Audio Engineering Society

Convention Paper

Presented at the 150th Convention
2021 May 25–28, Online

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Microphone Cross-talk Cancellation in Ensemble Recordings with Maximum Likelihood Estimation

Orchisama Das[1], Julius O. Smith[1], and Jonathan S. Abel[1]

[1] *Center for Computer Research in Music and Acoustics, Stanford University*

Correspondence should be addressed to Orchisama Das (`orchi@ccrma.stanford.edu`)

ABSTRACT

While recording an ensemble of musicians, it is often desired to isolate the instruments to avoid interference from other sources. Close-miking and acoustic isolation booths are some techniques for mitigating microphone cross-talk (or “bleed"). In this paper, we propose an algorithm for canceling microphone bleed in ensemble recordings in the post-processing stage. We propose a calibration stage to estimate the relative transfer function from each instrument to each mic. Then, we set up an optimization problem to simultaneously estimate the transfer functions and individual sources given the microphone signals and a noisy estimate of the transfer function obtained from the calibration stage. We show that minimizing this cost function gives us the maximum likelihood estimate when we assume the distributions to be normal. Finally, we test our proposed method to cancel microphone bleed in a synthesized environment, and compare our results to an existing multichannel Wiener filter method.

1 Introduction

Microphone cross-talk, or “bleed", has been a nuisance in the audio engineering community for decades. When two microphones pick up the same signal with a time delay, comb filtering artifacts are present. Furthermore, when dynamic, non-linear effects like compression are applied, the bleed from other instruments becomes even more prominent. An expensive solution to the microphone bleed problem is to use acoustic isolation panels between musicians. For example, drums can be recorded in a separate isolation booth (acoustically treated soundproof room). Obviously, this is not feasible in a live setting. Simpler solutions include using directional microphones with specific polar patterns to pick up radiation from a desired direction, and using the close-miking technique where microphones are placed at a distance of $5-50$ cm from the sound source [1]. Interference can become significant in such cases due to the effect of room acoustics and nearby strong reflective surfaces [2]. The complexity lies in the fact that there is usually an arbitrary number and distribution of instruments and microphones, and results are influenced by the radiation pattern of the instrument and room acoustics of the studio where the ensemble is recorded.

In recent years, the music research community has come up with novel solutions to this problem, which adds to a rich, existing body of literature on audio source separation [3, 4, 5, 6]. Some of these solutions are kernel additive modeling [7], cross-talk resistant adaptive noise canceler [8], inverse filtering [9] and non-negative signal factorization [10]. One of the most successful candidates is the Multi-Channel Wiener Filter (MCWF) proposed by Kokkinis et al. [11, 12]. However, this method only works in the determined case, i.e, when the number of microphones is equal to the number of sources, assumes close-miking, and calculates the power spectral density (PSD) of the interfering signal by assuming a gain and delay factor to represent the acoustic path, which is an oversimplification in rooms with strong reflective surfaces that produce prominent early reflections. An improvement to the PSD estimation technique using a Kalman filter was proposed in [13].

In this paper, we propose a method to cancel microphone bleed that can work in a close-microphone set up, having arbitrary number of sources, M, and mics, N, as long as $N \geq M$. First, in Section 3, we propose a calibration stage, where one source is played at a time and recorded by all the microphones, which is used to estimate an acoustic transfer function matrix, that represents the relative acoustic transfer function from each source to each mic. In Section 4, we come up with a cost function in the time-frequency domain that simultaneously optimizes the sources and the acoustic transfer function matrix, which we assume to be time-invariant. We show that minimizing this cost function gives us the maximum likelihood estimate when we maximize the joint likelihood of the measured microphone signals and estimated acoustic transfer functions over all time frames, and individually for each frequency bin. In Section 5, we evaluate the method against MCWF on a dataset of string quartet recordings for varying source-microphone distance, number of mics and number of sources. We discuss the results and scope for future work in Section 6 and conclude the paper in Section 7.

2 Model

For N mic signals $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]^\top$ and M sources $\boldsymbol{s} = [s_1, s_2, \ldots, s_M]^\top$, the n^{th} mic signal at time index k can be written as

$$x_n(k) = \sum_{m=1}^{M} s_m(k) * h_{nm}(k) + w(k) \tag{1}$$

where h_{nm} is the acoustic transfer function between the m^{th} source and the n^{th} mic, and $w(k)$ is additive noise, $w(k) \sim \mathcal{N}(0, \sigma_w^2)\ \forall\ k$, caused by microphone "self-noise". The acoustic transfer function contains the direct sound path, as well as frequency-dependent source radiation pattern, microphone directivity function and room acoustics. An example of this configuration is shown in Fig. 1. In the time-frequency domain, the convolution of the sources with the transfer function becomes multiplication of a time-invariant relative transfer function (RTF) matrix with the source vector for each time frame, τ, and frequency bin, ω.

