MAXIMUM A POSTERIORI ESTIMATION OF ROOM IMPULSE RESPONSES

Dinei Florencio and Zhengyou Zhang

Microsoft Research, Redmond, WA, USA

ABSTRACT

Estimating room impulse responses (RIRs) has a number of applications, including personalized audio, analyzing and improving acoustic behavior of concert halls, listening room compensation, sound source localization, and many others. RIRs have been estimated in essentially the same fashion for the last 50 years: Compute the cross correlation between a signal played at point A, and the signal received at point B. Best results are obtained when the signal played is white noise, or a maximum length sequence. No prior knowledge is exploited in computing the RIR, which is simply assumed to be the cross correlation between played and received signals. In contrast, research in *adaptive* RIR estimation (a.k.a. adaptive Acoustic Echo Cancellation) has made huge progress by (among other things) incorporating models for the RIR. In this paper we propose a new RIR estimation technique, based on a *maximum a posteriori* formulation. More specifically, we estimate the room reverberation time, as well as the room noise level, and use those as priors for the RIR estimation. Comparison with ground truth shows an average improvement of 12 dB compared to traditional methods.

Index Terms— RIR, room acoustics, room impulse response, AEC, echo cancellation.

1. INTRODUCTION

When a sound is played within an enclosure (e.g., a room), it propagates in very complex ways. It reflects off walls and other surfaces, it diffracts, it is absorbed by air and by these surfaces, and does all that in a frequency dependent manner [1, 2]. In general, this is further complicated by the directionality of the sound source as well as the directionality of the capturing device (e.g., a microphone or the human ear) [3]. Thus, typically, even if the exact geometry of a room is known, modelling the transfer function between two points in a room is extremely hard, except for crude approximations.

Knowledge of that transfer function is, however, extremely useful in a number of scenarios. Common application include acoustic echo cancellation (AEC) [4, 5, 6, 7], personalized audio [8, 9], analysing and improving acoustic behaviour of concert halls [10, 11], listening room compensation [12, 13, 14], sound source localization [15, 16, 17], and many others [18, 19, 20, 19, 21].

Due to the aforementioned difficulty in modeling acoustic behavior, room impulse responses (RIR) are typically experimentally measured directly in the real environment [22]. More specifically, a sound source and a microphone are placed at each end-point, a sound is played and recorded, and that recorded signal is used to compute (an estimate of) the RIR between those two points[1]. Furthermore, if we need estimates of the RIR in points other than the measured ones, RIR interpolation may be needed, which is, in itself, another challenging problem[23].

In this paper we introduce a technique that makes use of prior knowledge of acoustic behavior to improve the quality of the RIR estimates. Our *maximum a posteriori* (MAP) estimate incorporates a simple model for ambient noise and an exponential decaying model for reverberation. Yet, results are quite appealing: an average 12dB improvement over traditional RIR estimation.

A closely related problem is that of *adaptive* RIR estimation (typically referred as adaptive AEC). In adaptive AEC, one assumes the RIR is changing. That is much more challenging problem, and one has to balance the adaptation parameter according to the dynamics of the system, as well as the ambient noise. State of the art adaptive AEC use models of the RIR to control step size and related parameters [24, 25, 26, 27, 28]. While they greatly improve adaptation, for a fixed system (as in our problem), they underperform a simple matched filter (cross correlation), which (as us) uses the whole available signal at once. In a way, our paper brings the RIR estimation in line with the last 50 years of progress in adaptive AEC.

The remainder of this paper is organized as follows: Section 2 introduces our notation and explains the traditional RIR estimation method. Section 3 presents our MAP formulation, and Section 4 presents and discusses the experimental results. Finally, Section 5 presents our main conclusions and future directions.

