BLIND ESTIMATION OF ACOUSTIC TRANSFER FUNCTIONS WITH APPLICATION TO DEREVERBERATION USING CONVOLUTIVE TRANSFER FUNCTIONS

Anchi Yuan  
*Department of Power Mechanical Engineering*  
*National Tsing Hua University*Hsinchu, Republic of China (ROC)  
a0926129380@gmail.com

You-Siang Chen  
*Department of Power Mechanical Engineering*  
*National Tsing Hua University*Hsinchu, Republic of China (ROC)  
s108033851@m108.nthu.edu.tw

Mingsian R. Bai  
*Department of Power Mechanical Engineering*  
*National Tsing Hua University*Hsinchu, Republic of China (ROC)  
msbai@pme.nthu.edu.tw

*Abstract*—Although Acoustic Transfer Functions (ATFs) yield superior results to Relative Transfer Functions (RTFs) in array signal processing, accurately estimating ATFs is challenging due to the absence of source input. In this paper, we propose a novel blind ATF estimation method based on convolutive transfer functions (CTFs). The method commences with the estimation of the source location through the calculation of time difference of arrival (TDOA) utilizing the technique of Generalized Cross Correlation-Phase Transform (GCC-PHAT) in conjunction with a distributed array. Subsequently, the Weighted Prediction Error (WPE) algorithm is employed to de-reverberate signals captured by a hybrid compact-distributed array, utilizing the Delay and Sum (DAS) beamformer as an initial source signal estimate. Subsequently, the CTF coefficients are calculated using either the Wiener filter or the Kalman filter. Simulations conducted with a hybrid array have demonstrated the efficacy of the proposed method. A state-of-the-art Adaptive Multichannel Time Domain Least Mean Square (MCLMS) method was employed as a benchmark for comparison. Furthermore, the estimated ATFs were employed in signal dereverberation, thereby providing additional validation of our approach.

Keywords—convolutive transfer functions, weighted prediction error, delayed and sum beamformer, Wiener filter, Kalman filter

# Introduction

Blind System Identification (BSI) is a method of identifying systems without access to the input signal, relying solely on the output signal. This is a challenging but essential process for applications that require the use of Acoustic Transfer Functions (ATFs), such as acoustic echo cancellation [1], dereverberation [2], blind source separation [3], and beamforming in reverberant environments [4]. Conventional BSI techniques frequently operate within the time domain [5] or the Short-Time Fourier Transform (STFT) domain [6] [7]. This entails estimating the time-domain convolution by multiplying the source STFT with the room impulse response (RIR) STFT. Nevertheless, the validity of this multiplicative transfer function (MTF) approximation [8] is contingent upon the RIR length being shorter than the STFT window. This is a condition that is seldom met in practice, largely due to the inherent limitations of the STFT window in assuming local stationarity of audio signals. Furthermore, the use of longer STFT windows has been shown to result in increased estimation variance and computational complexity.

To address this issue, crossband filters (CBFs) for linear system identification [9] were introduced as an alternative to the MTF approach. CBFs represent the STFT coefficient output as a sum of convolutions between the STFT coefficients of the input signal and the RIR across frequency bins. In order to facilitate analytical tractability, the convolutive transfer function (CTF) approximation [10] was proposed. This approach allows each frequency's output STFT coefficient to be represented as a unique convolution between the input signal's STFT coefficients and the CTF.

This paper presents a method for blind ATF estimation utilizing CTF approximation. Initially, source localization is performed using Generalized Cross Correlation-Phase Transform (GCC-PHAT) [11] to estimate the time difference of arrival (TDOA) [12] of each microphone in a distributed array, thereby aiding in source localization. Subsequently, the source signal is subjected to a pre-processing phase involving dereverberation and extraction, utilizing Weighted Prediction Error (WPE) [13] and Delay and Sum (DAS) beamforming [14] techniques. The CTF coefficients are calculated using the extracted source signal with either a Wiener [15] or an adaptive Kalman filter [16]. Subsequently, the estimated CTF coefficients are convolved with a constant-magnitude filter along the frequency axis, and the inverse STFT yields time-domain ATFs or RIRs.

