A ROBUST SPARSE APPROACH TO ACOUSTIC IMPULSE RESPONSE SHAPING

Lakshmi Krishnan, Paul D. Teal

School of Engineering and Computer Science
Victoria University of Wellington
Wellington, New Zealand

Terence Betlehem

Callaghan Innovation
Lower Hutt
New Zealand

ABSTRACT

Impulse response shaping is a technique for partly equalizing impulse responses. In acoustics, it can be used for the reproduction of audio signals mitigated by distortions in a room. The most significant phenomenon among the distortions is reverberation, a straightforward characterization of which is the room impulse response. Room responses can be characterized but could contain measurement errors or noise. In addition, room responses vary with changes in atmospheric conditions such as temperature and humidity and also due to change in positions inside a room. The design of a shaping filter robust to at least some of these variations is likely to be very useful, which is considered in this work. The method uses a computationally efficient approach based on Basis Pursuit DeNoising (BPDN).

Index Terms— Sparse estimation, acoustic impulse response shaping, cross-talk, direct response, fast Dual Augmented Lagrangian Method (fast DALM)

1. INTRODUCTION

Acoustic impulse response shaping may be used to selectively equalize the effects of reverberation to points inside a room. It can be applied to single channel equalization and also multi-channel equalization and cross-talk cancelation. The shaping filters are implemented as pre-filters, placed before the loudspeakers, so that the listener hears sound that are perceptually improved. The technique can be framed as an optimization problem, that involves the minimization of a norm of some constraints involving shaping filters. In this paper, we propose a computationally efficient approach for robust impulse response shaping, based on an ℓ_1-norm approach. The constraints are chosen so as to maintain the direct path and the early reflections, that are perceptually useful [1], penalizing late reverberation that distort the signal. Cross-talk cancelation formulations that use a minimax approach to minimize the ℓ_∞-weighted norm of the error between the shortened response and a desired response are presented in [2, 3]; [2] addresses both cross-talk cancelation and impulse response shortening jointly. A different formulation based on the categorization of the room response into wanted part (direct path and early reflections) and unwanted part (late reverberation), which minimizes the ratio between the two parts, is presented in [4, 5]. It uses a gradient descent-based approach to find a feasible solution. A robust implementation based on this formulation is proposed in [6, 7], in which the shortening filter design is performed over multiple microphone positions at a radius from actual position, so that the filter is robust to small changes in microphone position. This approach utilizes a detailed theoretical analysis of the errors due to inaccurate measurements and small changes in microphone positions in [8]. Another robust formulation is to extend Relaxed Multichannel Least Squares (RMCLS) to control the level of coloration to maintain robustness in the presence of system identification errors (SIE) [9].

The work presented in this paper uses a computationally fast method [10] based on sparse estimation for a robust implementation. Shaping filters robust to changes in microphone positions are designed by averaging over channel realizations at multiple microphone positions as in [6]. Computational efficiency is an important requirement for a practically realizable shaping approach as typical room responses may have tens of thousands of samples resulting in huge optimization problems. In impulse response shaping, it is challenging to achieve both high computational speed and good performance, which is tackled here using a sparse approach. This paper is organized as follows. Section 2 formulates the optimization problem and the application of sparse estimation to acoustic impulse response shaping. Section 3 discusses the experimental results. The approach is shown to be more robust than both conventional shaping and inverse

filtering using a plot of cumulative density function (CDF) of cross-talk over a number of microphone positions.

2. PROBLEM DEFINITION

Acoustic room impulse response shaping can be achieved by solving an optimization problem to find a set of shaping filters, to reduce the unwanted effects of reverberation in the room and cross-talk from the microphones. In cross-talk cancelation problem, signals of s sources are delivered to m microphones through l loudspeakers. Consider the system shown in Fig. 1. Let L be the number of loudspeakers and M the number of microphones. The particular case for $L = 3$ and $M = 2$ is shown in Fig. 1. Here, c_{ml} represents the channel response from loudspeaker l to microphone m and h_{ls} represents the cross-talk canceling filter from source s to loudspeaker l.