2. TRADITIONAL RIR ESTIMATION

To estimate the room impulse response (RIR) between a loudspeaker and a microphone, a signal is played through the loudspeaker and recorded at the microphone. Without loss of generality, let us assume the first K samples of an RIR are a reasonable approximation of the RIR for the intended application. Let us denote the training signal as the $N \times 1$ vector $\mathbf{x}$, and the corresponding signal received at the microphone as $\mathbf{y}$. To simplify notation (and without loss of generality), we assume $\mathbf{x}$ has unit variance. Neglecting boundary effects to simplify the notation, we model the signal $\mathbf{y}$ as:

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{a} \tag{1}$$

where $\mathbf{h}$ is the $K \times 1$ vector corresponding to the room impulse response, $\mathbf{X}$ is the $N \times K$ matrix composed of K time-shifted versions and $\mathbf{x}$, and $\mathbf{a}$ is the ambient noise, which we assume to be independent from $\mathbf{x}$.

[1] In applications like AEC, the path of interest in is between the loudspeaker and a microphone, and both are under the control of the system. In other scenarios, loudspeakers and/or microphones may have to be placed at one of both the desired end points.

The traditional way of estimating the RIR is by playing a white training sequence $\mathbf{x}$, and correlating it with the received signal $\mathbf{y}$ [29]. More specifically, by computing:

$$\hat{\mathbf{h}} = \mathbf{X^T y}. \tag{2}$$

The reasoning for this approximation can be observed by inserting the expression for $\mathbf{y}$ from (1) into (2), i.e.:

$$\begin{aligned} \hat{\mathbf{h}} &= \mathbf{X}^T(\mathbf{X}\mathbf{h} + \mathbf{a}) & (3) \\ &= (\mathbf{X}^T\mathbf{X})\mathbf{h} + \mathbf{X}^T\mathbf{a}. & (4) \end{aligned}$$

By taking the expected value of (4), we get:

$$E\{\hat{\mathbf{h}}\} = \mathbf{R_{xx}}\mathbf{h} + \mathbf{r_{xa}} \tag{5}$$

where $\mathbf{R_{xx}} = E\{\mathbf{X}^T\mathbf{X}\}$ (i.e., the K-lag autocorrelation matrix of $\mathbf{x}$), and $\mathbf{r_{xa}} = E\{\mathbf{X}^T\mathbf{a}\}$ (i.e., the K-lag cross-correlation vector between $\mathbf{x}$ and $\mathbf{a}$). Note that, if the training sequence is white, than $\mathbf{R_{xx}} = \mathbf{I}$, and, since the noise is assumed independent of the signal $\mathbf{r_{xa}} = \mathbf{0}$. Thus, $\mathbf{X}^T\mathbf{y}$ is an unbiased estimate of the RIR $\mathbf{h}$. Furthermore, as the length of the training sequence increases, the estimate becomes increasingly close to the true $\mathbf{h}$, converging to the correct value as $N \to \infty$.

For finite training sequences, however, these are just approximations. Indeed, we can rewrite (4) as:

$$\hat{\mathbf{h}} = \mathbf{h} + \left(\mathbf{X}^T\mathbf{X} - \mathbf{I}\right)\mathbf{h} + \mathbf{X}^T\mathbf{a} \tag{6}$$

Thus, we see that there are two noise terms affecting the traditional way of estimating $\hat{\mathbf{h}}$, both related to the fact that $\mathbf{x}$ is finite. The first one reflects the non-diagonal nature of a finite signal covariance matrix, while the second term originate from the random correlations between two (independent) finite signals. The first can be alleviated by making the training sequence longer, and the second can be alleviated by making the loudspeaker signal stronger, or the sequence longer.

Note that the non-diagonal elements of the covariance matrix can also be reduced by using a special kind of white noise, called *maximum length sequences* (MLS) [30, 31, 32]. These special pseudo-random sequences, do improve the quality of the estimates. However, they sound like white noise, and cannot be modified much. As such, they are not appropriate for most applications where users will be present during the measurements (including our target application of initial AEC filter estimation).