The convergence performance is evaluated using the Normalized Root Mean Square Projection Mismatch (NRMSPM) between the ground truth RIR and the estimated RIR, in comparison to the Adaptive Multichannel Time Domain Least Mean Square (MCLMS) method [17]. The simulations encompass reverberation times ranging from 0.01 to 1.6 seconds, utilizing a hybrid array of 18 or 38 microphones. Furthermore, the application of signal dereverberation via the Multiple Input/Output Inverse Theorem (MINT) [18] is also included. The effectiveness of the proposed method is evaluated using objective metrics such as the Perceptual Evaluation of Speech Quality (PESQ) [19] and the Signal-to-Distortion Ratio (SDR) [20]. The simulations demonstrate that the proposed method exhibits a markedly superior performance compared to MCLMS.

# Ctf Signal Model

In an environment devoid of noise, the signal received by the microphone is presented in the time domain, as specified by

 (1)

where the *s*(*n*) and *a*(*n*) represent the source signal and the RIR, respectively. The symbol \* denotes the linear convolution. In (1), the RIR is typically estimated using the MTF in the STFT domain, as illustrated by

 (2)

where *yp,k* and s*p,k* represent the STFTs of their respective signals. Additionally, *ak* denotes the Fourier transformation of the RIR *a*(*n*). Furthermore, *p* ∈ [1, *P*] denotes the frame index, *N* indicates the STFT window size, and *k* ∈ [0, *N–*1] represents the frequency index. It should be noted, however, that this approximation is only valid if the length of the RIR *a*(*n*) is shorter than the STFT window size [9]. Accordingly, the cross-band filter model is employed in this study. The STFT coefficient *yp,k* is presented as the sum of multiple convolutions between the STFT-domain source signal and the filter over the frequency bins, as follows:

 (3)

The step size of the STFT frames is represented by *D*. In the event that *D* < *N*, *apˊ,k,k*ˊ will possess ⌈*N*/*D*⌉ *–* 1 non-causal coefficients [9]. The number of causal filter coefficients is dependent on the reverberation time. For the sake of simplicity in notation, we assume that the filter index *p*ˊ lies within the range [0, *L –*1], with *L* representing the filter length. This assumption requires that non-causal coefficients be relocated to the causal component, resulting in a fixed delay shift in the frame index of the received microphone signal [9]. The STFT analysis and synthesis windows are represented by[*w̃*](https://zh.wiktionary.org/zh-hant/Appendix:%E5%9B%BD%E9%99%85%E9%9F%B3%E6%A0%87%E7%AC%A6%E5%8F%B7#w%CC%83)(*n*) and *w*(*n*), respectively. The relationship between the STFT domain impulse response *apˊ,k,k*ˊ and the time domain impulse response *a*(*n*) is expressed as follows:

 (4)

which indicates the convolution with respect to the time index *n* evaluated at frame steps using

 (5)

In order to facilitate the analysis, we employ the CTF approximation, which exclusively considers the band-to-band filters with *k* = *kˊ*, as follows:

 (6)

Based on this, we propose a multi-channel configuration comprising *M* microphones as follows:

 (7)