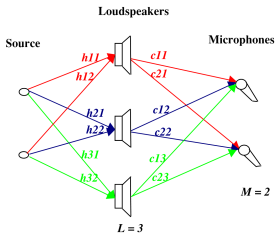

Fig. 1: A loudspeaker and microphone setup showing acoustic channels c_{ml} and cross-talk cancelation filters h_{ls}

For each path, the combined filter and channel response (called the Global Impulse Response, GIR) r_{ms} is found [2] as the sum of the convolutions of c_{ml} and h_{ls} over all the loudspeakers. For the special case shown in Fig. 1, a matrix equation for GIR's can be written as $r = Ch$ or

$$\begin{bmatrix} r_{11} \\ r_{21} \\ r_{12} \\ r_{22} \end{bmatrix}_{4N_\mathrm{r}\times 1} = \begin{bmatrix} \mathcal{C} & 0 \\ 0 & \mathcal{C} \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{21} \\ h_{31} \\ h_{12} \\ h_{22} \\ h_{32} \end{bmatrix}_{6N_\mathrm{h}\times 1} \tag{1}$$

where N_h is the length of each shaping filter h_{ls}, N_r is the length of each GIR r_{ms} and $\mathcal{C}$ is the matrix defined as

$$\mathcal{C} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \end{bmatrix} \tag{2}$$

of dimension $N_\mathrm{r}M \times N_\mathrm{h}L$, consisting of the Toeplitz convolution matrices C_{ml}. The optimization problem to be solved for finding the vector of shaping filters (h) given by

$$\min_{h} \quad \|W(Ch - r)\|_2^2 + \lambda \|h\|_1 \tag{3}$$

where h and r are defined as in (1) and W is a diagonal weighting matrix defined in [2]. The ℓ_1-norm of the variable to be estimated is the regularization term that ensures a sparse solution [11] and the ℓ_2-norm representing the constraint is the fidelity factor. λ is a regularization parameter that decides the relative importance of the two terms. The weighting coefficients are chosen so as to penalize the cross-talk (r_{ms}, $m \neq s$) and late reverberation and pre-echo in the direct channels (r_{mm}) heavily, but penalize the early reverberations lightly [2, 4]. The ℓ_1-regularized ℓ_2-norm minimization is chosen for computational efficiency reasons [10].

2.1. Sparse Approach

A sparse vector is one that has few non-zero elements. The sparse approach was adopted here for two reasons. First, using a regularized ℓ_1-norm of h improves the robustness of the solution by reducing the energy of h (though not as directly as does the ℓ_2-norm). Second, this allows use of the high computational efficiency and speed of iterative sparse reconstruction algorithms like fast DALM [12–14] to estimate a feasible solution for h. Computational efficiency is important because a realistic adaptive shaping approach has to deal with long room responses. The convolution matrices may then contain billions of elements. Such problems can lead to long execution times when the previously published approaches are used. In addition to computational efficiency and low energy filters, the sparse approach also allows sufficient length in h to control long delays in c, but with fewer non-zero taps.

The sparse approach finds a sparse estimate for h that solves the minimization problem posed in (3). The solution can be sparse if the length of h (all the filters h_{ls} concatenated) is greater than that of r (all the GIRs r_{ms}) so that the matrix C becomes an overcomplete dictionary [15]. The sparse estimation problem, thus, tries to find a sparse estimate for h that satisfies the equation $Ch = r$ with minimum error [15]. This is done by minimizing an objective function as in (3) which is known as Basis Pursuit DeNoising (BPDN). This optimization problem can be solved in a computationally efficient manner by using iterative sparse reconstruction algorithms such as Fast DALM [12] and FISTA [16].

2.2. Algorithm

The sparse reconstruction algorithm used is Fast DALM [12]. The algorithm minimizes the augmented dual of the

objective function of (3) given by [12]

$$\min_{h,y,z:z\in B_1^\infty} -Wr^Ty - h^T(z - WC^Ty) + \frac{\beta}{2}\left\|z - WC^Ty\right\|_2^2 + \frac{\lambda}{2}y^Ty \tag{4}$$

where B_1^∞ is the B_1 ball [12, 13], y is a dual variable, z is the dual variable corresponding to projection onto the B_1 ball, β is a regularization parameter and the third term containing an ℓ_2-norm is the augmentation term. The algorithm can be summarized as a set of iterative equations [12]. We use a Fourier implementation to replace multiplication by convolution matrices, improving computational speed further.

The approach in (3) can be extended to a robust estimation approach by performing the design over multiple realizations of each microphone position, so that the estimate is robust to perturbations of the microphone position [6]. The positions are taken to be within a fixed radius around the expected microphone positions [8] so that the estimate is still accurate for slight movement of the microphone within the radius. The objective function for such an implementation can be written as

$$\min_h \quad \sum_{n=0}^{N} \left\|W(C^{(n)}h - r)\right\|_2^2 + \lambda\|h\|_1 \tag{5}$$

where $C^{(n)}$ are the Toeplitz matrices generated from the perturbed as well as the actual channels and N is the number of perturbed channels. The perturbed channels are generated as

