ABSTRACT
An auto-relation aided multiple-input/output inverse filtering algorithm (A-RAM) is proposed for the inverse filtering of room acoustics in speech dereverberation. In A-RAM, we exploit the relationship between the received reverberant signals and the acoustic impulse responses as well as constraining the reconstructed signals of the two sub-systems. This results in an auto-relation which is then used as a constraint for the adaptive multiple input/output inversion theorem (A-MINT) algorithm. Simulation results, using both synthetic and recorded room impulse responses, show that the proposed A-RAM achieves a fast convergence compared to A-MINT.

1. INTRODUCTION
Reverberation occurs when an acoustic signal is radiating from a source to one or more microphones in an enclosed environment. As a result of multi-path propagation, the received acoustic signal(s) sound echoy. It is well known that reverberation reduces the intelligibility of the speech signal. In addition, the performance of automatic speech recognition algorithms for both terminal and mobile hands-free devices can be significantly degraded due to reverberation of the enclosure [1]. One possible way to address this problem is to deploy speech dereverberation algorithms through the use of room impulse response (RIR) estimation and inverse filtering. This class of dereverberation algorithms is popular because of its computational efficiency and its potential ability to deal with long reverberation time [2].

Estimation of RIRs can be achieved by blind system identification (BSI) algorithms. To achieve speech dereverberation, an inverse filtering process is subsequently required. Approaches for single-input/single-output (SISO) as well as single-input/multiple-output (SIMO) equalization techniques have been proposed. Algorithms developed for single-channel equalization include single-channel least squares (SCLS) and homomorphic equalization [3]. In the SCLS method, an adaptive algorithm which minimizes the squared error between the outputs of the inverse filtering system and the desired system is adopted. In homomorphic inverse filtering, the RIR is first decomposed into minimum phase and all-pass components. An inverse can then be found for the minimum phase component while the all-pass component is equalized using a matched filter [3]. However, it was found that using a matched filter for the equalization of the non-minimum phase components will result in audible residual echoes [4]. In addition, it has been studied in [1] that the least square error (LSE) inverse filters often require many coefficients which, in turn, introduces significant delay. It was subsequently concluded in [5], that although SCLS achieves less accurate inversion, it is more efficient in practice. Amongst one of the most popular algorithms proposed for dereverberation is the use of multiple-input/output inverse theorem (MINT) [6]. In the context of room acoustics, achieving an exact inverse of RIRs is challenging since an RIR is often non-minimum phase [7]. The MINT algorithm addresses this problem and achieves acoustic system inversion by exploiting spatial diversity using multiple microphones.

We present a new algorithm for the adaptive inverse filtering of room acoustics with application to speech dereverberation. The proposed method determines the inverse filters by taking into account how the dereverberated speech is reconstructed at each channel and segregating a SIMO system into two sub-systems among which each sub-system can estimate a clean speech. The difference between the two dereverberated speech signals is then used to update the learning algorithm.

2. REVIEW OF EXISTING ALGORITHMS
We consider a SIMO finite impulse response (FIR) system as shown in Fig. 1. The observed signal \( x_i(n) \) of each channel is a linear convolution between a source signal \( s(n) \) and the RIR \( h_i(n) = [h_i,0(n) \cdots h_i,L_i(n)]^T \) of an acoustic channel from the source to the \( i \)th microphone such that

\[
x_i(n) = h_i^T(n) s(n) + v_i(n), \quad i = 1, 2, \ldots, M,
\]

where \( s(n) = [s(n-1) \cdots s(n-L+1)]^T \). \( L \) is the length of the RIR, \( v_i(n) \) is the additive noise, \( M \) is the total number of channels and \([ \cdot ]^T\) denotes the transpose operation. For clarity of presentation, we assume \( v_i(n) = 0 \forall i \).

2.1. Review of the MINT algorithm
The main aim of inverse filtering is to estimate an inverse filter \( g(n) = [g_1^T(n) \ g_2^T(n) \cdots g_M^T(n)]^T \) that is formed by concatenating inverse filters \( g_i(n) \) each of length \( L_d \). In inverse filtering, the target impulse response \( d \) is defined as a Kronecker delta function

\[
d = [0 \cdots 0 1 0 \cdots 0],
\]

where \( L_d \) is the arbitrary delay. In speech dereverberation, it is desirable to have \( d(n) = g(n) * h(n) \) such that the dereverberated speech can be estimated as

\[
\widehat{s}(n) = g(n) * x_i(n) = g(n) * h(n) * s(n)
\]

\[
d(n) * s(n) = s(n).
\]

In order to address the non-minimum phase problem of RIRs, the
MINT algorithm utilizes a multichannel model such that
\[
d(n) = \sum_{i=1}^{M} g_i(n) * h_i(n).
\] (3)

The closed-form solution of (3) can be obtained by utilizing vector notations such that
\[
\hat{g}(n) = \left[ H^T(n) H(n) \right]^{-1} H^T(n) d(n),
\] (4)

where \( H(n) = [H_1(n) \ H_2(n) \ \cdots \ H_M(n)] \) while \( H_i(n) \) is the convolution matrix of \( h_i(n) \). It should be noted that although MINT does not require each RIR to be minimum phase, it assumes that no common or near-common zeros are shared among the channels.

