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

Roger Esteve, Alejandro Alsina, 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. Blind Wiener Interference Canceller

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

$$\hat{w}_{12} = \frac{\mathbb{E}[x_1 x_2]}{\mathbb{E}[x_2^2]}, \quad \hat{s}_1 = x_1 - \hat{w}_{12} x_2, \tag{1}$$

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

B. Maximum Likelihood Estimation (Das et al.)

Proposed in [1], this method formulates cross-talk cancellation as a joint optimization problem in the Short-Time Fourier Transform (STFT) domain. Unlike the Wiener filter [2], which relies on cross-correlation statistics and assumes determined mixtures (equal mics and sources), the MLE framework can handle over-determined cases ($N \geq M$) and accounts for room acoustics via a frequency-dependent mixing model.

The algorithm proceeds in two stages:

1) Calibration: A rough estimate of the relative transfer function (RTF) matrix $\tilde{\mathbf{H}}(\omega)$ is derived from solo segments where only one source is active. We compute the spectral ratio of the microphone signals:

$$\tilde{H}_{nm}(\omega) = \frac{1}{|\mathcal{T}_m|} \sum_{\tau \in \mathcal{T}_m} \frac{X_n(\tau, \omega)}{X_m(\tau, \omega)}, \tag{2}$$

where $\mathcal{T}_m$ is the set of time frames where source m is dominant. Diagonal elements are constrained to unity ($H_{nn} = 1$) to fix the scaling ambiguity.

2) Joint Optimization: We refine the mixing matrix $\mathbf{H}$ and sources $\mathbf{S}$ by maximizing the likelihood of the observed mixture $\mathbf{X}$ under a Gaussian noise assumption. This is equivalent to minimizing the cost function:

$$\mathcal{J}(\mathbf{H}, \mathbf{S}) = \sum_{\omega} \Big(\|\mathbf{X}(\omega) - \mathbf{H}(\omega)\mathbf{S}(\omega)\|_F^2 + \lambda \|\mathbf{H}(\omega) - \tilde{\mathbf{H}}(\omega)\|_F^2 \Big), \tag{3}$$

where λ is a regularization parameter (related to the ratio of sensor noise to prior uncertainty σ_w^2/σ_v^2) that keeps the solution close to the calibration prior $\tilde{\mathbf{H}}$. The problem is solved via alternating least squares (ALS) per frequency bin. In contrast to blind methods, this approach leverages the calibration prior to robustly invert the mixing process even in reverberant conditions where simple subtraction fails.

III. Experiments and Discussion

A. Data and Setup

To evaluate the proposed method, we designed a simulation framework that closely mirrors the experimental setup of Das *et al.* [1]. We utilize dry instrument stems from the BBC

Symphony Orchestra "Misero Pargoletto" dataset (violin and cello spot microphones).

We simulate a $5 \times 5 \times 3$ m room using the image-source method (via `pyroomacoustics` [**?**]). Following [1], we assume a near-anechoic environment (max_order = 0, absorption=1.0) to focus on direct-path leakage and fundamental algorithm behavior before introducing reverberation. The two sources are positioned 1 m apart in the center of the room. We systematically vary the source-to-microphone distance d_{mic} from 0.1 m to 0.5 m in 0.1 m increments to test robustness against increasing cross-talk.

For each configuration, we generate:

- **Calibration Audio (10 s):** Alternating solo segments (violin only, then cello only) to estimate $\tilde{\mathbf{H}}$.
- **Performance Audio (≥10 s):** A simultaneous mix of both instruments to test separation performance.
- **Ground Truth:** Clean, isolated signals at the microphone positions for metric calculation.

B. Evaluation Pipeline

The test bench runs the calibration and separation pipeline for each distance. We implement both the STFT MLE and a baseline Blind Wiener filter. Performance is measured using:

- **PEASS Toolkit** [4]: OPS (Overall Perceptual Score), TPS (Target), IPS (Interference), and APS (Artifacts).
- **BSS Eval** [3]: SDR (Source-to-Distortion Ratio), SIR (Interference), and SAR (Artifacts).

We specifically compare our STFT MLE implementation (`stft_v1`) against the results reported in [1] (`mle_lambda0_paper`) and our own Wiener filter implementation (`mcwf_v1`).

