

Sparse Array-Based Room Transfer Function Estimation for Echo Cancellation

Atulya Yellepeddi, *Student Member, IEEE* and Dinei Florencio, *Senior Member, IEEE*

Abstract

A number of applications in acoustics, such as echo cancellation, require learning the acoustic impulse response from each deployed loudspeaker to each microphone- the room transfer function. This has conventionally been done separately at each microphone for each loudspeaker. However, the signals arriving at the array share a common structure, which can be exploited to improve the impulse response estimates. In this work, we propose an algorithm that takes advantage of the array structure, as well as the sparsity of the reflections arriving at the array in order to form reliable estimates of the impulse response between each loudspeaker and microphone. The algorithm is shown to improve performance over the matched filter algorithm in echo cancellation applications, using both synthetic and real data.

Index Terms

Echo cancellers, Array signal processing, Sparse model

I. INTRODUCTION

Room transfer function (RTF) estimation, which is the problem of estimating the impulse response from a loudspeaker to a microphone through a reverberant room, has been an important problem in acoustic signal processing [1]. It is relevant in many applications, including in sound source localization [2], [3], sound field reproduction [4] and perception of 3-d audio [5]. In this work, we consider the application of initializing acoustic echo cancellation (AEC) algorithms.

Copyright ©2013 IEEE. Personal use of this material is permitted. Permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org

Atulya Yellepeddi was with Microsoft Research during this work. He is in the MIT-WHOI Joint Program and may be reached at 77 Massachusetts Avenue Room 36-699, Cambridge, MA 02139. Email: atulya@mit.edu

Dinei Florencio is with Microsoft Research. Email: dinei@microsoft.com

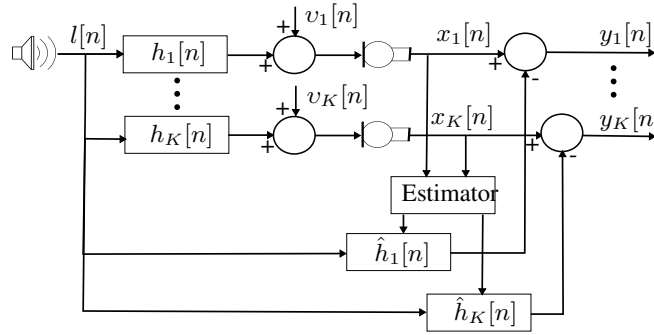


Fig. 1. System model for one loudspeaker

In AEC, filters are generally initialized with an estimate of the RTF, which is computed based on a calibration tone. These initial estimates are often computed using a simple matched filter algorithm, and then adapted using one of many techniques [6]–[9]. Modern applications of AEC like voice control for home- and car-entertainment systems [10]–[12] generally have systems with multichannel audio reproduction and microphone arrays, with unknown loudspeaker geometry and significant noise levels. This has brought significant research attention to the problem of exploiting side information for echo cancellation (e.g. [13]–[15]). In particular, the problem of array processing with AEC is dealt with in [16]–[18] among others. Generally, these involve jointly adapting a number of channels. The number of taps to adapt can be large if there are multiple loudspeakers, making this possibly computationally complex.

In this paper we consider using the (known) microphone array geometry to improve the RTF estimate at each microphone, and this is used to initialize the echo cancellation filters (as mentioned, it could be used for other purposes also). Since the algorithm computes an initial estimate, it is only run once, following which adaptation may be performed on each individual channel independently, with the knowledge that the estimates already include the array information. While not optimal, we expect that this approach is better in practice than treating each channel independently.

For the purposes of this paper, adaptation is not considered. We compare the initial estimates using the matched filter and using the proposed algorithm with respect to their performance in initial echo cancellation, with the understanding that low-complexity adaptation is to be performed afterward, which can only improve performance.

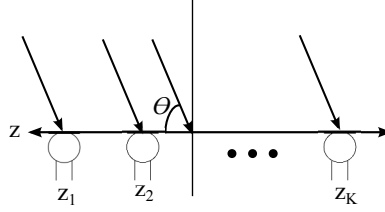


Fig. 2. Signals arriving at a 1-d Array

II. SYSTEM MODEL

Fig.1 represents the system model. A signal $l[n]$ is played at the loudspeaker¹. The RTF from the loudspeaker to the k th microphone is denoted as $h_k[n]$. Thus, the signal received at microphone k is:

$$x_k[n] = \sum_{m=0}^{N-1} h_k[m]l[n-m] + d_k[n] + v_k[n], \quad (1)$$

where $d_k[n]$ is the desired signal and $v_k[n]$ is the interfering noise. Note that $d_k[n]$ is assumed to be zero during the calibration phase.

When only one microphone and one loudspeaker are present, the minimum mean square error estimate of the RTF is known as the “matched filter” [19], which can be computed as:

$$h_k^{\text{MF}}[n] = \frac{\sum_m \tilde{l}[m] \tilde{x}_k[n+m]}{\sum_m (\tilde{l}[m])^2} \quad (2)$$

where the $\tilde{l}[n]$ refers to the whitened version (see [20]) of $l[n]$, and $\tilde{x}_k[n]$ refers to $x_k[n]$ after it is filtered by the same filter used to whiten $l[n]$.

The filter of (2) is the baseline initial estimate against which our algorithm is compared. It is shown that the estimate of each RTF obtained by exploiting the multiple microphones can be better than the above optimum.

A. Structure of the Received Signal

To understand the rationale behind our algorithm, we investigate the structure of the received signal. In the following, continuous-time signals are represented using parentheses. For example, $l(t)$ is the continuous time version (i.e., the continuous time interpolation) of the loudspeaker signal $l[n]$. Conversely, $l[n]$ can be seen as the sampled version of $l(t)$, i.e., $l[n] = l(nT_s)$, where T_s is the sampling interval.