$$c^{(n)} = c^{(0)} + p^{(n)} \tag{6}$$

where $p^{(n)}$ are the perturbations and $c^{(0)}$ is the unperturbed room response. The perturbations are generated by performing frequency and time shaping on white Gaussian noise to generate the perturbation characteristics. The frequency domain shaping is performed by multiplying the Fourier Transform of the Gaussian noise [8] by the power spectrum at perturbed positions in a diffuse field

$$P(\omega) = \|C(\omega)\|\sqrt{2 - 2\text{sinc}\left(\frac{\omega d}{v}\right)} \tag{7}$$

where d is the microphone displacement and v is the speed of sound. The Inverse Fourier Transform of the frequency shaped response is multiplied by the time shaping function [6] given by

$$p(t) = \begin{cases} 0 & t < t_0 - d/v \\ 1 & t_0 - d/v < t < t_0 + d/v \\ e^{\frac{-3\ln(10)(t-t_0-d/v)}{T60}} & t \geq t_0 + d/v \end{cases} \tag{8}$$

Table 1: Performance for different microphone displacements for a channel of length 10^4

d	N_h	λ	CDR (dB)
1	5×10^4	0.02	-44
2	5×10^4	0.02	-40
4	5×10^4	0.02	-33

where t_0 is the time taken by the direct component. This two step iteration generates channel responses synthetically with same acoustic properties as the room responses.

3. EXPERIMENTAL RESULTS

The robust problem in (5) was tested for channels measured in a small room at sampling rates of 16 kHz and 44.1 kHz with $L = 3$, $M = 2$ and $N = 14$. The synthesized positions were considered in a circle around the original microphone positions at specific displacements d of 1 cm, 2 cm and 4 cm for the robust filter estimation. The typical estimation results using a sampling rate of 16 kHz are tabulated in Table 1. The weighting matrix W was set according to [2]. The initial values for the algorithm were set as a delta function for each of the inverse filers. The performance metric used in this work is the Cross-talk to Direct Response ratio (CDR) which is the ratio of maximum cross-talk to maximum direct response (inverse of DSCR in [7]). The performance of the algorithm was tested for different values of the control parameter λ, as can be seen from Table 1. The simulations were run on a desktop computer with 4 GB RAM and 3 GHz Intel Processor using Matlab R2012b. The feasible values for λ for this problem were in the range $0 \leq \lambda \leq 0.2$. In this paper, we used $\lambda = 0.02$. Also, the results show that fast DALM performs quite well for the robust problem, with good crosstalk cancelation of around -40 dB, in addition to fast convergence.

From Table 1, it can be seen that the typical CDR value for the robust implementation is around -40 dB, which is almost 20 dB better than the values for the ℓ_2-norm implementations provided in [7]. For the implementations without considering spatial mismatch, CDR was found to be around -50 dB which is comparable to the values for the ℓ_p-norm implementations provided in [7]. The execution time taken for robust implementation with $N_r = 10\,000$ as in [7] was 1 minute, while for the non-robust shaping, it took 8.9 s. For the results shown in Table 1 with $N_r = 60\,000$, the computational time taken in all three cases were less than 10 minutes. For the non-robust shaping, the execution time was observed

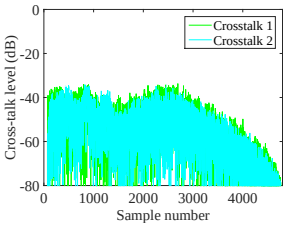

Fig. 2: Crosstalk responses for $N_{\mathrm{r}} = 5000$ and $N_{\mathrm{c}} = 1500$

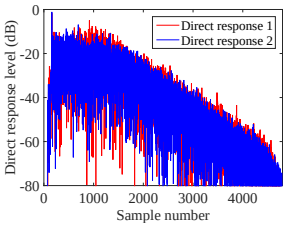

Fig. 3: Shaped direct responses for $N_{\mathrm{r}} = 5000$ and $N_{\mathrm{c}} = 1500$

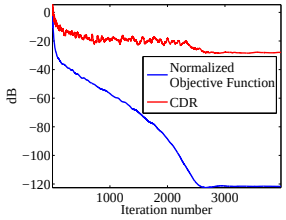

Fig. 4: Objective function value and CDR(dB) with iteration number for $N_{\mathrm{r}} = 5000$ and $N_{\mathrm{c}} = 1500$

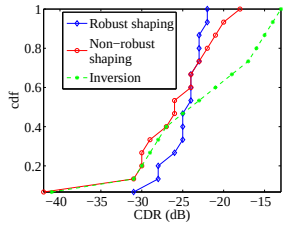

Fig. 5: Cumulative density functions of CDR at the original and perturbed positions for $N_{\mathrm{r}} = 9600$ and $N_{\mathrm{c}} = 4000$

to be 247 s. Thus, these results show that the sparse approach for shaping filter design provide good performance with results comparable to [7] in addition to high computational efficiency.