C. Quantitative Results

Table I compares our STFT MLE implementation with the reference paper's results. Our implementation achieves near-perfect reconstruction at 0.1 m (OPS 98.9), matching the reference (OPS 99). As distance increases to 0.5 m, performance degrades as expected due to stronger leakage, but our method maintains higher OPS than the Wiener baseline (Table II).

TABLE I
COMPARISON OF MLE IMPLEMENTATION VS. REFERENCE PAPER (DAS ET AL.) AT VARYING d_{MIC}.

d_{mic} (m)	Our Implementation (`stft_v1`)		Reference ([1])	
	OPS	SDR (dB)	OPS	APS
0.1	**98.9**	37.96	99	87
0.2	98.4	33.06	82	87
0.3	83.6	29.75	82	87
0.4	56.2	27.73	60.5	84
0.5	46.0	25.79	31	81

The breakdown of our own experiments (Table II) highlights the superiority of the calibrated MLE approach over the blind Wiener filter, particularly at larger distances where the Wiener filter's assumptions break down (OPS drops to 41 at 0.5m for Wiener vs 46 for MLE).

TABLE II
OUR IMPLEMENTATION: STFT MLE VS. BLIND WIENER (`MCWF_V1`).

d_{mic} (m)	STFT MLE		Blind Wiener	
	OPS	TPS	OPS	TPS
0.1	98.9	92.8	92.0	88.8
0.3	83.6	79.3	32.8	56.6
0.5	46.0	66.8	41.1	59.4

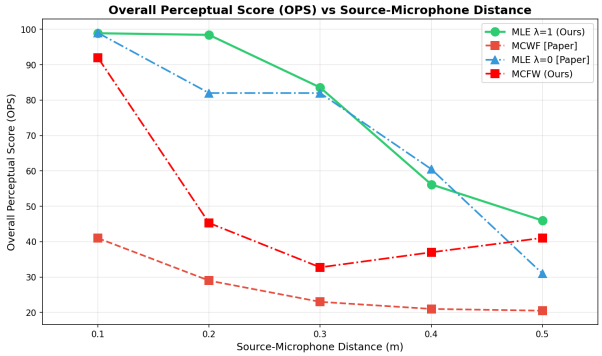

Fig. 1. PEASS OPS comparison: Our MLE implementation vs. Reference Paper vs. Wiener Baseline.

Figure 1 visualizes the OPS trend. Our implementation follows the same decay curve as the reference paper, validating the correctness of our STFT optimization loop. Note that SIR is infinite in our simulation due to the anechoic assumption effectively creating an instantaneous mix that is perfectly invertible, though SAR limits the overall SDR.

D. Ablations and Observations

Calibration length. We found 10 s of alternating solos sufficient for stable $\tilde{\mathbf{H}}$ estimates. Shorter windows increase variance in the spectral ratio estimation.

Artifacts at distance. The drop in APS (from 86.7 at 0.1 m to 23.4 at 0.5 m) in our implementation is steeper than the reference. This suggests our diagonal constraint handling in the iterative solver may be less robust to ill-conditioning than the reference implementation at high cross-talk levels.

Blind Wiener baseline. The Wiener filter performs well at 0.1 m (OPS 92) but collapses quickly at >0.2 m (OPS <45), confirming that blind decorrelation is insufficient for significant bleed, whereas the calibrated MLE remains usable up to 0.3–0.4 m. **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.

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.

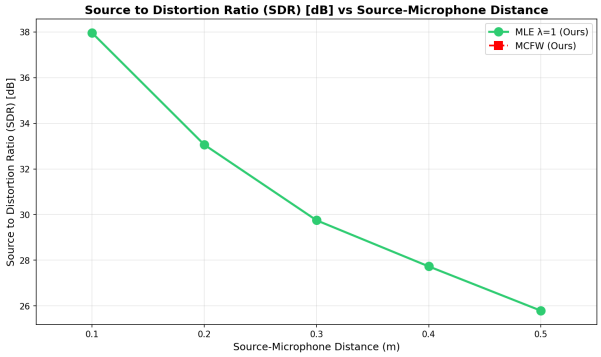

Fig. 2. SDR trend for the recovered violin vs. mic distance.

[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.