A SYSTEM-IDENTIFICATION-ERROR-ROBUST METHOD FOR EQUALIZATION OF MULTICHANNEL ACOUSTIC SYSTEMS

Wancheng Zhang, Emanuël A. P. Habets, and Patrick A. Naylor

Imperial College London, Department of Electrical and Electronic Engineering

ABSTRACT

In hands-free communications, speech received by a microphone is distorted by room reverberation that can reduce the intelligibility of speech. An approach to dereverberation is firstly to estimate the impulse responses of the acoustic channels between the speaker and the microphones and secondly to design a multichannel equalization system based on the estimated impulse responses. Traditional equalization techniques are designed without the consideration of estimation errors that are commonly introduced by the system identification process. In this work, a System-Identification-Error-Robust Equalization Method (SIEREM) for the equalization of multichannel room acoustic systems is presented. Experimental results for dereverberation using SIEREM applied to estimates of single-input multiple-output acoustic systems with known level of estimation errors show that the proposed equalization design significantly outperforms existing methods in the presence of both synthetic and real system identification errors.

Index Terms— dereverberation, room acoustics, channel estimation error, multichannel equalization, robustness

1. INTRODUCTION

In hands-free communications, in which a speech signal is acquired by microphones placed at some distance from the speaker in a room, the observed speech signals are distorted by room reverberation. Reverberation is the process of multipath propagation of an acoustic sound in an enclosed space and can reduce the intelligibility of speech. The observed reverberant signal generally therefore consists of a direct sound and reflections. The process of dereverberation is to recover the original (dry) speech signal from the signal(s) acquired by a single microphone or multiple microphones. Hereafter, we refer to the multiple acoustic channels from the source to multiple microphones as a multichannel acoustic system.

Acoustic channels are usually modeled as finite impulse response (FIR) filters. An approach to dereverberation is firstly to estimate the impulse responses of the acoustic channels and secondly to design a multichannel equalization system based on the estimated impulse responses [1]. The estimated impulse responses obtained through system identification (SI) always include errors that can be introduced by limitations and inaccuracies of the SI process.

The estimated multichannel acoustic system can be exactly inverted by a set of FIR filters using the multiple-input/output inverse theorem (MINT) [2], subject to specific conditions as explained in Section 3.1. However, the fact that the designed equalization system is able to equalize the estimated acoustic system does not mean that it is able to equalize the true acoustic system. When the designed equalization system is employed to equalize the true acoustic system, the response from the source to the output of the equalization system will deviate from the desired response due to the aforementioned estimation errors. In this paper, our objective is to find a robust equalization system design method that takes into account these system identification errors (SIEs).

The problem is formulated in Section 2. Section 3 reviews existing room acoustics equalization methods and evaluates their performance when SIEs are present. In Section 4, the proposed System-Identification-Error-Robust Equalization Method (SIEREM) is presented. The performance of the proposed method is evaluated in Section 5.

2. PROBLEM FORMULATION

Consider a speech signal \( s(n) \) propagating through an \( M \)-channel acoustic system \( \hat{h} = [h_1^T \ldots h_M^T]^T \) where \( n \) denotes the discrete time index. The acoustic channel between the source and the \( m \)th microphone is characterized by its impulse response \( h_m = [h_m(0) \ h_m(1) \ \ldots \ h_m(L - 1)]^T \), \( m = 1, \ldots, M \), where \( \cdot^T \) denotes the transpose operation. In this work we assume that the acoustic channels are time-invariant. Using the reverberant speech signals

\[
x_m(n) = s(n) * h_m(n) + v_m(n) \quad \text{for} \ m = 1, \ldots, M,
\]

estimates of the room impulse responses (RIRs) \( h_m \) can be obtained using an SI technique [3], where \( \ast \) denotes linear convolution and \( v_m(n) \) denotes additive noise at the \( m \)th microphone.