Figure 1 shows the $\hat{\mathbf{h}}$ estimation error as a function of training sequence duration, for ambient noise levels of 0 dB and 40 dB SNR (plotted as $10\log_{10}(\sum_i(\hat{\mathbf{h}}[i] - \mathbf{h}[i])^2 / \sum_i(\mathbf{h}[i])^2)$). The quality of estimates vary widely with room characteristics, distance, etc. The data in Figure 1 was obtained using the parameters that are closest to our target application in AEC, and are described in more detail in Section 4. As expected, the quality of the estimate increases by roughly 3dB for every doubling of the training sequence length. However, to get to reasonable estimates, reasonably long sequences have to be used. For example, at 40dB SNR level, a sequence with 262k samples, yields only a 23dB approximation of the true RIR. In other words, we need 16 seconds of a training sequence to get a just reasonable 23 dB approximation.

In many cases, however, we cannot afford too long sequences. While for acoustic analysis of a concert halls, it might be reasonable to play several minute-long training sequences, this is certainly not the case when the RIR is being used as initialization for a acoustic echo cancellation in a communication system.

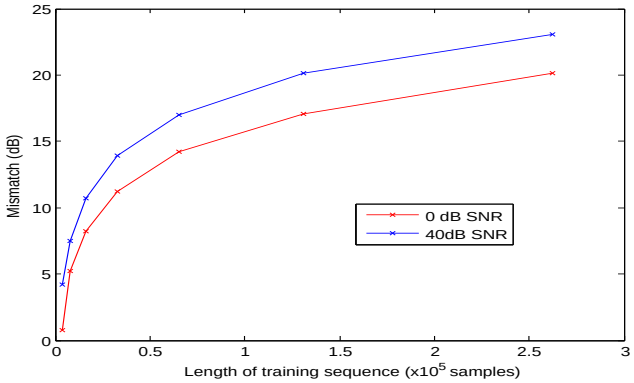

Fig. 1. Mismatch between true and conventional RIR estimation as a functions of training sequence length. Results correspond to a room with $T_{60} = 300ms$, a microphone at $2.2m$ from the loudspeaker, and $F_s = 16KHz$. Estimated filter is also $300ms$ long.

3. MAP ESTIMATION

We now propose a *maximum a posteriori* (MAP) formulation to estimate $\mathbf{h}$. Under the MAP formulation, we have:

$$\begin{aligned} \mathbf{h}_{MAP} &= \arg\max_{\mathbf{h}} \{f(\mathbf{h}|\mathbf{y})\} & (7) \\ &= \arg\max_{\mathbf{h}} \{f(\mathbf{h})f(\mathbf{y}|\mathbf{h})/f(\mathbf{y})\} & (8) \\ &= \arg\max_{\mathbf{h}} \{f(\mathbf{h})f(\mathbf{y}|\mathbf{h})\} & (9) \end{aligned}$$

where $f(\cdot)$ denotes the probability density function. Note that the probability $f(\mathbf{y}|\mathbf{h})$ can be computed by estimating the the probability of the noise $\mathbf{a}$ being $\mathbf{y} - \mathbf{X}\mathbf{h}$.

We assume that the ambient noise $\mathbf{a}$, and the RIR $\mathbf{h}$ are multivariate Gaussian random variables, i.e., $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathbf{a})$, and $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathbf{h})$, where $\Sigma_\mathbf{a}$ and $\Sigma_\mathbf{h}$ are the covariance matrices for $\mathbf{a}$ and $\mathbf{h}$, respectively. Note that, while a Gaussian is probably not the best model for $\mathbf{h}$, it does capture its key statistical characteristics, and it makes the mathematical treatment much simpler than if we assume a more complex model. Thus, based on the Gaussian assumption, we have:

$$f(\mathbf{h}) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma_\mathbf{h})}} \exp\left(-\frac{1}{2}\mathbf{h}^T\Sigma_\mathbf{h}^{-1}\mathbf{h}\right), \tag{10}$$

and

$$f(\mathbf{y}|\mathbf{h}) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma_\mathbf{a})}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{h})^T\Sigma_\mathbf{a}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{h})\right). \tag{11}$$

