In a setting free from noise and reverberations, the microphone's received signal is expressed in the time domain, as given by (1), where the source signal and the RIR are represented by (2) and (3) respectively, and (4) denotes linear convolution. (5) is commonly estimated using the MTF in the STFT domain, as shown by (6), where (7) and (8) are the STFTs of the corresponding signals, and (9) is the Fourier transform of the RIR, (10) is the frame index, (11) is the STFT window size, and (12) is the frequency bin index. It is worth noting that this approximation is only applicable when the RIR is shorter than the STFT window, which can be a matter of debate. In this paper we therefore empoly the cross-band filter model. The model represents the STFT coefficient as the sum of multiple convolutions between the STFT domain source signal and filter across frequency bins as follow: (13). Let (14) denote the STFT frame step. If (15) is small than (16), then (17) is non-causal, with eight non-causal coefficients [9] . The amount of causal filter coefficients is linked to the reverberation time. For the sake of notation simplicity, we assume that the filter index is in (18), with (19) being the filter length. This corresponds to shifting non-causal coefficients to the causal part, resulting in a constant delay shift of the frame index for the microphone received signal. The STFT analysis and synthesis windows are represented as (20) and (21), respectively. The STFT domain impulse response is related to the time domain impulse response by (22), which represents the convolution with respect to the time index  evaluated at frame steps with (23). To simplify the analysis, we employ the CTF approximation, which only considers band-to-band filters with (24) as follow: (25). Based on above, we consider a three microphones multi-channel version, where (26) and (27) denote microphone signal and CTF related to it. For frequency (28) and frame (29) , let(30), (31) and (32) denote the matrices of sensor signals, CTF coefficients and source signals respectively. As the proposed algorithm functions on a frequency basis, the frequency index is excluded for the remainder of this text. Therefore, (7) can be rephrased as (8).

In a noise-free and echoless environment, the signal received by the microphone is presented in the time domain, as specified by equation (1). The source signal and room impulse response (RIR) are denoted by equations (2) and (3), respectively, while (4) indicates linear convolution. Please note that technical abbreviations will be explained upon first use. (5) The estimation of << is often determined>> using the MTF in the STFT domain, as demonstrated by (6). Here, (7) and (8) represent the STFTs of their respective signals, while (9) denotes the Fourier transformation of the RIR. Additionally, (10) refers to the frame index, (11) indicates the STFT window size and (12) represents the frequency bin index. However, it should be emphasised that this approximation is only valid if the length of the RIR is shorter than the STFT window, which remains subject to debate. Therefore, we utilize the cross-band filter model in this study. The STFT coefficient model is presented as the sum of multiple convolutions between the STFT domain source signal and filter over frequency bins in the following manner: (13). STFT frame step is denoted by (14). If (15) is less than (16), then (17) exhibits eight non-causal coefficients [9]. The number of causal filter coefficients is connected to the reverberation time. For simplicity of notation, we assume that the filter index is in (18), with (19) as the filter length. This involves relocating non-causal coefficients to the causal component, leading to a fixed delay shift of the frame index for the received microphone signal. The STFT analysis and synthesis windows are denoted by (20) and (21), respectively. The STFT domain impulse response is linked to the time domain impulse response by (22), which indicates the convolution in relation to the time index assessed at frame steps using (23). To streamline the analysis, we utilize the CTF approximation, which solely takes into account band-to-band filters with (24) as outlined in (25). Following this, we examine a multi-channel version with three microphones, where (26) and (27) signify the microphone signal and the corresponding CTF, respectively. For frequency (28) and frame (29), matrices of sensor signals, CTF coefficients, and source signals are denoted as (30), (31), and (32), respectively. Technical abbreviations are explained upon first use. As the proposed algorithm operates on a frequency basis, the frequency index is omitted from the rest of this text. Therefore, (7) can be restated as (8).

In this section, we derive the method to estimate ATF blindly. Note that we only possess microphone signal and source signal extracted from DAS beamformer. Firstly, we estimated the CTF coefficients matrix by Wiener filter which minimizing the mean square error of (8) . It can be represented as follow: (1), where (2) denotes the expectation with respect to frame axis. The solution to (9) is given by (10).

Secondly, we estimated the CTF coefficients matrix by adaptive filter. Here we take Recursive Least Square (RLS) algorithm for example, which is already done in [14] in real domain with the notable difference that here the optimization process is carried on the complex domain. RLS algorithm minimizes sum of weighted error square described as follow:(11), where (13) denotes both adapting iteration and frame index, and (14) denotes forgetting factor which wighted the error with respect to iteration. Based on the objective function in (11), RLS algorithm is summarized in Algorithm 1. Once we get estimated CTF coefficients from each of two filters described above, we can progress to produce ATFs using CTF coefficients we haved obtained. Firstly, we generate an unit pulse sequence delayed by two points. Afterthat, we transform it to STFT domain obtaining (15) where (16) denotes frame index. Finally, we convolve it with estimated CTF coefficients to obtain ATF as follow: (20), where (22) denotes ordinal number of microphone. One can simply obtain RIR by means of inverse STFT with respect to ATF.

In this section, we present a method to estimate the ATF in a blind manner. It is important to note that we only have access to the microphone signal and source signal obtained via DAS beamformer. Initially, we estimate the CTF coefficients matrix via Wiener filter, which minimises the mean square error of (8). This can be expressed as (1), where (2) denotes the expectation with respect to the frame axis. The solution to (9) is provided by (10).

