

All source signals were drawn from the CSTR VCTK corpus of male and female speech [17]. They were uniformly positioned on a horizontal plane at a 1 m radius around the array. We set $\alpha = 0.9$ and $P_{\max} = 5$. For time–frequency analysis, we applied a STFT with a 128 ms Hamming window and a 64 ms frame shift.

We compared the proposed method to two reference methods:

- *Baseline MVDR*, which assumes spatially white interference (i.e. uniform power across all source directions) with a fixed interference covariance matrix.
- *Oracle-PSD MVDR*, which employs time-frequency adaptive beamforming by constructing an interference covariance matrix in each TF bin from the true source PSDs.

Performance was evaluated using two objective metrics: signal-to-interference ratio improvement (SIR; see Appendix A) and short-time objective intelligibility (STOI) [18]. Improvements in these metrics – denoted as SIR improvement (SIRi) and STOI improvement (STOIi) – are computed as the difference between the beamformer's output and input values. Evaluations are conducted under both simulated anechoic and real-world reverberant conditions.

V. Experimental Results

A. Simulated anechoic conditions

Figures 2 and 3 show the SIRi and STOIi, respectively, as functions of the number of sources N for the array sizes $M \in \{2, 3, 4\}$ under ideal anechoic condition. Figure 2 demonstrates that the oracle-PSD MVDR establishes the ideal upper bound on SIRi in anechoic scenarios, achieving the highest interference suppression for every M and N. The baseline MVDR, attains only modest SIRi gains and degrades rapidly as the system becomes underdetermined ($N > M$). In contrast, the proposed method nearly attains the oracle performance. Figure 3 illustrates that the baseline MVDR's intelligibility improvements degrade sharply once it is underdetermined ($N > M$), reflecting its limited spatial degrees of freedom. By comparison, the proposed method degrades gradually, consistently closing over half the gap to the oracle and preserving speech intelligibility under underdetermined scenarios. These results demonstrate substantial gains in interference suppression (SIRi) and speech intelligibility (STOIi) under underdetermined anechoic conditions.

Figure 4 compares the spectrogram at each processing stage under the underdetermined scenario ($M = 3$, $N = 5$). The proposed method exhibits deeper nulls at time-frequency regions dominated by interfering sources than the baseline MVDR, reflecting enhanced frequency-domain selectivity in underdetermined scenarios.

B. Real-World Reverberant Experiments

To assess performance in realistic environment, we varied the reverberation time with fixed microphone of $M = 3$. Figures 5 and 6 plot SIRi and STOIi versus the number of sources N for each T_{60}.

Under mild reverberation ($T_{60} = 160$ ms), the results closely follow the trends seen in the simulated anechoic conditions. For $N = 6$, both the oracle-PSD MVDR and the proposed method achieve SIRi of 12.5 dB, whereas the baseline MVDR yields only of 7 dB. Notably, the baseline MVDR peaks near the determined case ($N = M = 3$), while the proposed method

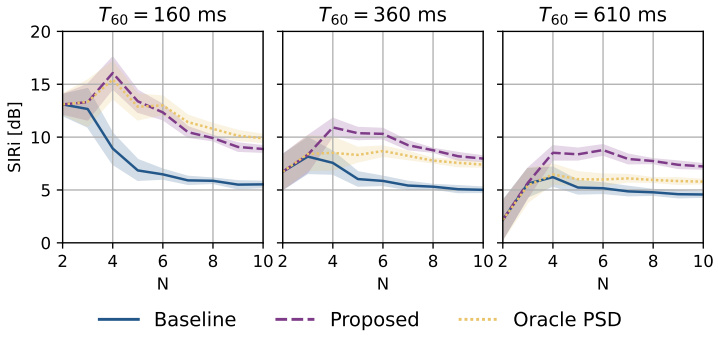

Fig. 5: Measured SIRi for $T_{60} \in \{160, 360, 610\}$ ms ($M = 3$); shaded areas show 95% CIs.

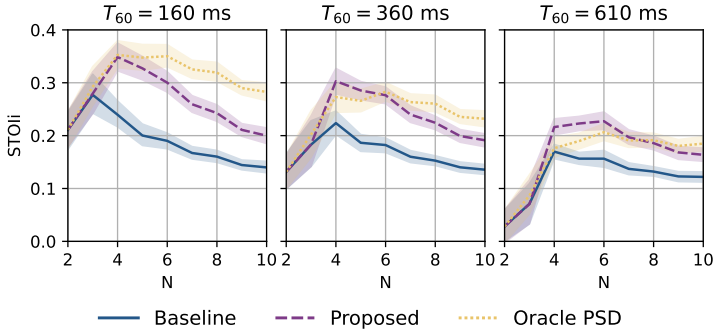

Fig. 6: Measured STOIi for $T_{60} \in \{160, 360, 610\}$ ms ($M = 3$); shaded areas show 95% CIs.