Inserting (10) and (11) in (9), taking the log, and disregarding the constants and the log itself (since we only care about the $\mathbf{h}$ at the maximum, but not the maximum value *per se*), we get:

$$\mathbf{h}_{MAP} = \arg\max_{\mathbf{h}} \left\{ -\mathbf{h}^T \Sigma_{\mathbf{h}}^{-1} \mathbf{h} - (\mathbf{y} - \mathbf{X}\mathbf{h})^T \Sigma_{\mathbf{a}}^{-1} (\mathbf{y} - \mathbf{X}\mathbf{h}) \right\} \tag{12}$$

Taking the derivative of the function being maximized in the right side of (12) and making it equal to zero, we get:

$$\Sigma_{\mathbf{h}}^{-1} \mathbf{h}_{MAP} - \mathbf{X}^T \Sigma_{\mathbf{a}}^{-1} (\mathbf{y} - \mathbf{X}\mathbf{h}_{MAP}) = \mathbf{0} \tag{13}$$

Which can be rewritten as:

$$\mathbf{h}_{MAP} = \left(\mathbf{X}^T \Sigma_{\mathbf{a}}^{-1} \mathbf{X} + \Sigma_{\mathbf{h}}^{-1} \right)^{-1} \mathbf{X}^T \Sigma_{\mathbf{a}}^{-1} \mathbf{y} \tag{14}$$

Thus, to estimate $\mathbf{h}_{MAP}$, besides the received signal $\mathbf{y}$, we need estimates of the covariance matrices of the ambient noise and of the taps of the RIR. These are, in effect, our priors on the the ambient noise, and on the RIR itself. We discuss our method for obtaining them next.

3.1. Estimating the priors $\Sigma_{\mathbf{a}}$ and $\Sigma_{\mathbf{h}}$

The covariance matrix $\Sigma_{\mathbf{a}}$ captures the characteristics of the ambient noise. Assuming such noise is stationary, a reasonable estimate can be easily computed during silence periods, e.g., before playing the training sequence $\mathbf{x}$. Thus, in effect it is truly *a priori* knowledge.

As a proof of concept, and a demonstration of the potential of the technique, we have chosen to use a reasonably simple model for $\Sigma_{\mathbf{h}}$. We already simplified the model by assuming the taps of $\mathbf{h}$ are jointly Gaussian. While the result in (14) would be valid for a generic $\Sigma_{\mathbf{h}}$, we further simplify our method by assuming the taps are independent, and that they decay at an exponential rate with a time constant proportional to the room reverberation time T_{60}. The independence assumption makes $\Sigma_{\mathbf{h}}$ a diagonal matrix. Then, each element of the diagonal corresponds to the estimate of the variance of that tap [33, 34]:

$$\sigma_{\mathbf{h}}^2[k] = A.10^{-6k/F_s T_{60}}, \tag{15}$$

where A is a scaling constant to account for the starting sound level. Note that this simplified model disregards the fact that the direct path does not obey the same exponential decay as the rest of the reverberation, and the sparseness of early reverberation. The constants A and T_{60} are estimated by computing the cross correlation between $\mathbf{x}$ and $\mathbf{y}$ (i.e., the RIR using the traditional method), and fitting a function in the form of (15) to the later part of the correlation (to avoid the direct path and early reflections). More specifically, we compute $\hat{\mathbf{h}}$ using Eq. (2), and estimate A and T_{60} as:

$$\{A, T_{60}\} = \arg\min_{A, T_{60}} \sum_{k=k_s}^{K} \left((\hat{\mathbf{h}}[k])^2 - A.10^{-6k/F_s T_{60}} \right)^2, \tag{16}$$

where k_s is the starting point, chosen to help ignore the direct path and the early reflections.