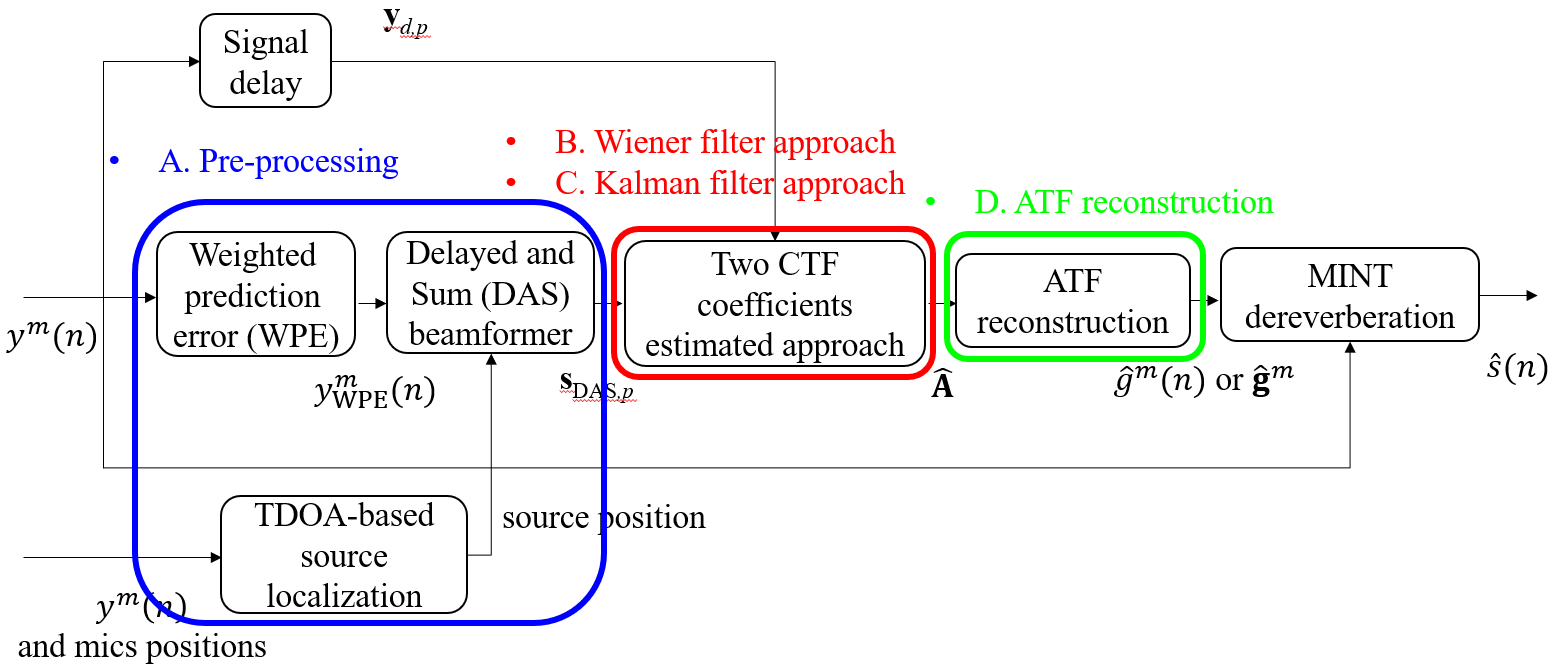
where *yi p,k* and *ai p,k* represent the *i*-th microphone signal and the corresponding CTF, respectively. Consequently, the source signals can be expressed in matrix form as follows:

 (8)

In order to streamline the discussion, the frequency index will be omitted from here on in, as the proposed algorithm operates on a frequency basis.

# Proposed Method

This section presents a method for estimating the ATF in a blind manner. It is important to note that the only data available for analysis is the microphone signal **y***d,p*, which is delayed by ⌈*N*/*D*⌉ *–* 1 frames [9], and the source signal **s***DAS,p* obtained via DAS beamformer. It is noteworthy that two techniques are offered for the estimation of the CTF coefficients. In order to facilitate comprehension of each step of the proposed method, we present a flow chart of the method in Fig.1, which we hope will assist readers in grasping the procedure.



1. Flow chart of the proposed method

## Pre-processing procedures

In order to localize the position of the source, the values of TDOA are calculated using the GCC-PHAT algorithm. Subsequently, the source location can be obtained as illustrated in [12].

In order for the subsequent CTF estimation algorithm to function effectively, it is essential that the source signal be free from contamination and echoes. Nevertheless, in practical applications, obtaining a source signal that is free from contamination and echoes is frequently a significant challenge. Accordingly, this paper utilizes the WPE algorithm for dereverberation, as detailed in [13]. Subsequently, the DAS beamformer is employed, utilizing the source location obtained from the aforementioned TDOA-based source localization method, to extract a clean source signal **s**DAS,p from the WPE outputs of all channels.

## Wiener Filtering CTF Coefficients Estimation Approach

In order to estimate the CTF coefficients matrix, the Wiener-based derivations are employed, which serve to minimize the expectation of the mean squared error. This is expressed by the following equation:

 (9)

where E[·] represents the expectation with respect to the frames. Consequently, equation (9) can be rewritten as follows:

 (10)

where *tr*{·} represents the matrix trace. The associated covariance matrices are provided below:

 (11)

By taking the derivative of (10) with respect to **A***H*, we obtain:

 (12)

The optimal Wiener solution can be obtained as

 (13)

In practical implementation, the recursive averaging method is employed to obtain **Rsy** and **Rss**, as demonstrated by the following equations:

 (14)

where α denotes the forgetting factor for the recursive averaging process.