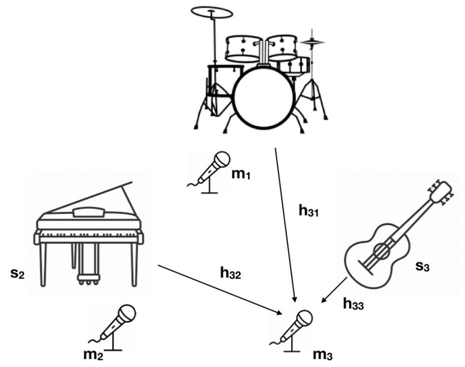


Fig. 1: Example of a studio setup

$$\boldsymbol{x}_\tau(\omega) = \boldsymbol{H}(\omega)\boldsymbol{s}_\tau(\omega) + \boldsymbol{w}_\tau(\omega)$$
$$\boldsymbol{H}(\omega) = \begin{bmatrix} H_{11}(\omega) & H_{12}(\omega) & \ldots & H_{1M}(\omega) \\ H_{21}(\omega) & H_{22}(\omega) & \ldots & H_{2M}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ H_{N1}(\omega) & H_{N2}(\omega) & \ldots & H_{NM}(\omega) \end{bmatrix} \tag{2}$$

Although in reality, the acoustic transfer function is not time-invariant due to changes in temperature and pressure, for all practical purposes, the time variation is slow enough to make such an assumption. Similarly, movements made by the musicians also affect the acoustic path, but these are usually small compared to the mean free path traveled by sound waves. On the other hand, assuming the transfer function to be time-invariant has the advantage of significantly reducing the number of unknowns to be estimated, and gives a smoothly varying RTF.

We want to estimate the source vector $\boldsymbol{s}_\tau(\omega)$ given the microphone signals, $\boldsymbol{x}_\tau(\omega)$. However, it is clear that without knowing $\boldsymbol{H}(\omega)$, we cannot solve this system of equations. Assuming we have some knowledge of the transfer function matrix, i.e., a noisy estimate of the transfer function matrix with each element in the noise matrix $\boldsymbol{\nu}$ independent and identically distributed with the same mean and variance,

$$\widetilde{\boldsymbol{H}}(\omega) = \boldsymbol{H}(\omega) + \boldsymbol{\nu}(\omega);\ \boldsymbol{\nu} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_\nu^2)\ \forall\ \omega \tag{3}$$

our goal is to jointly estimate the optimal values $\boldsymbol{H}^*(\omega)$, $\boldsymbol{s}^*_\tau(\omega)$ given $\widetilde{\boldsymbol{H}}(\omega), \boldsymbol{x}^*_\tau(\omega)$. The two noise terms, $\boldsymbol{w}$ and $\boldsymbol{v}$, are assumed to be uncorrelated, since they arise from different physical phenomena.

3 Calibration

To calculate $\widetilde{\boldsymbol{H}}(\omega)$, we propose a calibration stage. When all the microphones in the studio have been set up, a sound-check can be performed where one instrument is active at a time and picked up by all the microphones. For example, while recording drums, each drum part can be struck separately and captured simultaneously by all the mics. In the close-microphone case, the sound captured by the mic closest to the m^{th} source can be approximated as the source itself, and its spectral ratio with the sound captured by the n^{th}e mic should give an approximate value of $\widetilde{H}_{nm}(\omega)$. We average the spectral ratios of frames that have significant energy to get the off-diagonal elements, $\widetilde{H}_{nm}(\omega)$. The diagonal elements, $\widetilde{H}_{nn}(\omega)$, are approximated to be 1. For the m^{th} microphone signal,

$$\begin{aligned}
\mathrm{sig}_m &= \left[\tau \in 1,\ldots,T : E_{\tau,m} \geq \frac{\max(E_{1,\cdots,T,m})}{\sqrt{2}}\right] \text{ where} \\
E_{\tau,m} &= \frac{1}{L}\sum_{l=0}^{L-1} |x_m(\tau L + l)|^2 \\
\tilde{H}_{nm}(\omega) &= \begin{cases} 1 \text{ if } n = m \\ \frac{1}{N_{\mathrm{sig}}} \sum \tau \in \mathrm{sig}_m \frac{x_{n\tau}(\omega)}{x_{m\tau}(\omega)} \text{ if } n \neq m \end{cases}
\end{aligned} \tag{4}$$

If calibration in this way is not possible, an alternative is to detect solo excerpts of the m^{th} source in the piece that has been recorded [12], and calculate $\widetilde{h}_{nm}(\omega)$ as the spectral ratio averaged over all time frames where only source m is active.