The estimate of an acoustic system always includes some errors due to under-modeling of channel order and/or the existence of the additive noise. However, as is common practice in the current literature, we assume the channel orders are known or can be correctly estimated. We also note that the acoustic system \( \hat{h} \) can only be estimated up to an unknown scaling factor [3]. Without loss of generality and to be consistent with earlier work [3], we express the estimated acoustic system \( \hat{h} = [h_1^T \ldots h_M^T]^T \) as

\[
\hat{h} = \frac{1}{\gamma} (h - e),
\]

where \( \gamma \) is defined as

\[
\gamma = \frac{\hat{h}^T \hat{h}}{h^T h},
\]

and \( e = [e_1^T \ldots e_M^T]^T \) with \( e_m = [e_m(0) \ e_m(1) \ \ldots \ e_m(L - 1)]^T \). Following the definition of \( \gamma \) in (3), the error vector \( e \) in (2) represents the projection misalignment vector [3] that can be geometrically interpreted as the vector from the projection of \( h \) on the linear manifold of \( \hat{h} \) to \( h \) and therefore is independent of the scaling factor introduced by the SI process. In general, an equalization system \( g = [g_1^T \ldots g_M^T]^T \) where \( g_m = [g_m(0) \ g_m(1) \ \ldots \ g_m(L_i - 1)]^T \) can be computed that satisfies

\[
\sum_{m=1}^{M} \hat{h}_m(i) * g_m(i) = d(i) \quad \text{for} \ i = 0, \ldots, L + L_i - 2,
\]

where \( d(i) \) defines the target impulse response. We now define the equalized impulse response (EIR):
The EIR, $b(i)$, may still distort the speech signal severely due to the second term in (6). It is this problem that is addressed in this paper.

As an example, the true impulse response $h_1$ and the resulting error vector

$$e_1 = h_1 - \gamma h_1,$$

which is part of $e$, are shown in Fig. 1(a) and (b) respectively; (c) shows the energy decay curves (EDCs) of $h_1$ and $e_1$, where the EDC of an impulse response $a = [a(0)\ a(1)\ \ldots\ a(L_\alpha - 1)]^T$ is defined as

$$e(i) = \frac{1}{||a||_2} \sum_{j=1}^{L_\alpha - 1} \alpha^2(j),$$

where $||\cdot||_2$ denotes $\ell_2$-norm.

It can be seen in Fig. 1(b) that the overall temporal shape of the $e_1$ can be modeled by a random sequence with an exponential decay.

3. REVIEW OF EXISTING METHODS

3.1. LS and MINT

The equalization system $g$ can be obtained by solving the system of equations (4), where the target impulse response is given by

$$d(i) = \begin{cases} 0 & \text{if } 0 \leq i < \tau; \\ 1 & \text{if } i = \tau; \\ 0 & \text{otherwise}, \end{cases}$$

with an integer delay $\tau$. In matrix form, (4) can be written as

$$\mathbf{H}g = d,$$

where $\mathbf{H} = [\mathbf{H}_1 \ \cdots\ \mathbf{H}_M]$ and $d = [d(0) \cdots d(L + L_i - 2)]^T$ with $\mathbf{H}_m$ an $(L + L_i - 1) \times L_i$ convolution matrix of $h_m$.

When two or more channels are employed, exact solution(s) to (10) exist when the following two conditions are both satisfied [2]:

1. $L_i \geq L_c$.
2. $L_i \geq L_c = \lceil \frac{1}{3\delta^2} \rceil$ [6], where $\lceil \cdot \rceil$ denotes the smallest integer larger than or equal to $\kappa$. If both conditions are satisfied, the solution with minimum $\ell_2$-norm can be obtained using

$$g = (\mathbf{W}_m)^{-\dagger} d,$$

where $\{\cdot\}^-$ denotes Moore-Penrose pseudo-inverse [8]. If any one of both conditions are violated, (11) gives a multichannel LS solution.

3.2. Weighted LS

The weighted LS method [9] is used for the over-determined cases when the condition 2 is violated. The weighted LS solution is obtained by minimizing the following cost function

$$J = ||\mathbf{W}(\mathbf{H}g - d)||_2^2,$$

where $\mathbf{W} = \text{diag}(\mathbf{w})$ denotes the diagonal weighting matrix with $\mathbf{w} = [w(0) \cdots w(L + L_i - 2)]^T$.