## Kalman Apaptive Filtering CTF Coefficients Estimation Approach

In the second technique, the CTF coefficient matrix is estimated by applying the Kalman filter. It is noteworthy that this paper adopts the Kalman filter as an adaptive filter, rather than utilizing it as a state space control filter. Notwithstanding this modification, the fundamental concept remains unchanged. The process equation of the stationary Kalman adaptive filter for each microphone, in the absence of process noise, is as follows:

 (15)

where ∈ ℂLх1 signifies the optimal weight vector and has a connection with the CTF coefficients matrix as

 (16)

where denotes the *m*-*th* row of **A***p*. The measurement equation of the stationary Kalman adaptive filter for each microphone is as follows:

 (17)

where denotes the measurement noise for each microphone, and

 (18)

where is the covariance of measurement noise.

By employing the process and measurement equations delineated in (15) and (17), the Kalman gain can be derived through the minimization of the error covariance matrix [16]. Table 1 provides a concise overview of the stationary Kalman adaptive filter approach.

1. CTF Estimation Using Stationary Kalman Adaptive Filtering

|  |
| --- |
| Input: , **s**DAS,*p*  1) Initialize estimated Kalman weight, error covariance matrix, Kalman gain and measurement noise covariance as  where *η* and *ρ* is a small positive constant.  2) For each microphone, *m* = 1, 2, …,  For each instant of frame, *p* = 1, 2, …, compute |

## ATF Reconstruction

Once the matrix of CTF coefficients has been estimated from the aforementioned approaches, the subsequent step is to proceed with the production of the ATFs. The initial step is to generate a unit pulse sequence, which is subject to a delay of ( *L*-1)*D* points. Subsequently, the sequence is transformed into the short-time Fourier transform (STFT) domain, resulting in the following:

 (19)

It is evident that the magnitude across different frame indices *p* remains constant along the frequency axis, depending on the analysis window utilized. Ultimately, the estimated CTF coefficients are convolved with the aforementioned unit pulse sequence in the STFT domain, thereby yielding the following signal:

 (20)

where *p* ∈ [0, *PATF*]. The estimated RIRs , *n* ∈ [0, *NATF*], can be obtained by applying the inverse STFT to . Subsequently, the estimated ATFs, represented by the vector with each element corresponding to different frequency bins, can be obtained by performing a fast Fourier transform (FFT) on the estimated RIRs . is expressed as:

 (21)

# Simulations

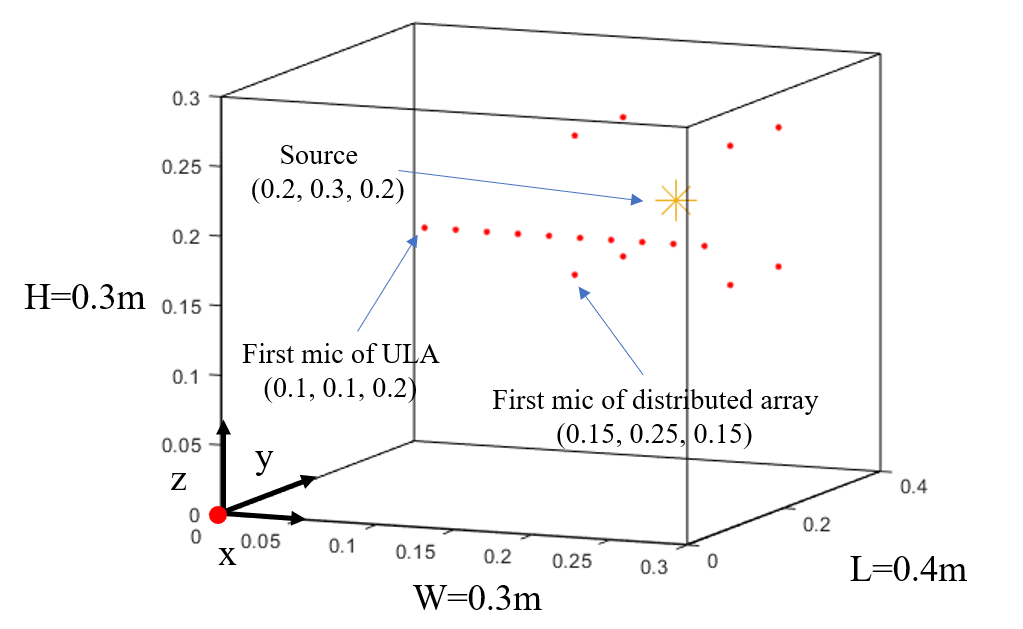
For purposes of comparison, our approach has been contrasted with the state-of-the-art BSI method, specifically the MCLMS method. Moreover, this chapter presents optimizations of filter parameters and applications of dereverberation utilizing MINT, which requires accurate estimated ATFs or RIRs.