4 Maximum Likelihood Estimate

4.1 Deriving the cost function

Let $\widetilde{\boldsymbol{h}}(\omega) = [\widetilde{H}_{11}, \widetilde{H}_{12}, \ldots, \widetilde{H}_{NM}]^\top$. We can maximize the joint likelihood of the measured microphone signals and estimated acoustic transfer functions conditioned on the source signals and the actual transfer function, over all T frames, and individually for each frequency bin. In the following derivations, the frequency index ω has been omitted for clarity, and the subscript has been used to denote a time frame.

$$\begin{aligned}
J(\boldsymbol{s_1},\ldots,\boldsymbol{s_T},\boldsymbol{h}) &= \max p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T,\widetilde{\boldsymbol{h}}|\boldsymbol{s}_1,\ldots,\boldsymbol{s}_T,\boldsymbol{h}) \\
&= \max p(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T|\boldsymbol{s}_1,\ldots,\boldsymbol{s}_T,\boldsymbol{h})\, p(\widetilde{\boldsymbol{h}}|\boldsymbol{s}_1,\ldots,\boldsymbol{s}_T,\boldsymbol{h})
\end{aligned} \tag{5}$$

We can assume $\boldsymbol{x}_1, \ldots \boldsymbol{x}_T$ to be independent and identically distributed and the microphone signal at current frame, $\boldsymbol{x}_\tau$, to depend only on the source signal at current frame, $\boldsymbol{s}_\tau$. Similarly, the measurement of $\widetilde{\boldsymbol{h}}$ is independent of $\boldsymbol{s}_\tau \forall\ \tau$. Then, we can maximize the joint likelihood, or equivalently, minimize negative log of the joint likelihood.

$$\begin{aligned}
J(\boldsymbol{s_1},\ldots,\boldsymbol{s_T},\boldsymbol{h}) &= \max \left(\prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{s}_t,\boldsymbol{h})\right) p(\widetilde{\boldsymbol{h}}|\boldsymbol{h}), \\
&= \min \left[-\sum_{t=1}^{T} \ln p(\boldsymbol{x}_t|\boldsymbol{s}_t,\boldsymbol{h}) - \ln p(\widetilde{\boldsymbol{h}}|\boldsymbol{h})\right]
\end{aligned} \tag{6}$$

Since we have assumed normally distributed measurement noise, the above distributions are normal with the following parameters : $\boldsymbol{x}_t|\boldsymbol{s}_t,\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{H}\boldsymbol{s}_t, \sigma_w^2\boldsymbol{I})$ and $\widetilde{\boldsymbol{h}}|\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{h}, \sigma_v^2\boldsymbol{I})$. The negative log likelihood function is

$$\begin{aligned}
J(\boldsymbol{s_1},\ldots,\boldsymbol{s_T},\boldsymbol{h}) &= \max \left[\exp\left(\frac{-\sum_{t=1}^{T}(\boldsymbol{x}_t - \boldsymbol{H}\boldsymbol{s}_t)^{\mathsf{H}}(\boldsymbol{x}_t - \boldsymbol{H}\boldsymbol{s}_t)}{2\sigma_w^2}\right)\cdot\right. \\
&\left.\exp\left(\frac{-(\widetilde{\boldsymbol{h}} - \boldsymbol{h})^{\mathsf{H}}(\widetilde{\boldsymbol{h}} - \boldsymbol{h})}{2\sigma_v^2}\right)\frac{1}{\sqrt{(2\pi)^{T+1}\sigma_w^2\sigma_v^2}}\right] \\
&= \min \frac{1}{\sigma_w^2}\sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{H}\boldsymbol{s}_t||^2 + \frac{1}{\sigma_v^2}||\widetilde{\boldsymbol{h}} - \boldsymbol{h}||^2 \\
&= \min \sum_{t=1}^{T} ||\boldsymbol{x}_t - \boldsymbol{H}\boldsymbol{s}_t||^2 + \frac{\sigma_w^2}{\sigma_v^2}\mathrm{Tr}\left((\widetilde{\boldsymbol{H}} - \boldsymbol{H})^{\mathsf{H}}(\widetilde{\boldsymbol{H}} - \boldsymbol{H})\right)
\end{aligned} \tag{7}$$

where $\mathrm{Tr}(\cdot)$ is trace of a matrix, and $(\cdot)^{\mathsf{H}}$ is the Hermitian transpose. This cost function looks similar to a least squares formulation with ℓ_2-norm regularization. However, in this case, both $\boldsymbol{H}$ and $\boldsymbol{s}_\tau$ are unknown.

