Cross-Talk Cancellation for Close-Miked String Duos via STFT MLE Calibration

Roger Esteve, Alejandro Alzina, Guerau Orus
DSAP course project. Autumn 2024-25. Masters MET&MATT,
Barcelona School of Telecommunication Engineering (ETSETB), UPC

***Abstract*—We study microphone bleed suppression for close-miked ensemble recordings using a lightweight calibration pipeline. The system targets string duos (violin and cello) recorded in a virtual 5 m × 5 m room, following the setup of Das *et al.* (AES 2021). Three algorithms are compared: a blind Wiener-style interference canceller, a time-domain regularized least-squares (RLS) calibration with alternating refinements, and a frequency-domain maximum-likelihood estimator (MLE) in the STFT domain. Calibration uses 10 s of alternating solos; performance is evaluated on a joint performance segment. Objective metrics include PEASS (OPS/TPS/IPS/APS) and BSS Eval (SDR/SIR/SAR). On simulated data with microphone–source distances from 0.1 m to 0.5 m, the STFT MLE yields OPS > 98 and SDR up to 38 dB at 0.1 m, degrading gracefully with distance. The code is incomplete for real recordings; we outline remaining steps to reach a deployable toolchain.**

I. Introduction

Close-microphone bleed complicates live mixing, rehearsal feedback, and source separation in small ensembles. When each player has a dedicated microphone, even modest leakage degrades downstream effects and monitoring. We address two-channel bleed reduction with minimal assumptions and light computation, aiming for a practical path from simulation to rehearsal-stage deployment.

We adopt the virtual studio EnsembleSet data (BBCSO) and simulate room acoustics with pyroomacoustics to reproduce the conditions of Das *et al.* [1]. Our contributions are:

- A reproducible data generator that matches the AES 2021 microphone geometry (5 m room, two sources 1 m apart, mic–source distances 0.1–0.5 m) with 10 s calibration solos and held-out performance audio.
- An RLS calibration baseline with alternating updates for sources and mixing, plus a blind Wiener interference canceller for comparison.
- A frequency-domain MLE (STFT) implementation that follows the trust-region formulation of [1], delivering strong perceptual scores on simulated duos.
- An evaluation harness computing PEASS and BSS Eval metrics and exporting results to CSV for sweep analysis.

II. Techniques

A. Signal Model

We assume an instantaneous mixture for two sources and two microphones:

$$\mathbf{X}(t) = \mathbf{H}\,\mathbf{S}(t) + \mathbf{W}(t), \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{2\times T}$ are microphone signals, $\mathbf{S}$ are sources, $\mathbf{H}$ is the mixing/cross-talk matrix, and $\mathbf{W}$ is noise. In the STFT domain, $\mathbf{H}$ becomes frequency-dependent, $\mathbf{H}(\omega)$.

B. Blind Wiener Interference Canceller

Following [2], we treat the non-diagonal mic as interference and estimate scalar filters that minimize output power:

$$\hat{w}_{12} = \frac{\mathbb{E}[x_1 x_2]}{\mathbb{E}[x_2^2]}, \quad \hat{s}_1 = x_1 - \hat{w}_{12} x_2, \tag{2}$$

and symmetrically for $\hat{s}_2$. This requires no calibration but assumes low leakage and uncorrelated sources.

C. Time-Domain Regularized LS Calibration

With short calibration solos, we estimate $\mathbf{H}$ via ridge regression:

$$\hat{\mathbf{H}} = \left(\mathbf{X}\mathbf{S}^\top + \lambda \mathbf{H}_0\right)\left(\mathbf{S}\mathbf{S}^\top + \lambda \mathbf{I}\right)^{-1}, \tag{3}$$

reducing to standard RLS when the prior $\mathbf{H}_0$ is absent. An alternating scheme refines $\mathbf{S}$ and $\mathbf{H}$ (ALS) with early stopping on reconstruction cost. Noise variance is estimated from residuals to report SNR and conditioning diagnostics.

D. STFT-Domain MLE (Das et al.)

The core system follows [1]: (i) compute STFTs of calibration solos; (ii) derive a prior $\tilde{\mathbf{H}}(\omega)$ from spectral ratios when one source dominates each time frame; (iii) per frequency bin, solve

$$\begin{aligned}\min_{\mathbf{H},\mathbf{S}} \; & \|\mathbf{X} - \mathbf{H}\mathbf{S}\|_2^2 \\ & + \lambda \|\mathbf{H} - \tilde{\mathbf{H}}\|_2^2, \end{aligned} \tag{4}$$

using alternating updates with Hermitian solves. Inference inverts $\hat{\mathbf{H}}(\omega)$ per bin and applies an iSTFT. Active-frame masks are known in simulation (solo segments); automatic energy-based masks are available for real recordings.