## Settings and parameters

Three distinct room settings were devised to generate ground truth RIRs with varying reverberation times using the RIR generator [21], with the specifications of each room enumerated in Table 2. The system employs a hybrid compact-distributed array comprising eight microphones in a cuboidal distributed part, while the compact section is a uniform linear array (ULA) with microphones spaced at 0.02 m intervals. The number of microphones adopted in ULA is adjusted according to the reverberation time. Speech signals sampled at 16 kHz were employed as sources to generate microphone signals, which were subsequently convolved with the ground truth RIRs. The configuration of Room 1, as an example, is illustrated in Fig.2.

1. Specifications of room settings for Simulations

|  |  |  |  |
| --- | --- | --- | --- |
| Room  Settings | Room1 | Room2 | Room3 |
| Range of *T*60 (sec) | 0.01 | 0.1 | 0.2~1.6 |
| Dimensions of the room (m) | 0.3 × 0.4 ×0.3 | 3 × 3×2.5 | 5× 6×2.5 |
| Number of microphones of ULA | 10 | 30 | 30 |
| First sensor location of ULA (m) | 0.1, 0.1, 0.2 | 1.1, 1, 1 | 2.1, 2, 1 |
| Dimensions of distributed array (m) | 0.1×0.1×0.1 | 1 × 1 ×1 | 1 × 1 ×1 |
| First sensor location of distributed array (m) | 0.15, 0.25, 0.15 | 1, 1, 1 | 1, 1, 1 |
| Source location (m) | 0.2, 0.3, 0.2 | 1.7, 1.8, 1.3 | 2.1, 2.15, 1.1 |



1. Configurations of the Room 1

In the chapter of simulations, the values of the free parameters α, η and ρ were consistently set to 0.999, 0.5 and 0.001 respectively, as they were found to be appropriate for all conditions.

## Results and discussions

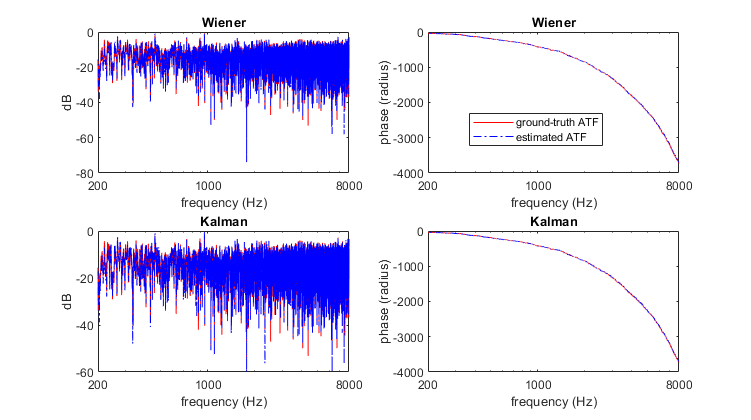
The magnitude and phase of the estimated ATF for all frequency bins, along with the amplitude of the estimated RIR for reverberation times equaling 0.01 and 1.6 seconds, are compared with their ground truth values and presented in Fig.3. However, when the reverberation time exceeds 0.1 seconds, the MCLMS method encounters difficulties in converging due to the challenge of identifying the minimum eigenvalue of its microphone signal covariance matrix, as illustrated in Fig.4. As a result, MCLMS simulations are only conducted for reverberation times below 0.1 seconds. Fig.3. illustrates the efficacy of the proposed method, with extremely small absolute magnitude and phase errors across all frequency bins and a remarkable correspondence between the amplitude of the estimated RIR and its ground-truth counterparts.



(a)



(b)



(c)



(d)

1. Magnitude and phase of the estimated ATF and amplitude of the estimated RIR of all algorithms with several chosen *T*60 (a) ATF with *T*60= 0.01s (b) RIR with *T*60= 0.01s (c) ATF with *T*60= 1.6s (d) RIR with *T*60= 1.6s