4.2 Optimal Solution

It can be verified that this cost function is convex in $\boldsymbol{H}$ and $\boldsymbol{s}_\tau$. Therefore, the optimal values $\boldsymbol{H}^*, \boldsymbol{s}_\tau^*$ can be found by calculating the gradient of the cost function and setting it to zero, thus giving the following solution,

$$\begin{aligned}\boldsymbol{H}^* &= \left(\widetilde{\boldsymbol{H}} + \frac{\sigma_v^2}{\sigma_w^2}\sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{s}_t^{\mathsf{H}}\right)\left(\boldsymbol{I} + \frac{\sigma_v^2}{\sigma_w^2}\sum_{t=1}^{T} \boldsymbol{s}_t \boldsymbol{s}_t^{\mathsf{H}}\right)^{-1} \\ \boldsymbol{s}_\tau^* &= (\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{H}^*)^{-1}\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{x}_\tau\end{aligned} \tag{8}$$

where $\boldsymbol{I}$ is an $M \times M$ identity matrix. It is to be noted that σ_v^2/σ_w^2 acts as a hyperparameter. A small value of σ_v^2/σ_w^2 gives more weight to the initial estimate of the transfer function, $\widetilde{\boldsymbol{H}}$.

Equation (8) is a function of $\boldsymbol{s}_\tau$ and can be solved using any numerical root finder, such as `MATLAB`'s `fsolve` [14]. Since the computation is independent over frequency bins, the optimization can be parallelized. The computations can further be sped up by replacing summation over vectors with matrix operations. More specifically, we can write the time-varying microphone signals as an $N \times T$ matrix $\boldsymbol{X}$, and the source signals as an $M \times T$ matrix $\boldsymbol{S}$, with time frames along the columns. The operations in (8) then become

$$\begin{aligned}\boldsymbol{H}^* &= \left(\widetilde{\boldsymbol{H}} + \frac{\sigma_v^2}{\sigma_w^2}\boldsymbol{X}\boldsymbol{S}^{\mathsf{H}}\right)\left(\boldsymbol{I} + \frac{\sigma_v^2}{\sigma_w^2}\boldsymbol{S}\boldsymbol{S}^{\mathsf{H}}\right)^{-1} \\ \boldsymbol{S}^* &= (\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{H}^*)^{-1}\boldsymbol{H}^{*\mathsf{H}}\boldsymbol{X}\end{aligned} \tag{9}$$

5 Experiments and Results

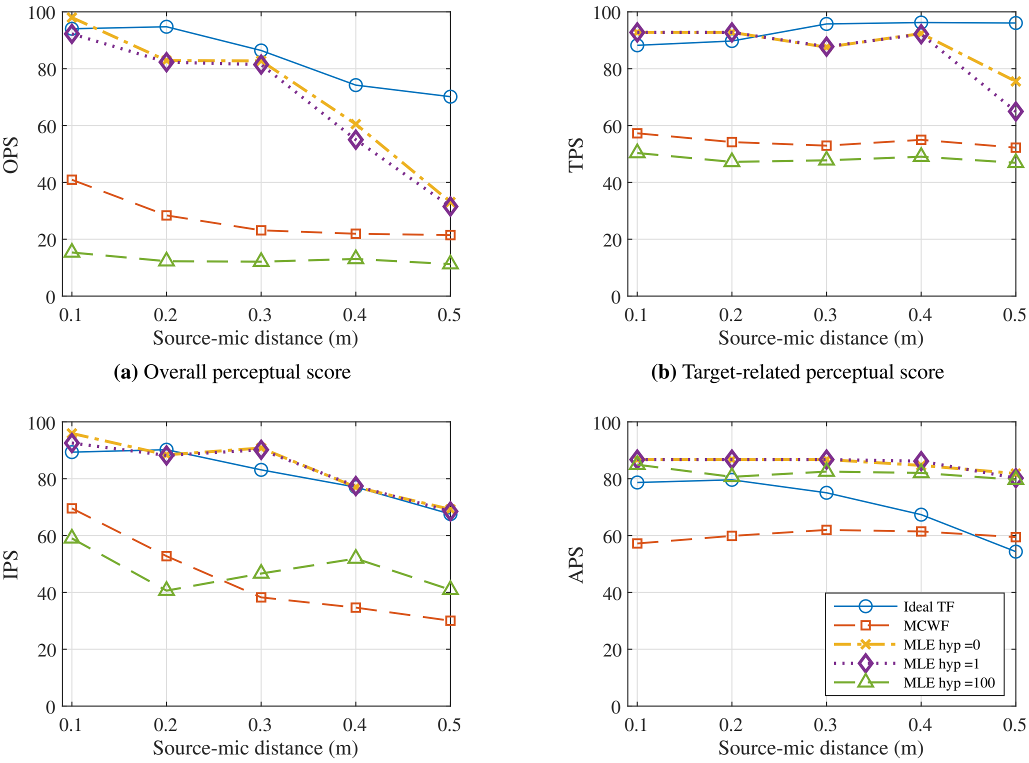
For a realistic recording scenario, we used a dataset of anechoic string quartet recordings from TU Berlin [15]. We placed the instruments in a virtual shoebox room of dimensions $3 \times 4 \times 3.25$ m^3. For the following analysis, we primarily worked with two instruments – the viola (Va) was placed at $(1.9, 2.5, 1.0)$ m and the violoncello (Vcl) was placed at $(1.7, 2.8, 0.8)$ m. Omnidirectional microphones were placed directly in front of the instruments on the same axis. RIR Generator [16] was used to simulate the room acoustics with the image-source method [17]. The reverberation time (RT_{60}) was fixed to be 0.8 s and the length of the impulse response generated was 128 samples at a sampling rate of 48 kHz. The short length of the impulse response is adequate to capture the early reflections, which primarily affect the cross-talk between microphones. The impulse response for each source-microphone pair was convolved with the anechoic recordings to generate the captured microphone signals. While the image-source method is not full-proof, it simulates microphone cross-talk adequately since it reconstructs the early-reflections of a shoebox room correctly.