Next, we estimate the CTF coefficients matrix using adaptive filter. Here, we will use the Recursive Least Squares (RLS) algorithm as an example, which has already been performed in the real domain in [14], with the notable difference that the optimization process is now carried out in the complex domain. The RLS algorithm reduces the sum of the weighted error square, described as follows: (11), where (13) represents both the adaptation iteration and the frame index, and (14) represents the forgetting factor that weights the error with respect to the iteration. Based on the objective function outlined in (11), the RLS algorithm can be summarised in Algorithm 1. Once we have estimated the CTF coefficients from each of the two filters described above, we can move on to producing the ATFs using these coefficients. Firstly, we generate a unit pulse sequence which is delayed by two points. We then transform it to the STFT domain, obtaining Equation (15), where (16) denotes the frame index. Finally, we convolve it with the estimated CTF coefficients to obtain the ATF as follows: (20), where (22) indicates the ordinal number of the microphone. The room impulse response (RIR) can be obtained through the inverse short-time Fourier transform (STFT) with respect to the acoustic transfer function (ATF).

To test the accuracy of the proposed ATF estimation method, we dereverb the microphone signal using estimated RIR by means of MINT [15] in the time domain. For comparison, we also dereverb the microphone signal using state-of-the-art dereverberation method, namely WPE [11] . Ground-truth RIRs were generated with the Room Impulse Response Generator [19] . The size of the room was 5 m × 6 m × 2.5 m. The first sensor of thirty-microphone ULA was located at (1 m, 1.5 m, 1 m) with spacing 0.02 m along first axis. Speech signals sampled at 16 kHZ were used as sources convolved with the ground-truth RIRs to generate microphones signals. The speech sources were located at (2 m, 2.6 m, 1 m). Three reverberation times were tested, namely  = 0.2 s, 0.3 s and 0.4 s. The STFT window was a Hamming window of 1024 samples (64 ms), with 75% overlap. The free parameter  was set at a fixed value of 0.999, 0.99 and 0.01 through the whole set of simulations, as it was shown to be suitable for all the conditions.

The Perceptual Evaluation of Speech Quality (PESQ) [16], Signal-to-Distortion Ratio (SDR) [17] and Scale-Invariant Signal-to-Noise Ratio (SI-SNR) [18] are used as dereverberation performance metrics on the four types of signal, namely RLS, Wiener filter, WPE and unprogressed. Fig. 1 plots the PESQ obtained for the 3 reverberation times. It can be seen in plot that the proposed two methods achieve higher PESQ than WPE, and that as the reverberation time increases, the PESQs of all the four signals decrease.

Fig. 2 plots the SDR obtained for the 3 reverberation times. We can see in this plot that the proposed two methods still achieve higher SDR than WPE, and that as the reverberation increases, the SDRs of all the four signals decreases.

Finally, Fig. 3 plots the SI-SNR obtained for the 3 reverberation times. We can see in this plot that the proposed two methods achieve lower SI-SNR than WPE except for Wiener filter at the last reverberation time

To assess the precision of the proposed method for estimating ATF, we use MINT [15] to dereverberate the microphone signal in the time domain employing the estimated RIR. For comparative purposes, we also dereverberate the microphone signal utilizing the state-of-the-art dereverberation technique WPE [11]. The Room Impulse Response Generator [19] was employed to create ground-truth RIRs. The room dimensions were 5 m × 6 m × 2.5 m. The first sensor of a 30-microphone ULA was positioned at (1 m, 1.5 m, 1 m) with a 0.02 m spacing along the first axis. To generate microphone signals, speech signals sampled at 16 kHz were used as sources convolved with the ground-truth RIRs. The speech sources were located at (2 m, 2.6 m, 1 m). We tested three different reverberation times, namely 0.2 s, 0.3 s, and 0.4 s. To perform the STFT, we used a 1024-sample Hamming window (64 ms) with 75% overlap. The free parameters  were fixed at a consistent value of 0.999, 0.99, and 0.01 throughout the simulations, as it proved to be suitable for all conditions.

Four types of signal in abbreviation, namely proposed method with RLS algorithm, proposed method with Wiener filter, WPE and unprogressed, were evaluated for their dereverberation performance metrics using the Perceptual Evaluation of Speech Quality (PESQ) [16], Signal-to-Distortion Ratio (SDR) [17], and Scale-Invariant Signal-to-Noise Ratio (SI-SNR) [18]. Fig. 1 illustrates the obtained PESQ for the three reverberation times. It can be observed from the plot that the two proposed methods achieve higher PESQ compared to WPE. Furthermore, as the reverberation time increases, the PESQs of all four signals decrease.

Fig. 2 illustrates the SDR obtained for three reverberation times. It is evident in this graph that the proposed two methods still achieve higher SDR than WPE, and that the SDRs of all four signals decrease with increasing reverberation time.

Finally, Fig. 3 displays the SI-SNR obtained for three reverberation times. We can observe from this graph that, excluding the Wiener filter at the final reverberation time, the proposed two techniques produce lower SI-SNR than WPE, which is our desired outcome.

In this paper, we propose a blind estimation method for ATF based on the CTF model. The proposed method can accurately estimate ATF and outperforms the WPE method in dereverberation applications. Currently, based on the microphone array signal processing techniques, we are developing the application of the convolutional beamforming using a weighted power minimization distortionless response beamformer (WPD), which will replace WPE and DAS in our proposed method and reduce the computation time.