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)

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, with parameters optimized via particle swarm optimization (PSO). Simulations and experiments conducted with a thirteen-microphone 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, particle swarm optimization

# 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]. Furthermore, the parameters of the aforementioned filters are optimized using Particle Swarm Optimization (PSO) [17]. 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 [18]. The simulations encompass reverberation times ranging from 0.01 to 1.6 seconds, utilizing a hybrid array of 38 microphones. Furthermore, the application of signal dereverberation via the Multiple Input/Output Inverse Theorem (MINT) [19] is also included. The effectiveness of the proposed method is evaluated using objective metrics such as the Perceptual Evaluation of Speech Quality (PESQ) [20] and the Signal-to-Distortion Ratio (SDR) [21]. Experiments conducted in a room with a reverberation time of 0.128 seconds, utilizing 13 microphones, 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

 

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:

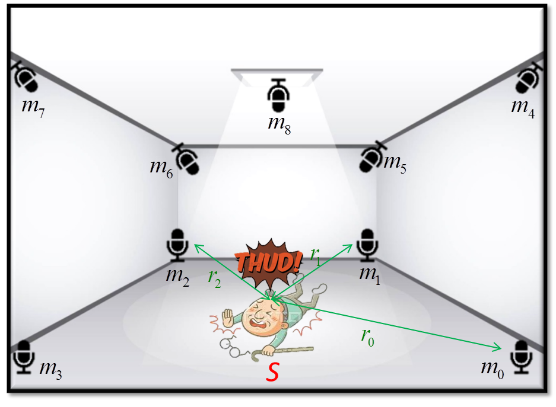
 8

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*, which is delayed by ⌈*N*/*D*⌉ *–* 1 frames [9], and the source signal **s***DAS* obtained via DAS beamformer. It is noteworthy that two techniques are offered for the estimation of the CTF coefficients.

## Pre-processing procedures



1. Relative positions of microphones and the sound source in TDOA-based source localization algorithm

Fig.1 depicts the relative positions of the microphones and the sound source in a TDOA-based source localization algorithm. For the purposes of this discussion, let *S* represent the source, *mm* (*m*=1,…,*M*) denote the microphones, and *m*0 be the designated reference microphone. The values of TDOA can be 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 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. The essence of the Wiener filtering approach can be encapsulated as follows:

|  |
| --- |
| **Algorithm 1** CTF estimation using Wiener filtering |
| Input: **y**d,p, **s**DAS,p1) Initialize forgetting factor α and covariance matrices as2) For each instant of frame, p = 1, 2, …, compute |

## 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.

## ATF Reconstruction

## Summary of Proposed Method

# Simulations

# Experiments

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.