For the Short-Time Fourier Transform (STFT), we used a frame size of 1024 samples with a Hann window, a hop size of 512 samples, and FFT length of 4096 samples. We tested our method with three different values of the hyperparameter, as well as with the known transfer function. A hyperparameter value of zero essentially gives $\boldsymbol{H}^* = \widetilde{\boldsymbol{H}}$, and the separated source is the least squares solution. As the hyperparameter value increases, $\boldsymbol{H}^*$ deviates further from $\widetilde{\boldsymbol{H}}$. The MCWF parameters used are the same as in [12], but the frame size and hop size are kept same as the proposed maximum likelihood estimator (MLE). For estimation of $\widetilde{\boldsymbol{H}}$ with the spectral ratio method, a different 2 s excerpt from the same piece was used.

The evaluation was done with the PEASS toolbox [18] which gives perceptually motivated scores for the separated sources. The mean scores (averaged over all instruments) are reported in this paper. Overall perceptual score (OPS) gives an indication of how close the separated signal is to the target signal. Target-related perceptual score (TPS) gives an indication of the target distortion. Interference-related perceptual score (IPS) is an indication of how much interference has been eliminated from the separated signal, and Artifact-related perceptual score (APS) is an indication of the artifacts in the separated signal. These scores were chosen over the more commonly reported objective scores of SDR, SAR and SIR from the BSS toolbox [19], due to their higher correlation with subjective preferences. Additionally, we provide some sound examples at `https://ccrma.stanford.edu/~orchi/Mic-bleed/aes21.html` for an informal listening test. It is to be noted that the gains of these audio files have been normalized, so that the amplitudes lie within ±1.

5.1 Effect of source-microphone distance

For the determined case, ($N = M = 2$), we varied the distance between the source and the microphones linearly from $10-50$ cm in steps of 10 cm. The results are