Our estimation of the covariance matrix of the taps of the RIR, $\Sigma_{\mathbf{h}}$, is based on the actual received signal $\mathbf{y}$. Thus, rigorously speaking, it is not *prior* knowledge. The prior is effectively the exponential decay assumption. In other words, there are two estimation steps. In the first, we use the received signal to estimate $\Sigma_{\mathbf{h}}$. Then, on a second step, we estimate $\mathbf{h}_{MAP}$ using our estimate of $\Sigma_{\mathbf{h}}$ as prior. Because the first step is simply an estimation of two real numbers (A and T_{60}), we assume it is reasonably robust, and don't consider this as a significant source of error, and focus mostly on the second step. Although subtle, we make that distinction to clarify our use of the "MAP" term in naming our estimation algorithm.

4. EXPERIMENTAL RESULTS

As we mentioned earlier, obtaining true RIRs is challenging, so the standard way of evaluating RIR estimation methods is based on synthetic data. We compare the results of our MAP method with the "traditional" correlation method, both using a random training sequence as well as a *maximum length sequence*. We denote these by TRAD and MLS, respectively, while refer to our method as MAP.

For synthetic data, we can directly compare the estimated RIR with the "actual" one, and precisely measure the estimation error. Table 1 shows the results for ambient SNRs varying from 0 to 30dB, and sequences lengths of 4095, 8191, and 16383. Note that MLS sequences are restricted to certain lengths (e.g., $2^N - 1$), so we evaluate results only at those lengths, even though the MAP and TRAD methods do not have such restriction. The simulated room is a rectangular box measuring $2.9 \times 3.7 \times 5.4$ meters, with a T_{60} of 300ms, and a microphone at 2.2 m from the loudspeaker. Sampling frequency is 16KHz, and the estimated filter length is 300ms (thus $K = 4800$). The "true" RIR $\mathbf{h}$ was obtained using the image method [35]. For TRAD and MAP, a different training sequence is generated each trial, as a pseudo-random gaussian noise with the desired length. For the MLS the training sequence is computed as in [31]. For all cases, the received signal $\mathbf{y}$ was then computed by Eq. (4), and used as input for the estimation method. Results are averaged over 100 trials with different environment (white) noise.

For the MAP method, we compute A and T_{60} by using (16) with $k_s = 800$ and $K = 4800$, and use a quadratic error parametric fitting based on the *minsearch* function from Matlab 2013Ra.

As it can be seen in Table 1, the MAP estimation is better than TRAD and MLS in all cases. The gains over TRAD vary from 5.5 to 19.4 dB, with an average improvement of 11.8 dB. Compared to MLS, the gains are smaller, varying from 1.3 to 6.8 dB, with an average improvement of 2.8 dB.

Figure 2 shows an example of the true and estimated RIRs by the different methods. As expected, the TRAD method is much noisier, and that estimation noise does not decrease towards the tail of the RIR. The estimation based on MLS has significantly less noise than TRAD. However, although hard to see in the picture, that noise does not decrease significantly towards the tail either. Finally, the MAP estimation is clearly closer to the ground truth, and the estimation noise is decreasing with k. This is important, particularly in cases where the length of the true impulse response is not known a priori.

4.1. RIR estimates and AEC

While this paper focuses mostly on the RIR estimation itself, in this section we quickly comment on our underlying application, as it has some implications on our method. AEC is typically done by adaptive systems. This is partially due to the need to adapt to environment changes, and partially due to the need for a reasonable start up time. In systems where the direct path between loudspeaker and microphone is known a priori (e.g., speakerphones, etc) that path is

	Length = 4095			Length = 8191			Length = 16383		
SNR	TRAD	MLS	**MAP**	TRAD	MLS	**MAP**	TRAD	MLS	**MAP**
0	1.61	5.52	**9.08**	4.80	8.68	**11.19**	7.91	11.41	**13.41**
10	3.68	11.66	**14.34**	7.05	16.90	**18.19**	10.60	19.61	**21.47**
20	4.07	13.26	**15.49**	7.52	20.50	**22.78**	10.75	23.14	**28.08**
30	4.12	13.47	**15.78**	7.68	21.12	**23.26**	11.16	23.75	**30.55**

Table 1. Estimation mismatch, in dB, between the true RIR and the corresponding estimates using traditional correlation (TRAD), using correlation based on a MLS, and using the proposed method (MAP).