Typical plots obtained for cross-talk and shaped responses are shown in Fig. 2 and Fig. 3 respectively. The shaped responses in Fig. 3 shows the peak is around 10 dB above the smaller values and the response decays rapidly after the allowed early reflections. The shaping approach maintains the direct path (peak) and some of the early reflections whilst reducing late reverberation, thus achieving good shaping. Fig. 4 shows the objective function value in dB versus iteration number demonstrating the fast convergence of fast DALM. Fig. 4 also shows the maximum cross-talk to maximum direct response ratio in dB with iteration number, demonstrating the cross-talk cancelation performance of the algorithm.

The cumulative density function of the CDR values were plotted for the original channel and the 14 perturbed versions (Fig. 5) to study the efficacy of the robust implementation. The plots show that the robust implementation (blue) shows satisfactory performance at the original as well as perturbed positions. If the filters are designed only using the measured channel, the performance at the perturbed channels is shown in green. This demonstrates the value of optimising over perturbed channels so as to generate a robust solution. Classical channel inversion [17] provides considerably degraded cross-talk cancelation compared to the channel shaping approach at the perturbed positions. Thus, it was concluded that sparse shaping filter design over multiple positions is an efficient method for robust acoustic crosstalk cancelation and GIR shaping.

4. CONCLUSION

A robust approach using sparse estimation of inverse filters for acoustic crosstalk cancelation is discussed in this paper. The optimization problem is solved using an iterative sparse reconstruction algorithm called fast DALM, which is demonstrated to be effective in impulse response shaping and crosstalk cancelation. It was found that this algorithm performed well, providing good cross-talk cancelation and GIR shaping at the original as well as perturbed microphone positions, in addition to faster computation compared to gradient projection method, previously used for solving this problem. Therefore, it was concluded that the sparse robust shaping filter design using fast DALM as estimation algorithm is a good choice for crosstalk cancelation and impulse response shaping.

5. REFERENCES

[1] I. Kodrasi, S. Goetze, and S. Doclo, "A perceptually constrained channel shortening technique for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, May 2013, pp. 151–155.

[2] T. Betlehem, P. D. Teal, and Y. Hioka, "Efficient crosstalk canceler design with impulse response shortening filters," in *Proc.IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, March 2012, pp. 393–396.

[3] H. I. K Rao, V. John Mathews, and Y. C Park, "A minimax approach for the joint design of acoustic crosstalk cancellation filters," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2287–2298, Nov 2007.

[4] A. Mertins, T. Mei, and .M Kallinger, "Room impulse response shortening/reshaping with infinity- and pnorm optimization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 249–259, 2010.

[5] R. Mazur, J. O. Jungmann, and A. Mertins, "Optimized gradient calculation for room impulse response reshaping algorithm based on p-norm optimization," in *Proc.IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, March 2012, pp. 185–188.

[6] J.O. Jungmann, R. Mazur, and A. Mertins, "Perturbation of room impulse responses and its application in robust listening room compensation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, May 2013, pp. 433–437.

[7] J. O. Jungmann, R. Mazur, M. Kallinger, T. Mei, and A. Mertins, "Combined acoustic mimo channel crosstalk cancellation and room impulse response reshaping," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1829–1842, Aug 2012.

[8] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 311–319, May 2000.

[9] Zhang W. Habets E.A.P Lim, F. and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1379–1390, Sep 2014.

[10] L. Krishnan, T. Betlehem, and P. D. Teal, "A sparsity based approach for acoustic room impulse response shortening," in *Proc. IEEE Statistical Signal Processing Workshop(SSP2014)*, Jun 2014.

[11] D. L. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[12] A. Yang, A. Ganesh, Z. Zhou, S. Sastry, and Y. Ma, "Fast ℓ_1 minimisation algorithms and an application in robust face recognition: a review," *Technical Report UCB/EECS-2010-13*, 2010.

[13] R. Tomioka, T. Suzuki, and M. Sugiyama, "Superlinear convergence of dual augmented lagrangian algorithm for sparsity regularized estimation," *J. Machine Learning Research*, vol. 12(2011), pp. 1537–1586, May 2011.

[14] P. E. Gill and D. P. Robinson, "A primal-dual augmented lagrangian," *J. Computational Optimization and Applcns*, vol. 51, no. 1, Jan 2012.

[15] S. F. Cotter and B. D. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Trans. Communications*, vol. 50, no. 3, pp. 374–377, Mar 2002.

[16] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[17] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, no. 2, pp. 145–152, Feb 1988.