The solution is given by

$$g = (\mathbf{WH})^{-\dagger} \mathbf{W}d.$$

When conditions 1 and 2 are both satisfied and $w(i) \neq 0$, for $i = 0, \ldots, L + L_i - 2$, then the solutions given by (13) and (11) are the same [10].

3.3. Discussion

The approaches reviewed above are generally designed without the consideration of SIEs. The performance of these methods when SIEs are present is now studied for a 2-channel system. The $\hat{h}_1$ and $\hat{h}_2$ obtained in Section 2 are used in this study. The normalized projection misalignment (NPM) of $\hat{h} = [\hat{h}_1^2\ \hat{h}_2^2]^T$ with respect to $h = [h_1^2\ h_2^2]^T$ is $-12$ dB, where NPM is defined as [11]

$$\text{NPM} = \frac{||h - \gamma h||_2^2}{||h||_2^2}$$

with $\gamma$ defined in (3). Parameters are set to $L_i = L_c$, $\tau = 0$.

It should be noted that the exact solution(s) to (10) always exist when $L_i \geq L - 1$ [7]. Unfortunately, this cannot be guaranteed when $L_i \geq L_c$. However, it has been proved in [6] that exact solution(s) exist for almost all cases.
Fig. 2. EDCs of $h_1$ and EIRs obtained with MINT and the proposed SIEREM.

4. PROPOSED METHOD

In this Section, a System-Identification-Error-Robust Equalization Method (SIEREM) is proposed that takes the robustness of equalization systems to SIEs into consideration.

Since the SI process introduces an unknown scaling factor $\gamma$, our equalization system $g$ is designed to equalize $1/(1/\gamma)\hat{h}$, rather than $h$. However, assuming that equalizing $1/(1/\gamma)\hat{h}$ gives
\[
b'(i) = \sum_{m=1}^{M} \frac{1}{\gamma} h_m(i) * g_m(i),
\]
(15)
using (5) we see that the resulting EIR $b(i) = \gamma b'(i)$, i.e., the acoustic system is equalized only up to a scaling factor $\gamma$. Therefore, we obtain
\[
g' = \arg \min_{g} \| W (H + \frac{1}{\gamma} E) g - d \|^2_2
\]
(16)
where $E$ is formed by $e$ and has the same form as $\hat{H}$. Since $e$ is unknown, $E$ is also unknown. In order to find $g$ that minimizes (16), we replace $(1/\gamma)\hat{h}$ by an estimate $\hat{E}$.

As shown in Section 2, the overall temporal shape of $e_m(i)$ can be modeled to a first approximation by a random sequence with an exponential decay rate similar to that of $h_m$. Assuming the power of SIEs is uniformly distributed among the channels, we can use $e_m = [\tilde{e}_m(0) \ldots \tilde{e}_m(L-1)]$, where
\[
\tilde{e}_m(i) = \beta \cdot e_m(i) \cdot e^{-\alpha i},
\]
(17)
to form $\hat{E}$, where $e_m(i)$ is an uncorrelated Gaussian random sequence with zero mean and unit variance, $\beta$ a multiplicative factor related to the power of $\tilde{e} = [\tilde{e}_1^T \ldots \tilde{e}_M^T]^T$ and the decay rate $\alpha$ is related to the reverberation time $T_{60}$ by $[5]$
\[
\alpha = \frac{3 \ln(10)}{T_{60} \cdot f_s},
\]
(18)
It should be noted that the decay rate $\alpha$ can be estimated from $\hat{h}_m$.

With different realizations of the sequence $\varepsilon_m(i)$, (17) provides good and bad estimates of $(1/\gamma)\hat{E}$, and with $(1/\gamma)\hat{E} = \hat{E}$, the performance of $g$ obtained from (16) with different realizations of the sequence $\varepsilon_m(i)$ vary much. However, we desire a $g$ which performs well on average. Therefore, the $g$ that minimizes
\[
J = E \left\{ \| W (H + \hat{E}) g - d \|^2_2 \right\}
\]
(19)
is preferred, where $E\{\cdot\}$ represents expectation operation.