2.2. Review of the A-MINT Algorithm

In order to reduce the computational complexity, one of the most recent equalization algorithms is the adaptive MINT (A-MINT) algorithm [8]. This algorithm is based on the least-mean-square (LMS) algorithm and is developed based on a cost function defined by
\[
J_\alpha(n) = \| d(n) - H(n) \hat{g}(n) \|^2,
\] (5)

where \( H(n) \) is defined after (4), \( \hat{g}(n) = [\hat{g}_1(n) \ \cdots \ \hat{g}_M(n)]^T \) and \( \hat{g}_i(n) \) is the estimate of \( g_i(n) \). The gradient of this learning algorithm is given by
\[
\nabla J_\alpha(n) = \frac{\partial J_\alpha(n)}{\partial \hat{g}(n)} = -2H^T(n)d(n) + 2H^T(n)H(n)\hat{g}(n),
\]

which is subsequently employed in the update equation
\[
\hat{g}(n+1) = \hat{g}(n) + \mu \nabla J_\alpha(n) |_{\hat{g} = \hat{g}(n)},
\]

where \( \mu \) is the step-size.

We note that the advantage of A-MINT is not limited to a reduction in computational complexity, but also its ability to adapt to changes of the RIRs. However, as shown in Section 4, one of the main weaknesses of this algorithm is its slow convergence.

3. THE PROPOSED MULTICHANNEL SYSTEM INVERSION ALGORITHM

We propose an auto-relation aided MINT (A-RAM) algorithm by segregating, as illustrated in Fig. 2, a \( M \)-channel system into two sub-systems where the first sub-system contains channel indices 1 to \( M_1 \) while the second sub-system contains channel indices \( M_1 + 1 \) to \( M \) such that \( 1 < M_1 < M \). Although partitioning a SIMO model is not limited to two sub-systems, we limit our discussion as such for clarity. The motivation of segregating into two sub-systems is to allow multiple reconstruction of the speech signal. We therefore propose to improve the performance of the A-MINT algorithm by imposing an additional constraint \( J_\alpha(n) \) such that we minimize the cost function
\[
J_\alpha(n) = \| d(n) - \hat{H}_1(n)\hat{g}_1(n) \|^2 + \| d(n) - \hat{H}_2(n)\hat{g}_2(n) \|^2 + \beta J_\alpha(n),
\] (6)

where \( \hat{H}_1(n) = [\hat{H}_1(n) \ \cdots \ \hat{H}_{M_1}(n)] \), \( \hat{H}_2(n) = [\hat{H}_{M_1+1}(n) \ \cdots \ \hat{H}_M(n)] \), \( \hat{g}_1(n) = [\hat{g}_1(n) \ \cdots \ \hat{g}_{M_1}(n)]^T \) and \( \hat{g}_2(n) = [\hat{g}_{M_1+1}(n) \ \cdots \ \hat{g}_M(n)]^T \), such that \( d(n) \) is defined in (1) and \( \beta \) is the scaling factor.

To illustrate the concept of the additional constraint \( J_\alpha(n) \), we consider initially \( M = 2 \) channels where each sub-system contains one channel such that \( M_1 = 1 \). We achieve inverse filtering by taking into account how the reverberant speech from each microphone is generated from \( s(n) \) where \( x_i(n) = s(n) * h_i(n) \). The estimated speech for each channel, and hence each sub-system, is given by convolving \( x_i(n) \) with its inverse filter, i.e.,
\[
\hat{s}_i(n) = x_i(n) * \hat{g}_i(n) = s(n) * h_i(n) * \hat{g}_i(n),
\] (7)

\[
\hat{s}_j(n) = x_j(n) * \hat{g}_j(n) = s(n) * h_j(n) * \hat{g}_j(n).
\] (8)

Since the aim of speech dereverberation requires both dereverberated signals \( \hat{s}_i(n) \) and \( \hat{s}_j(n) \) to be equal to \( s(n) \), we obtain the important relation \( x_i(n) * \hat{g}_i(n) = x_j(n) * \hat{g}_j(n) \) which can be expressed in vector notation by
\[
x_i^T(n) \hat{g}_i(n) = x_j^T(n) \hat{g}_j(n),
\] (9)

where \( x_i(n) = [x_i(n) \ x_i(n-1) \ \cdots \ x_i(n-L+1)]^T \) and \( x_j(n) = [x_j(n) \ x_j(n-1) \ \cdots \ x_j(n-L+1)]^T \). We note that, as opposed to the cross-relation defined for BSI [2], the auto-relation given by (9) governs the reconstruction of \( s(n) \) in the context of channel equalization.