continues to improve beyond that point – reaching its maximum at $N = 4$ for ($T_{60} = 160$ ms) and between $N = 4$ and 6 for the more reverberant conditions ($T_{60} = 360$ ms and 610 ms). This behaviour highlights the advantage of the proposed method: by exploiting TF-domain sparsity and progressively updating interference PSDs, the algorithm effectively forms additional nulls to cancel extra interferers. As reverberation increases to $T_{60} = 360$ ms and 610 ms, all methods exhibit diminished gains, with their SIRi and STOIi curves shifting downward. In the most reverberant condition ($T_{60} = 610$ ms), the proposed method outperforms the oracle-PSD MVDR in SIRi, and in some cases, also in STOIi. This counterintuitive outcome arises because the oracle-PSD MVDR, despite using true per-source PSDs, assumes anechoic propagation for interference and ignores the diffuse reverberant field. In contrast, the proposed method directly estimates the directional PSD including reflections for each beamformer, thereby more accurately modelling and suppressing reverberation induced interference. Compared to the baseline method, the proposed algorithm maintains robust interference suppression when $N > M$ and degrades gracefully as T_{60} increases, thus preserving intelligibility improvements under realistic acoustic conditions.

The heatmaps in Figs. 7 and 8 depict the difference between the proposed and baseline methods (proposed minus baseline) in SIR and STOI, respectively. In particular, the SIR-difference heatmap (Fig. 7) shows that under underdetermined scenarios ($N > M$), the proposed method achieves interference suppression gains up to approximately 7 dB for $N = 5$, $M = 3$, and $T_{60} = 160$ ms. Similarly, the STOI-difference heatmap (Fig. 8) indicates intelligibility improvements up to 0.14 under the same condition, with appreciable gains persisting in more reverberant environments. These results confirm that in underdetermined real-world scenarios the proposed method consistently outperforms the baseline MVDR beamformer in both interference reduction and speech intelligibility; notably, it achieves these improvements while maintaining a distortionless response without relying on any auxiliary information.

VI. Conclusion

In this paper, we proposed a novel approach that integrates PSD estimation in beamspace and MVDR beamforming. By exploiting speech sparsity in the time-frequency domain, the proposed method enhances MVDR performance in underdetermined scenarios while maintaining the distortionless-response without the need for additional prior information. This distortionless property makes the approach particularly well suited to downstream applications such as automatic speech recognition and beneficial for hands-free audio applications. Both simulations and real-world experiments demonstrate that the proposed method consistently outperforms the baseline MVDR and closely approaches the SIR improvement of a time-frequency adaptive MVDR beamformer constructed with oracle source PSD. These results underscore the practical utility and robustness of the proposed method for underdetermined sound source separation in realistic acoustic environments.

Appendix

The SIR is defined as:

$$\mathrm{SIR} = 10 \log_{10} \frac{\sum_t x^2_{\mathrm{target}}(t)}{\sum_t x^2_{\mathrm{int}}(t)}, \tag{20}$$

where x_{target} is the target component and x_{int} is the interference component.

Acknowledgment

This research was partially funded by the Acoustics and Vibration Research Centre at the University of Auckland and the Kajima Foundation's Support Program for International Joint Research Activities (2024-kyodoshin-05).

References

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time–frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[2] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*, S. Makino, H. Sawada, and T.-W. Lee, Eds. Dordrecht: Springer Netherlands, 2007, pp. 217–241.

[3] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[4] J. Spriet, S. Doclo, M. Moonen, and J. Wouters, "Speech distortion weighted multichannel Wiener filtering techniques for noise reduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds., Springer, 2005, pp. 199–228.

[5] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2ch bss using the em algorithm in reverberant environment," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 147–150.

[6] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1240–1250, 2013.