III. Experiments and Discussion

A. Data and Setup

Audio stems come from the BBC Symphony Orchestra “Misero Pargoletto” excerpt (violin and cello spot mics). We simulate a $5 \times 5 \times 3$ m room with pyroomacoustics; anechoic mode (max_order = 0) matches the AES setup. Sources are 1 m apart; mic–source distance d_{mic} is swept from 0.1 to 0.5 m. Calibration uses 10 s of alternating solos; the remaining

TABLE I
STFT MLE (v1) PERFORMANCE VS. MIC DISTANCE.

d_{mic} (m)	OPS	TPS	IPS	APS	SDR (dB)	SIR (dB)	SAR (dB)
0.1	98.9	92.8	96.8	86.7	37.96	∞	37.96
0.2	98.4	90.6	95.1	82.9	33.06	∞	33.06
0.3	83.6	79.3	75.6	62.0	29.75	∞	29.75
0.4	56.2	64.6	67.9	21.8	27.73	∞	27.73
0.5	46.0	66.8	62.3	23.4	25.79	∞	25.79

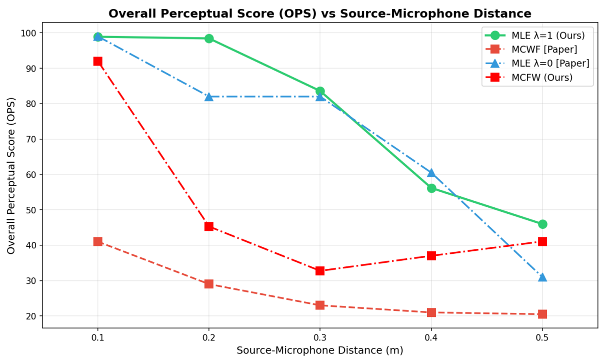

Fig. 1. PEASS OPS across microphone distances (simulated).

≥10 s constitutes the performance mixture. Sampling rate is inherited from the stems (44.1 kHz). Outputs follow the ICASSP evaluation recipe:

- **PEASS** (`pyass`): OPS, TPS, IPS, APS on 1 s and 30 s excerpts.
- **BSS Eval** (`mir_eval`): SDR, SIR, SAR on the recovered violin against its clean reference.

Results are logged to `run/results.csv`; figures `ops_comparison.png` and `sdr_comparison.png` summarize the sweep.

B. Quantitative Results

Table I reports the STFT MLE performance across mic distances. OPS and TPS stay > 60 even at 0.5 m, while SDR degrades from 38 dB to 26 dB. SIR is numerically unbounded (∞) in the anechoic simulation, indicating near-complete interference rejection; in real rooms this will be finite.

Figure 1 visualizes OPS trends; Figure 2 shows SDR decay with distance. Both emphasize strong performance at 0.1–0.2 m and graceful degradation thereafter.

C. Ablations and Observations

Calibration length. The 10 s solos suffice for stable $\tilde{\mathbf{H}}$ estimates; shorter excerpts increase conditioning issues.
Regularization. $\lambda = 0.01$ in the STFT solver balances fidelity and stability; larger values underfit high-frequency leakage.
Blind Wiener baseline. Works only for mild bleed (close to diagonal $\mathbf{H}$) and fails when delays or stronger cross-talk appear, confirming the need for calibrated inversion.
Limitations. Experiments are simulation-only; real recordings will introduce room modes, time-varying balance, and synchronization drift. Active-mask estimation on real calibration takes energy heuristics and may mis-detect overlaps; robustness remains to be validated.

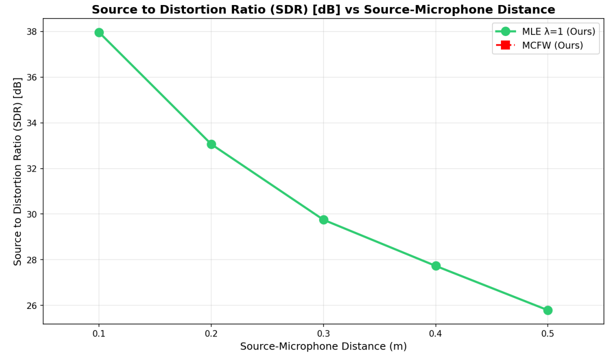

Fig. 2. SDR trend for the recovered violin vs. mic distance.

IV. Conclusions

We reproduced the STFT-domain MLE of Das *et al.* for two-channel bleed reduction and built a simulation/evaluation pipeline grounded in BBCSO stems. The approach achieves high perceptual quality at practical close-mic distances and provides a structured path to field tests. Remaining work includes: (i) validating on real duo recordings; (ii) handling more than two sources/mics; (iii) adding mild reverberation and time-delay modeling; and (iv) benchmarking against deep learning baselines. The current codebase nonetheless offers a strong starting point for DSAP course deployment.

References

[1] A. Das, M. J. Murphy, and D. Dorran, “Microphone bleed reduction in close-microphone applications,” in *AES Conf.*, 2021.
[2] A. Kokkinis, M. Poulos, and J. Kanellopoulos, “A Wiener filter approach to microphone leakage reduction in close-microphone applications,” *J. Audio Eng. Soc.*, vol. 60, no. 9, pp. 706–718, 2012.
[3] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.
[4] V. Emiya *et al.*, “Subjective and objective quality assessment of audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2046–2057, 2011.