(a) Overall perceptual score

(b) Target-related perceptual score

(c) Interference-related perceptual score

(d) Artifact-related perceptual score

Fig. 2: Perceptual scores for varying source-microphone distance

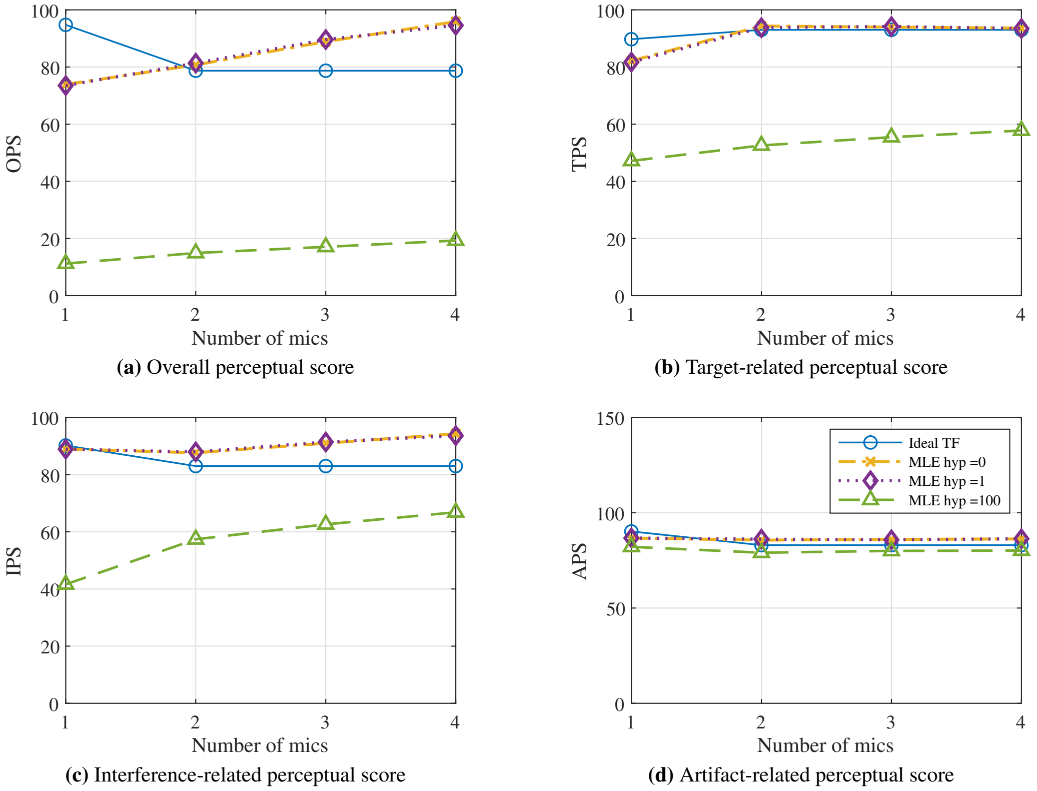
shown in Fig. 2. The OPS, TPS and IPS of all methods decline as the source-microphone distance is increased, since the close-microphone assumption starts failing. The proposed ML estimator with $\sigma_v^2/\sigma_w^2 \in \{0,1\}$ outperforms MCWF by a large margin. However, performance deteriorates when the hyperparameter value is large ($\sigma_v^2/\sigma_w^2 = 100$). This is because the initial transfer function matrix estimated with the spectral ratio is fairly accurate, and not trusting it leads to overfitting. In this case, choosing a large hyperparameter value hurts us. The advantages of using optimization is not clear in the figure; however, listening to the sound examples provided when the source-microphone distance is 0.3 m, makes it obvious. Optimization gets rid of subtle artifacts in the separated sources.

5.2 Effect of number of microphones

We varied the number of microphones capturing each instrument ($N \in \{1,2,3,4\}, M = 2$) to simulate an overdetermined system. The closest mic was kept on the same axis as the source at a distance of 20 cm, and the other microphones were distributed uniformly on an arc of angle 60° of the same radius. The results are shown in Fig. 3. The OPS, TPS and IPS improve as the number of microphones is increased. This is expected, since more information is available in overdetermined systems. The APS remains constant. Again, the performance of the ML estimator with $\sigma_v^2/\sigma_w^2 \in \{0,1\}$ are closely matched.

5.3 Effect of number of sources

For a determined system ($N = M$), we varied the number of sources, $M \in \{2,3,4\}$ to include all four instruments in the quartet. The four instruments – viola, violoncello and two violins were virtually placed at locations $(1.9,2.5,1.0)$, $(1.7,2.8,0.8)$, $(1.3,2.8,1.0)$, $(1.1,2.5,1.0)$ respectively. The virtual microphones were placed at a distance of 20 cm from the sources. The mean perceptual evaluation scores of the separated sources are shown in Fig. 4. The OPS, TPS and IPS of



(a) Overall perceptual score

(b) Target-related perceptual score

(c) Interference-related perceptual score

(d) Artifact-related perceptual score

Fig. 3: Perceptual scores for varying number of microphones

the ML estimator show a decreasing trend as the number of sources increases, which is expected. However, the APS score remains fairly constant, which is desired. MCWF outperforms the proposed method in terms of OPS when $M \in \{3,4\}$, but its other scores are sub-par.