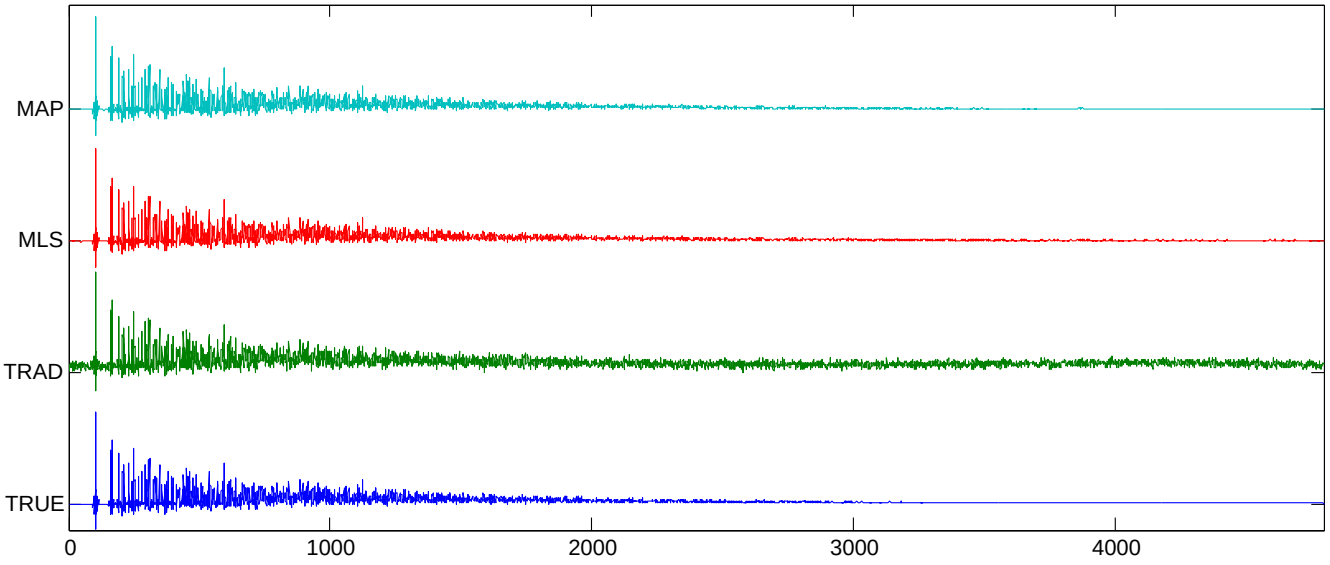

Fig. 2. RIR and corresponding estimations for a room with $T_{60} = 300ms$, and 20dB SNR. From bottom to top: RIR ground truth, traditional RIR estimate, MLS RIR estimate, and MAP RIR estimate.

typically measured in the factory (or during design), and used as the starting point for the adaptive filter. In systems where the location of loudspeaker is not controlled (e.g., PCs, game consoles, etc), this initial estimate has to be computed after deployment. Indeed, modern communication systems where the placement of the microphone in relation to the loudspeaker is not constrained, typically play a calibration tone to obtain an initial estimate of the AEC filter. Such calibration tone may be disguised (e.g., the sinusoidal sweep at the "startup" sound in Skype), or not (e.g., the Mozart snippet played during the initial Kinect calibration). Regardless of being disguised or not, "white noise" is not a desirable sound to be played at loud volumes. For that reason, as much as MLS provide improvements over pseudo-random sequences, they are very specific and cannot be modified. While the results reported for TRAD and MAP are also based on white noise sequences, these can be easily modified to accommodate virtually any sequence, by whitening the received (and reference) signals before computing the RIR estimates. While some degradation can be expected, the overall trend is similar. In other words, while we include results for using MLS in Table 1, we do not consider them as a valid alternative to many applications. Furthermore, we point out that, for cases where MLS are valid, the MAP results can be further improved by using them as the training sequence as well.