The $g$ that minimizes (19) can be obtained by computing the derivative of $J$ with respect to $g$ and subsequently solving
\[
\frac{\partial J}{\partial g} = 0_{[M,L_i \times 1]}.
\]
(20)
Using (19) and (20), we can obtain
\[
g = (H^T W^T WH + E(\hat{E}^T W^T \hat{E}))^{-1} H^T W^T W d.
\]
(21)
The matrix $R = E(\hat{E}^T W^T \hat{E})$ is a diagonal matrix with $r(j)$, for $j = 1, \ldots, M L_i$, on its diagonal, where
\[
r((m - 1) \cdot L_i + j) = \beta^2 \sum_{i=0}^{L-1} w^2(i + j - 1) e^{-2\alpha i}
\]
(22)
for $m = 1, \ldots, M$ and $j = 1, \ldots, L_i$.

Now we derive the relationship between the multiplicative factor $\beta$ and the NPM of the estimate $\hat{h}$. Since $\varepsilon_m(i)$ is uncorrelated with $\hat{h}_m(i)$, it can be assumed that $E\{[\hat{E}^T \hat{h}] = 0$, i.e., the vector $\hat{e}$ is on average orthogonal to $\hat{h}$.

As shown in [12], we can express the NPM as follows,
\[
NPM = \| e \|^2_2 / \| \hat{h} \|^2_2 = \sin^2(\theta),
\]
(23)
where $\theta$ is the angle between the vectors $h$ and $\hat{h}$. Alternatively, we have
\[
\| e \|^2_2 / \|\gamma \hat{h}\|^2_2 = \tan^2(\theta).
\]
(24)
In order to make on average the NPM caused by $\hat{e}$ equal to the true NPM, we require that
\[
E\{\| e \|^2_2 \} = \frac{1}{\gamma} \| e \|^2_2.
\]
(25)
Using (17), (23), (24) and (25), we can express $\beta$ as
\[
\beta = \frac{\tan \left[ \arcsin \left( \sqrt{NPM} \right) \right]}{\sqrt{M \cdot \frac{e^{-2\alpha \tau_i}}{e^{-2\alpha \tau_i - 1}}} \cdot \| \hat{h} \|_2}.
\]
(26)
Following the above derivation, we conclude that given the decay rate $\alpha$ and the NPM of the SI process we are able to design an equalization system that takes into account the SIEs.

5. PERFORMANCE EVALUATION

In this Section, the performance of the proposed SIEREM is compared with existing methods.

For the intelligibility of the equalized speech signal the suppression of late reflections is more important than the suppression of early reflections [13]. Therefore, it is natural to use some weighting function $w(i)$ for which the amplitudes related to the late reflections are larger than those of the early reflections. We have adopted the following intuitively conceived weighting function that has been verified by informal listening tests to provide good perceptual quality in the equalized speech signal:
\[
w(i) = \begin{cases} 
1 & \text{if } 0 \leq i \leq \tau; \\
e^{\alpha(i-\tau)} - 1 & \text{if } i > \tau.
\end{cases}
\]
(27)
In order to evaluate the performance of the equalization system the following measures are used in this work. The first performance measure is the EDC as defined in (8). The second performance measure is the early-to-late reverberation ratio (ELR), also known as the Clarity Index, which is defined as [5]
measures in Table 1. The decay rate (or reverberation time) and the NPM level. These parameters are not a priori known and therefore need to be estimated. Due to space limitations, elaborate testing results of the robustness of the SIEREM with respect to these design parameters cannot be provided here. In our tests, we have found that the SIEREM is not significantly sensitive to the $T_{60}$ and NPM; estimation accuracy of 20% in $T_{60}$ and ±10 dB in NPM is sufficient for typical operation.

6. Conclusion

In this paper we discussed the equalization of multichannel room acoustic systems, when the multichannel room impulse responses include estimation errors introduced by the system identification process. It was shown that existing equalization techniques fail to equalize the acoustic systems in the presence of system identification errors. Therefore, a new method which takes into account system identification errors was proposed. The evaluation results indicate that the proposed method is more robust to system identification errors than MINT.

Acknowledgment

The authors wish to thank Dr. Jimi Wen for insightful discussion on perceptual characteristics of the RIRs.

7. References