For the above illustrative example where each of the two sub-systems contains only one channel, a stable inverse filter \( \hat{g}_i(n) \) in (7) and (8) does not exist due to the non-minimum phase property of the RIR. For this reason, we extend (9) into multichannels such that each of the two sub-systems contains more than one channel operating within a multichannel MINT framework, i.e.,
\[
\sum_{i=1}^{M_1} x_i^T(n) \hat{g}_i(n) = \sum_{i=M_1+1}^{M} x_i^T(n) \hat{g}_i(n),
\] (10)

where \( M_1 \) is the last channel index of the first sub-system.

Since we desire the output of each sub-system (containing the reconstructed speech signal) to be equivalent, the differential signal...
between these two sub-systems is then given by
\[ e(n) = \sum_{i=1}^{M_1} x_i^T(n) \hat{g}_i(n) - \sum_{i=M_1+1}^{M} x_i^T(n) \hat{g}_i(n) = \chi_1^T \cdot \text{A-RAM algorithm when } \beta = 0. \] This is due to the increased number of near-common zeros in the three-channel relations between the two sub-systems is defined as
\[ J_{ar}(n) = e^2(n). \] A LMS algorithm is proposed to solve this optimization problem using an update equation given by
\[ \hat{g}(n+1) = \hat{g}(n) - \mu \nabla J_{ar}(n)|_{\hat{g} = \hat{g}(n)}, \] where \( J_{ar}(n) \) is defined in (6).

It is observed that (6) is an auto-rotation aided MINT cost function, in which the first two terms are employed for A-MINT [8] while the last term \( J_{ar}(n) \) is included to better estimate the inverse filters \( \hat{g}_i(n) \) by taking into account how reverberant speech \( x_i(n) \) is generated via the unknown system through (2), (7) and (8). We note that the trivial solution of \( \hat{g}_i(n) = \hat{g}_i(n) = 0 \) for \( J_{ar}(n) = 0 \) is avoided by the A-MINT constraints. It can also be observed that the closed-form solution of (6) consists of the MINT solution of each sub-system. The proposed A-RAM therefore achieves fast convergence by constraining the optimization problem using \( J_{ar}(n) \). This cost function imposes a constraint to A-MINT and it limits the solution search of \( \hat{g}_i(n) \) to a confined multi-dimensional space such that the reconstructed speech from each sub-system is equivalent. With this confined search space, and will be shown in Section 4, it is therefore expected that the convergence of the algorithm is improved over one without such a constraint. However, it should be noted that the computational load will be higher due to this additional constraint.

Since the number of channels in each sub-system can be equivalent or otherwise, the partitioning of the sub-systems can be symmetric or asymmetric. In addition, it has been studied in [9] that the performance of BSI and equalization algorithms can be limited by the near-common zeros across the channels where such near-common zeros is defined as zeros in the z-domain lying within a vicinity \( \delta \geq 0 \) in terms of Euclidian distance. Since the existence of near-common zeros across all the channels decreases with increasing number of channels, it is therefore expected that the equalization performance between the two sub-systems will differ.

For a symmetric partition, the number of channels in each sub-system is equivalent to \( M/2 \). For an asymmetric partition, one sub-system will contain more channels than the other. It is therefore expected that the sub-system consisting of more channels will have lesser near-common zeros which in turn will outperform the other sub-system having lesser number of channels. Figure 3 illustrates the variation of the number of near-common zeros with the length of RIRs for two subsystems. In this illustrative example, the first sub-system contains three channels while the other contains two channels. The RIRs were generated using the Polack’s model given by
\[ h_i(n) = b_i(n) \exp \left( \frac{-k_{i,\text{int}}}{\alpha} \right) u(n - k_{i,\text{int}}), \] where \( b_i(n) \) is a zero mean Gaussian sequence, \( k_{i,\text{int}} = 100 \) is the initial delay of the \( i \)th channel RIR. The variable \( u(n) \) is a unit step function and \( \alpha \) is related to reverberation time \( T_{60} \) given by
\[ \alpha = (T_{60} f_s) / (3 \ln 10), \] where \( f_s = 16 \text{ kHz}, T_{60} = 400 \text{ ms} \). As can be seen in Fig. 3, the first sub-system containing more channels has lesser number of near-common zeros than that of the second sub-system for all cases of \( L \). For this reason, we propose to partition the microphone array in an asymmetric manner and select the dereverberated speech from the sub-system having more channels.