5. CONCLUSIONS

Estimating Room impulse responses find application in a number of areas. Estimation time, however, is a concern, as it can be too long for many applications. This time can be reduced by the use of MLS sequences, but these imply additional constraints on the sequences. In this paper we presented a new method for estimation of room impulse responses. The method is based on a MAP formulation, using the observed ambient noise and an exponential reverberation decay as priors. While priors have been used in the past in *adaptive* AEC, they are used to control the adaptation, and do not extend naturally to fixed (or initial) RIR estimation. Using our method, improvements of 11.8 dB over traditional methods, and of 2.8 dB over MLS sequences were achieved. In contrast to MLS, the proposed technique can be easily modified to support non-white training sequences, making it particularly appropriate for applications where users will be present, as for example in initial AEC filter estimation.

We are currently using this MAP estimation to speed up product development[36]. Additionally, we are also investigating extensions to multichannel adaptive AEC[7], implications to microphone arrays [37]and evaluating subjective quality of the results using CrowdMOS [38, 39]

6. REFERENCES

[1] M Junger and D Feit, *Sound, structures, and their interaction*, vol. 240, MIT press Cambridge, MA, 1972.

[2] M Schroeder, "Computer models for concert hall acoustics," *American Journal of Physics*, vol. 41, pp. 461, 1973.

[3] L Beranek and T Mellow, *Acoustics: Sound Fields and Transducers*, Academic Press, 2012.

[4] A Yellepeddi and D Florencio, "Sparse array-based room transfer function estimation for echo cancellation," *Signal Proc. Letters, IEEE*, vol. 21, no. 2, pp. 230–234, 2014.

[5] S Kuo and D Morgan, "Active noise control: a tutorial review," *Proceedings of the IEEE*, vol. 87/6, 1999.

[6] J Wung, T Wada, B-H Juang, B Lee, M Trott, and R Schafer, "A system approach to acoustic echo cancellation in robust hands-free teleconferencing," in *Proc. of WASPAA*, 2011.

[7] Z Zhang, Q Cai, and J Stokes, "Multichannel acoustic echo cancelation in multiparty spatial audio conferencing with constrained kalman filtering," in *Proc. of IWAENC*, 2008.

[8] M Song, C Zhang, D Florencio, and H Kang, "Personal 3D audio system with loudspeakers," in *Proc. of ICME*, 2010.

[9] Y Huang, J Chen, and J. Benesty, "Immersive audio schemes," *Signal Processing Magazine, IEEE*, vol. 28, no. 1, pp. 20–32, Jan 2011.

[10] J Pätynen, S Tervo, and T Lokki, "Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses," *The Journal of the Acoustical Society of America*, vol. 133, pp. 842, 2013.

[11] H Morgenstern and B Rafaely, "Analysis of acoustic mimo systems in enclosed sound fields," in *Proc. of ICASSP*, 2012.

[12] S Goetze, E Albertin, M Kallinger, A Mertins, and K Kammeyer, "Quality assessment for listening-room compensation algorithms," in *Proc. of ICASSP*, 2010.

[13] J Antons, S Arndt, R Schleicher, S Moller, D O'Shaughnessy, T Falk, et al., "Cognitive, affective, and experience correlates of speech quality perception in complex listening conditions," in *Proc. of ICASSP*, 2013.

[14] F Xiong, J Appell, and S Goetze, "System identification for listening-room compensation by means of acoustic echo cancellation and acoustic echo suppression filters," in *Proc. of ICASSP*, 2012.

[15] Y Rui, D Florencio, W Lam, and J Su, "Sound source localization for circular arrays of directional microphones," in *Proc. of ICASSP*, 2005.