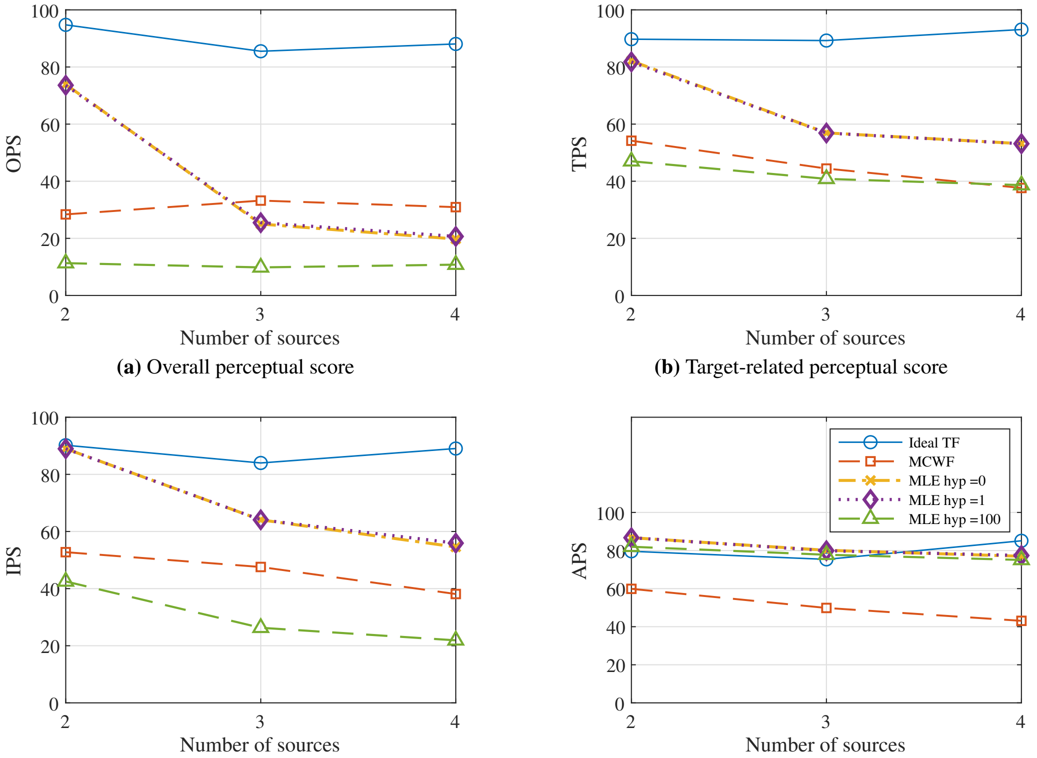
6 Discussion and Future Work

The proposed MLE method shows promising results on a simulated dataset. It is to be noted that, in some cases, the MLE scores overtake that of the system inverted with the known transfer function. While this may seem implausible, perfect reconstruction of the source signals given the ideal transfer function is unlikely because of noise. This is a common issue with Zero-Forcing equalizers [20].

We suspect the performance of the calibration stage using the spectral ratio method will degrade when tested on recordings made in a real studio. This is because the position of the source relative to the microphone will not remain constant as the performers make slight movements. Moreover, direction-dependent source radiation may have an effect. The actual scenario will be far more complex than what the Image-Source method can replicate. The benefits of optimization can be demonstrated more clearly in such cases. Hence, the proposed ML estimator needs to be tested on real-world recordings, and a listening test needs to be performed to see if human subject ratings corroborate the objective findings.

It will be interesting to compare different methods of calibration, including ones that approximate the unknown impulse response as FIR filters, and use blind channel identification methods for Single-Input-Multi-Output (SIMO) systems, when only one source is active at a time [21]. More sophisticated methods for Multi-Input Multi-Output (MIMO) blind system identification using second and higher order statistics are also available [22].

Since the roots of a non-linear equation need to be



(a) Overall perceptual score

(b) Target-related perceptual score

(c) Interference-related perceptual score

(d) Artifact-related perceptual score

Fig. 4: Perceptual scores for varying number of sources for a determined system

found for each frequency bin, the proposed method with optimization is inherently slow. The FFT size needs to be larger than the frame size to avoid time-domain aliasing due to convolution. More specifically, if a frame size of N samples is being filtered by a signal of length L samples, then the FFT size should be $N+L-1$ at least. A smaller FFT size means fewer frequency bins have to be processed at the cost of processing more time-frames, so the computational advantage of a short FFT is not necessarily significant. Further tests need to be performed to see how the performance changes with FFT size. The performance of both MCWF and MLE was seen to be relatively independent of the reverberation time (or absorption characteristics) of the room.

7 Conclusion

In this paper, we have proposed a novel method for microphone “bleed" or cross-talk cancellation in multichannel recordings. A maximum likelihood solution has been proposed that aims to simultaneously optimize the relative transfer function between each source and microphone, and separate the sources. A calibration stage is used for estimating an initial relative transfer function matrix. A hyperparameter determines how much the initial transfer function matrix is trusted during the optimization. Methods for parallelizing the optimization have been suggested.