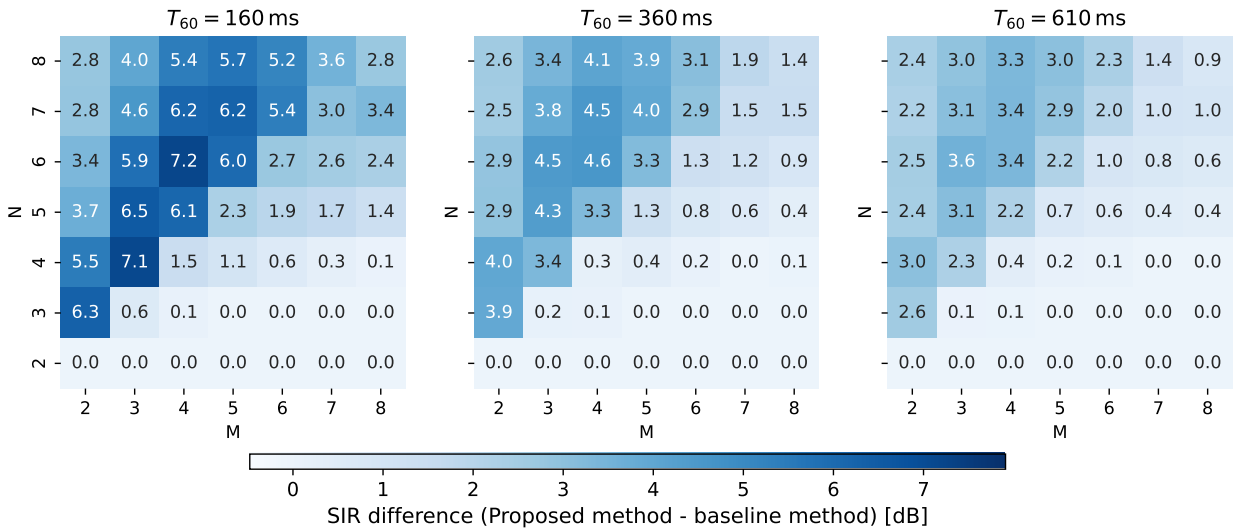

Fig. 7: Heatmap of SIR difference (proposed method minus baseline method) across number of sources N and number of microphones M under varying reverberation times $T_{60} \in \{160, 360, 610\}$ ms.

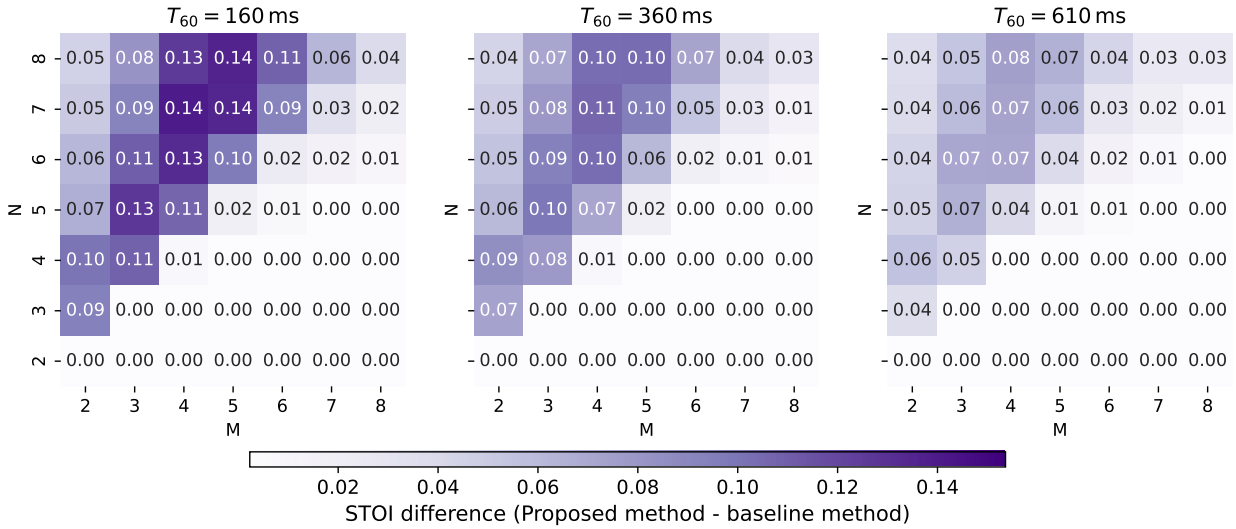

Fig. 8: Heatmap of STOI difference (proposed method minus baseline method) across number of sources N and number of microphones M under varying reverberation times $T_{60} \in \{160, 360, 610\}$ ms.

[7] R. Wang, T. Fujimura, and T. Toda, "Target speaker extraction under noisy underdetermined conditions using conditional variational autoencoder, global style token, and neural post-filter," *APSIPA Transactions on Signal and Information Processing*, vol. 14, e2, 2025.

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

[9] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "Adl-mvdr: All deep learning mvdr beamformer for target speech separation," *arXiv preprint arXiv:2008.06994*, 2020.

[10] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[11] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[12] J. Li and P. Stoica, *Robust adaptive beamforming* (Wiley Series in Telecommunications and Signal Processing ; v.88), eng. Hoboken, NJ: John Wiley, 2006.

[13] Y. Kubo, T. Nakatani, M. Delcroix, K. Kinoshita, and S. Araki, "Mask-based MVDR beamformer for noisy multisource environments: Introduction of time-varying spatial covariance model," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK: IEEE, May 2019, pp. 6855–6859.

[14] K. Yamaoka, N. Ono, and S. Makino, "Time-frequency-bin-wise linear combination of beamformers for distortionless signal enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3461–3475, Nov. 2021.

[15] M. Brandstein and D. Ward, *Microphone Arrays*. Springer Berlin Heidelberg, 2001.

[16] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC 2014)*, Antibes–Juan-les-Pins, France, Sep. 2014.

[17] J. Yamagishi, V. Christophe, and M. Kirsten, *CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)*, Sound, Data retrieved from Edinburgh DataShare, Nov. 2019.

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