[16] R Parisi, R Russo, M Scarpiniti, and A Uncini, "Localization of audio sources by multiple binaural sensors," in *Digital Signal Processing (DSP), 18th IEEE Int. Conf, on*, 2013.

[17] S Tervo, J Patynen, and T Lokki, "Acoustic reflection localization from room impulse responses," *Acta Acustica*, vol. 98, no. 3, pp. 418–440, 2012.

[18] S Tervo and T Tossavainen, "3D room geometry estimation from measured impulse responses," in *Proc. of ICASSP*, 2012.

[19] F Ribeiro, D Florencio, D Ba, and C Zhang, "Geometrically constrained room modeling with compact microphone arrays," *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. 20, no. 5, pp. 1449–1460, 2012.

[20] D Ba, F Ribeiro, C Zhang, and D Florencio, "L1 regularized room modeling with compact microphone arrays," in *Proc. of ICASSP*, 2010.

[21] F Ribeiro, D Florencio, P Chou, and Z Zhang, "Auditory augmented reality: Object sonification for the visually impaired," in *Proc. of MMSP*, 2012.

[22] J Klein, M Pollow, P Dietrich, and M Vorländer, "Room impulse response measurements with arbitrary source directivity," in *40th Italian (AIA) Annual Conference on Acoustics*, 2013.

[23] R. Mignot, L. Daudet, and F. Ollivier, "Room reverberation reconstruction: Interpolation of the early part using compressed sensing," *Audio, Speech, and Language Processing, IEEE Trans. on*, vol. PP, no. 99, pp. 1–1, 2013.

[24] C Contan, M Zeller, W Kellermann, and M Topa, "Excitation-dependent stepsize control of adaptive volterra filters for acoustic echo cancellation," in *Proc. of EUSIPCO*, 2012.

[25] E Habets, S Gannot, and I Cohen, "Robust early echo cancellation and late echo suppression in the stft domain," in *Proc. of IWAENC*, 2008.

[26] K Helwani, H Buchner, and S Spors, "Multichannel adaptive filtering with sparseness constraints," in *Acoustic Signal Enhancement; Int. Workshop on*, 2012.

[27] S Makino, Y Kaneda, and N Koizumi, "Exponentially weighted stepsize nlms adaptive filter based on the statistics of a room impulse response," *Speech and Audio Processing, IEEE Trans. on*, vol. 1, no. 1, pp. 101–108, 1993.

[28] M Rupp, "Convergence properties of adaptive equalizer algorithms," *Signal Processing, IEEE Trans. on*, vol. 59, no. 6, pp. 2562–2574, 2011.

[29] G Turin, "An introduction to matched filters," *Information Theory, IRE Trans. on*, vol. 6, no. 3, pp. 311–329, 1960.

[30] J Vanderkooy, "Aspects of mls measuring systems," *Journal of the Audio Engineering Society*, vol. 42, no. 4, pp. 219–231, 1994.

[31] M Vorländer and M Kob, "Practical aspects of mls measurements in building acoustics," *Applied Acoustics*, vol. 52, no. 3, pp. 239–258, 1997.

[32] I Mateljan, "Signal selection for the room acoustics measurement," in *Proc. of WASPAA*, 1999.

[33] E Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. of ICASSP*, 2005.

[34] K Lebart, J Boucher, and P Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[35] S Schimmel, M Muller, and N Dillier, "A fast and accurate shoebox room acoustics simulator," in *Proc. of ICASSP*, 2009.

[36] P Moquin, K Venalainen, and D Florencio, "Determination of room impulse response for synthetic data acquisition and ASR testing," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 2265, 2014.

[37] D A Florencio and H Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," in *Proc. of ICASSP*, 2001.

[38] F Ribeiro, D Florencio, C Zhang, and M Seltzer, "CrowdMOS: An approach for crowdsourcing mean opinion score studies," in *Proc. of ICASSP*, 2011.

[39] F Ribeiro, D Florencio, and V Nascimento, "Crowdsourcing subjective image quality evaluation," in *Proc. of ICIP*, 2011.