4. SIMULATION AND EXPERIMENTAL RESULTS
In the following, the signal-to-reverberation ratio (SRR) [1] defined as
\[ \text{SRR} = 10 \log_{10} \frac{\sum_{n=N-k}^{N-k+N-1} s^2(n)}{\sum_{n=N-k}^{N-k+N-1} s(n) - \overline{s}(n)^2} \] is employed to evaluate the equalization performance. The variable \( s_d(n) \) is defined as \( s_d(n) = s(n) * h_d(n) \) where \( h_d(n) \) is the direct path component of the RIR, \( K \) is the number of frames and \( N \) is the length of each frame. For comparison, the A-MINT algorithm [8] is chosen as a reference. Since A-MINT and A-RAM both are adaptive algorithms, the comparison is achieved by comparing the SRR measures iteratively. Throughout the following simulation results, a five-channel microphone array system is partitioned into two sub-systems while the first sub-system contains channel indices 1 to 3, i.e., \( M_1 = 3 \). The performance of the first sub-system from the proposed A-RAM algorithm is chosen to compare with A-MINT.

4.1. Channel equalization using true impulse responses generated from Polack’s model
In this simulation example, we assume we have perfect knowledge of the RIRs, i.e., \( H(n) = h \) while the RIRs with length \( L = 512 \) are generated using the Polack’s model using the same parameters as specified in (14). A WGN is adopted as the input signal. The step-size is chosen to be \( \mu = 0.02 \) for both A-MINT and A-RAM. Figure 4 illustrates their performance when \( \beta = 0 \) and 0.05 with the number of near-common zeros shown in brackets at a tolerance \( \delta = 6 \times 10^{-5} \). We note from (6) that when \( \beta = 0 \), the proposed A-RAM algorithm is equivalent to two independent A-MINT algorithms. As can be seen from Fig. 4, the five-channel A-MINT outperforms the A-RAM algorithm when \( \beta = 0 \). This is due to the increased number of near-common zeros in the three-channel
sub-system compared to the five-channel A-MINT since the number of channels is reduced. However, when the auto-relation function \( J_a(n) \) is taken into account with an empirically determined \( \beta = 0.05 \), the degradation in convergence performance due to the presence of near-common zeros is now offset by the additional constraint imposed on the reconstructed speech from each sub-system. Although the algorithms cannot reach a steady-state since no noise has been added to the RIRs, we can still observe that A-RAM gains approximately 10 dB improvement during the initial convergence.

4.2. Channel equalization using real impulse responses and speech input

In this experiment, we show the performance of A-RAM using recorded RIRs obtained from the Center for Signal Processing Laboratory at Nanyang Technological University, Singapore. The RIRs were measured using a microphone array with interspacing 14 cm at a height of 1.2 m. The signal was collected at a sampling frequency 44.1 kHz. The obtained RIRs were subsequently truncated to a length of \( L = 512 \) samples. In addition, to investigate the significance of the constraint \( J_a(n) \) when the RIRs are not estimated perfectly, we provide an illustrative example by assuming

\[
\hat{h}_i = h_i + v_i, \quad i = 1, 2, \ldots, M, \tag{16}
\]

where \( v_i \) is a vector of estimation error. Her, the \( v_i \) is generated using WGN such that \( 20 \log_{10} ||h_i||/||v_i|| = 20 \) dB with \( M = 5 \) and \( M_1 = 3 \). As can be seen in Fig. 5, with the auto-relation constraint when \( \beta \) is empirically determined as 0.02 and 0.05, the performance of the proposed A-RAM algorithm is improved significantly compared to the five-channel A-MINT algorithm. It can be observed that the proposed A-RAM algorithm converges much faster than that of A-MINT gaining an improvement in the steady-state performance by approximately 15 dB. This result illustrates that the proposed A-RAM algorithm is not only able to achieve higher convergence rate, but also more robust to the estimation error incurred during system identification. We note that since there is no direct relationship between the SRR and the noise added to the RIRs, the SRR increases beyond that of 20 dB.

5. Conclusion

We proposed a sub-system based inverse filtering system for room acoustics. The proposed A-RAM algorithm takes into account the reverberant speech received at the receivers which allow us to exploit the auto-relation of each channel. Extending this auto-relation of individual channel to two sub-systems, we obtained an additional constraint to minimize the difference between the reconstructed speeches governed by the sub-systems. It has been shown that the proposed A-RAM algorithm has a higher rate of convergence in terms of the speech dereverberation compared to A-MINT.

6. REFERENCES