1. The eigenvalues of the microphone covariance matrix in MCLMS

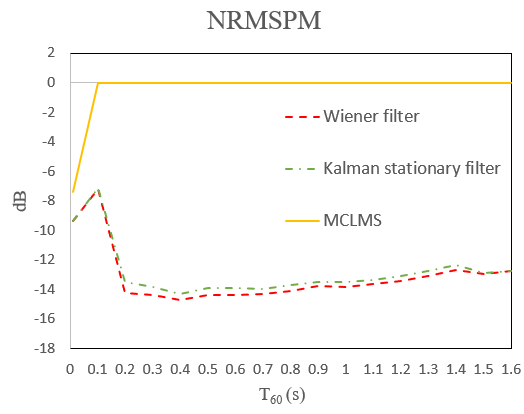
Fig.5 demonstrate the normalized root mean square projection misalignment (NRMSPM) of the estimated ATFs for all algorithms. Again, MCLMS simulations are only performed for reverberation times below 0.1 seconds. Consequently, the NRMSPM values for higher reverberation times are set to zero for MCLMS. NRMSPM is defined as follows:

 (22)

where *N* represents the number of Monte Carlo runs, **g** denotes a long vector connected by the ground-truth RIR of each channel, and (•)(*i*) denotes a value obtained from the *i*-*th* run. The projection misalignment vector is represented as follows:

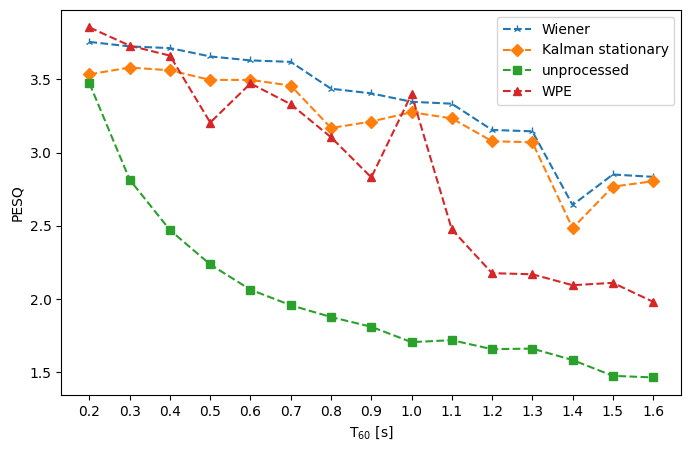
 (22)

where denotes a long vector linked by the estimated RIR of each channel.

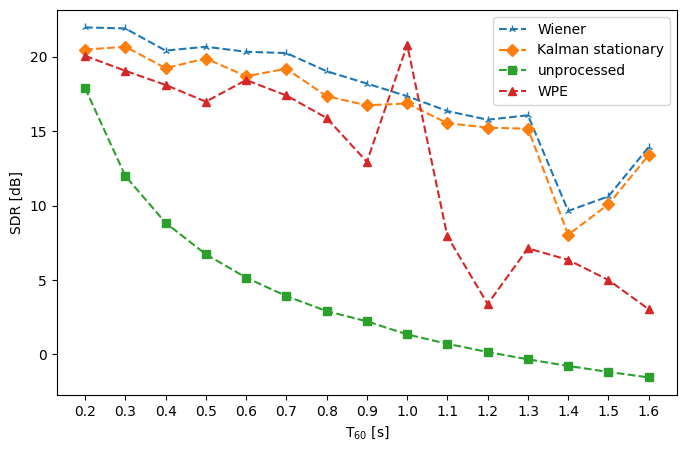


1. NRMSPM of the estimated RIRs for all algorithms at different *T*60

Fig.6 presents the Perceptual Evaluation of Speech Quality (PESQ) and Signal-to-Distortion Ratio (SDR) values calculated between the source signal obtained after MINT dereverberation and the ground truth source signal. The RIRs were estimated from the *T*60 value, which ranged from 0.2 to 1.6 seconds, with an interval of 0.1 seconds. In addition, the WPE and unprocessed microphone signals are evaluated against the ground truth source signal for comparison in terms of PESQ and SDR. The results demonstrate that the signals processed by the MINT dereverberation method achieve higher scores than the WPE and unprocessed signals, thereby substantiating the superior performance of the proposed method.