The method was tested against a well-known multichannel Wiener filter solution on a dataset of string quartet recordings that were virtually placed in a studio using the image-source method. Perceptual scores were evaluated for varying source-microphone distance, as well as for varying number of microphones capturing each source and varying number of sources. The ML estimator generally outperforms MCWF. Unlike the MCWF, the ML estimator can also work with set ups where the number of microphones is equal to, or greater than the number of sources being captured, as long as the close microphone assumption is maintained. Performance of the ML method declines as the source-microphone distance increases, improves as number of microphones capturing an instrument increases, and deteriorates with increasing number of sources. Future work includes investigating alternate methods for the calibration stage and testing on data recorded in a real studio.

Acknowledgment

The authors would like to thank Elias K. Kokkinis for sharing his MATLAB code for the multichannel Wiener filter based cross-talk canceler, and Fabian Brinkman for sharing the string quartet recordings with us.

References

[1] Owsinski, B., *The mixing engineer's handbook*, Nelson Education, 2013.

[2] Kokkinis, E. K., Georganti, E., and Mourjopoulos, J., "Statistical properties of the close-microphone responses," in *Audio Engineering Society Convention 132*, Audio Engineering Society, 2012.

[3] Hyvärinen, A. and Oja, E., "Independent component analysis: algorithms and applications," *Neural networks*, 13(4-5), pp. 411–430, 2000.

[4] Ozerov, A. and Févotte, C., "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3), pp. 550–563, 2009.

[5] Adel, H., Souad, M., Alaqeeli, A., and Hamid, A., "Beamforming techniques for multichannel audio signal separation," *arXiv preprint arXiv:1212.6080*, 2012.

[6] Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T., and Shikano, K., "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Advances in Signal Processing*, 2003(11), p. 569270, 2003.

[7] Prätzlich, T., Bittner, R. M., Liutkus, A., and Müller, M., "Kernel additive modeling for interference reduction in multi-channel music recordings," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 584–588, IEEE, 2015.

[8] Clifford, A., Reiss, J. D., et al., "Microphone interference reduction in live sound," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, 2011.

[9] Uhle, C. and Reiss, J., "Determined source separation for microphone recordings using IIR filters," in *129th AES Convention*, 2010.

[10] Carabias-Orti, J. J., Cobos, M., Vera-Candeas, P., and Rodríguez-Serrano, F. J., "Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings," *EURASIP Journal on Advances in Signal Processing*, 2013(1), p. 184, 2013.

[11] Kokkinis, E. K. and Mourjopoulos, J., "Unmixing acoustic sources in real reverberant environments for close-microphone applications," *Journal of the Audio Engineering Society*, 58(11), pp. 907–922, 2010.

[12] Kokkinis, E. K., Reiss, J. D., and Mourjopoulos, J., "A Wiener filter approach to microphone leakage reduction in close-microphone applications," *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3), pp. 767–779, 2011.

[13] Meyer, P., Elshamy, S., and Fingscheidt, T., "Multichannel speaker interference reduction using frequency domain adaptive filtering," *EURASIP Journal on Audio, Speech, and Music Processing*, 2020(1), pp. 1–17, 2020.

[14] MathWorks, "MATLAB Optimization Toolbox," 2020, the MathWorks, Natick, MA, USA.

[15] Böhm, C., Ackermann, D., and Weinzierl, S., "A Multi-channel Anechoic Orchestra Recording of Beethoven's Symphony No. 8 op. 93," *Journal of the Audio Engineering Society*, 68(12), pp. 977–984, 2021.

[16] Habets, E. A., "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, 2(2.4), p. 1, 2006.

[17] Allen, J. B. and Berkley, D. A., "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, 65(4), pp. 943–950, 1979.

[18] Emiya, V., Vincent, E., Harlander, N., and Hohmann, V., "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), pp. 2046–2057, 2011.

[19] Vincent, E., Gribonval, R., and Févotte, C., "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language processing*, 14(4), pp. 1462–1469, 2006.

[20] Lucky, R. W., "The adaptive equalizer," *IEEE Signal Processing Magazine*, 23(3), pp. 104–107, 2006.

[21] Xu, G., Liu, H., Tong, L., and Kailath, T., "A least-squares approach to blind channel identification," *IEEE Transactions on Signal Processing*, 43(12), pp. 2982–2993, 1995.

[22] Chen, B. and Petropulu, A. P., "Frequency domain blind MIMO system identification based on second-and higher order statistics," *IEEE Transactions on Signal Processing*, 49(8), pp. 1677–1688, 2001.