(a)



(b)

1. (a) PESQ and (b) SDR of MINT de-reverberated signal using RIR estimated by proposed method, WPE signal and unprocessed signal

# Conclusions

This thesis presents a blind estimation method for ATFs based on the CTF model. Two techniques are developed for estimating the CTF coefficient matrices using the Wiener filter and the Kalman filter, respectively. The magnitude and phase of the estimated ATF are compared with those of the ground truth ATF, and the NRMSPM of the estimated RIRs is also calculated using its ground-truth counterpart. The findings demonstrate that the proposed method yields more precise ATF estimation than the baseline MCLMS approach in simulations. This enhanced accuracy can be attributed to the challenge of identifying the minimum eigenvalue of the microphone signal covariance matrix in MCLMS. Moreover, the dereverberation outcomes generated by MINT are evaluated in comparison with the ground truth source signal through the use of PESQ and SDR. The results illustrate a significant improvement in the processed signal in comparison to the other signals included in the analysis.

##### Acknowledgment

The authors would like to acknowledge National Tsing Hua University, ROC.

##### References

1. J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, “Advances in Network and Acoustic Echo Cancellation,” *New York: Springer,* 2001.
2. Y. Huang, J. Benesty, and J. Chen, “A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment,” *IEEE Trans. Speech Audio Process,* vol. 13, no. 5, pp. 882–895, 2005.
3. S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, “The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech,” *IEEE Trans. Audio, Speech, Lang. Process,* vol. 14, no. 3, pp. 774–784, May 2006.
4. S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Trans. Audio, Speech, Lang. Process,* vol. 49, no. 8, pp. 1614–1626, Aug. 2001.
5. Y. Huang and J. Benesty, “Adaptive multi-channel least mean square and Newton algorithms for blind channel identification,” *Signal Processing,* vol. 82, no. 8, pp. 1127–1138, Aug. 2002.
6. Y. Huang and J. Benesty, “A class of frequency-domain adaptive approaches to blind multichannel identification,” *IEEE Trans. Signal Processing,* vol. 51, no. 1, pp. 11–24, Jan. 2003.
7. M. K. Hasan, J. Benesty, P. A. Naylor, and D. B. Ward, “Improving robustness of blind adaptive multichannel identification algorithms using constraints,” *Proc. European Signal Processing Conference (EUSIPCO),* 2005.
8. Y. Avargel and I. Cohen, “On multiplicative transfer function approximation in the short-time Fourier transform domain,” *IEEE Signal Processing Letters,* vol. 14, no. 5, pp. 337–340, 2007.
9. Y. Avargel and I. Cohen, “System identification in the short-time Fourier transform domain with crossband filtering,” *IEEE Trans. Audio, Speech, Lang. Process,* vol. 15, no. 4, pp. 1305–1319, 2007.
10. R. Talmon, I. Cohen and S. Gannot, “Relative transfer function identification using convolutive transfer function approximation,” *IEEE Trans. Audio, Speech, Lang. Process,* vol. 17, no. 4, pp. 546–555, 2009.
11. B. Van Den Broeck, A. Bertrand, P. Karsmakers, B. Vanrumste, H. Van hamme and M. Moonen, “Time-domain generalized cross correlation phase transform sound source localization for small microphone arrays,” *2012 5th European DSP Education and Research Conference (EDERC),* pp.76-80, 2012.
12. T. Li, Z. Deng, G. Wang and J. Yan, "Time Difference of Arrival Location Method Based on Improved Snake Optimization Algorithm, “*2022 IEEE 8th International Conference on Computer and Communications (ICCC),* pp. 482-486, 2022.
13. T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B. -H. Juang, “Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction,” *IEEE Trans. Audio, Speech, Lang. Process,* vol. 18, no. 7, pp. 1717–1731, 2010.
14. Harry L. Van Trees, “Optimum array processing: Part Iv of detection, estimation, and modulation theory,” *New York: Wiley,* 2002*.*
15. J. Benesty, S. Makino, J. Chen, Y. Huang, and S. Doclo, “Study of the Wiener filter for noise reduction,” *Speech enhancement,* pp. 9-41, 2005.
16. R. E. Kalman, “A New Approach to Linear Filtering and Prediction Problems,” *Journal of Basic Engineering,* 1960.
17. Y. Huang and J. Benesty, “Adaptive blind channel identification: Multi-channel least mean square and Newton algorithms,” *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. II-1637-II-1640, 2002.
18. M. Miyoshi, and Y. Kaneda, “Inverse filtering of room acoustics,” *IEEE Trans. Audio, Speech, Lang. Process,* vol. 36, no. 2, pp. 145-152, Feb. 1988.
19. A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,”*IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings,* vol. 2,pp. 749-752, 2001.
20. 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.
21. Emanuël Habets, “Room Impulse Response Generator,” *Internal Report,* pp. 1-17, 2006*.*