¹For simplicity, only one loudspeaker is considered to compute the RTF, though more than one may be deployed in reality. This is compatible with the calibration phase, where each speaker is played separately.

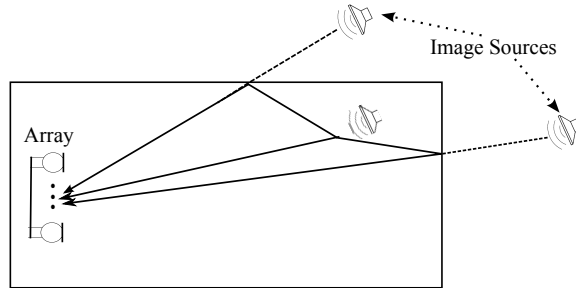


Fig. 3. Modeling reflections in a room

Consider a plane wave impinging upon the linear array in Fig.2. If the continuous time signal at the origin is $s(t)$, then the signal at microphone k is $s(t - \tau(z_k, \theta))$, where $\tau(z_k, \theta) = -z_k \cos \theta / c$, and c is the speed of sound. In other words, the signal produced at each microphone by the plane wave is the same, except for a delay which depends only on the array geometry and the direction of arrival.

Now consider a room as in Fig.3. The loudspeaker signal $l(t)$ is reflected off a number of surfaces or objects and all of these reflections are received at the microphone array. Each reflection corresponds to a “image source” [21] as shown. Now, further assume that: 1) the dimensions of the array are small enough and the sources are far enough away from the array that the wavefronts can be approximated as planar (farfield approximation); 2) every reflector is flat-fading², and; 3) P sources (real or image) are enough to provide a good approximation of the arriving signal.

Under these assumptions, the signal received at sensor k can be written as:

$$x_k[n] = \sum_{p=1}^P \alpha_p l(nT_s - T_p - \tau(z_k, \theta_p)) + \tilde{v}_k[n] \quad (3)$$

where T_p and α_p are, respectively, the delay and attenuation suffered by wavefront p (including propagation loss and microphone and loudspeaker directivities), and $\tilde{v}_k[n]$ includes both the actual noise $v_k[n]$ and possibly any unmodeled signal components. The corresponding channel is:

$$h_k[n] = \sum_{p=1}^P \alpha_p \text{sinc} \left(n - \frac{T_p + \tau(z_k, \theta_p)}{T_s} \right) \quad (4)$$

where the function $\text{sinc}(x) = \sin(\pi x) / (\pi x)$ arises as a result of sampling an impulse. In other words, the RTF of each microphone can be obtained by writing the received signal as a combination of planar wave copies of the loudspeaker signal. This is the principle that we use to develop our algorithm.

²See, however, Section III-B

III. THE PROPOSED ALGORITHM

Using the model of the previous subsection, we develop an algorithm that takes advantage of the array information to compute the filters $\hat{h}_k[n]$. Let L_x denote the length of the received calibration signal at each sensor, i.e, the length of $x_k[n]$ during the calibration period. Define the received signal vector \mathbf{x} of dimension $KL_x \times 1$ as:

$$\mathbf{x} = [x_1[0], \dots, x_1[L_x - 1], \dots, x_K[0], \dots, x_K[L_x - 1]]^T \quad (5)$$

That is, stack up the received signals during the calibration period at all the sensors into one tall vector. Similarly, define a basis vector $\mathbf{b}_{T,\theta}$, also of dimension $KL_x \times 1$, as:

$$\mathbf{b}_{T,\theta} = \begin{bmatrix} l(0 - T - \tau(z_1, \theta)) \\ \vdots \\ l((L_x - 1)T_s - T - \tau(z_1, \theta)) \\ \vdots \\ l(0 - T - \tau(z_K, \theta)) \\ \vdots \\ l((L_x - 1)T_s - T - \tau(z_K, \theta)) \end{bmatrix}. \quad (6)$$

In other words, $\mathbf{b}_{T,\theta}$ represents the set of signals that would be received by the K array elements from an (unattenuated) source, either real or image, located at an angle θ from the array axis, and at a distance such that the propagation delay from the source to the center of the array is T .

Equation (3) implies that the vector \mathbf{x} of (5) can be written as a linear combination of the basis vectors (6), except for the noise and unmodeled components $\tilde{v}[n]$. Thus, the problem posed is recovering the basis vectors and corresponding weights which will result in the best representation of the observed vector \mathbf{x} . We begin by discretizing the space of θ and T , and forming a basis³ \mathbf{B} for the desired echo space $\mathbf{B} = [\mathbf{b}_{T_1,\theta_1}, \mathbf{b}_{T_2,\theta_1}, \dots]$.

It may appear as though we have made the problem more complex by adding nearly every possible basis vector to the optimization (as opposed to, for example, trying to somehow detect the useful basis vectors). However, observe that while the number of basis functions is fairly large, the number of reflections we actually expect to receive is relatively small, due to the sparseness of the RTF (which has been used in

³Rigorously, sampling the T space finely enough does produce a basis, but sampling θ does not.

literature- for example, to interpolate RTF between two source locations [22], to estimate room geometry [23] and even in echo cancellation [24]). Thus, we impose a sparsity constraint on the solution. Basis Pursuit Denoise (BPDN) with a convex relaxation is a well-known method to solve the problem of approximate representation with sparse constraints [25]. This method solves the following optimization problem:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \|\mathbf{w}\|_1, \quad \text{s.t.} \quad \|\mathbf{B}\mathbf{w} - \mathbf{x}\|_2 \leq \sigma. \quad (7)$$

where σ is a reconstruction error, which should ideally be slightly larger than the standard deviation of $\tilde{v}[n]$. That would ensure a reconstruction error close to the modeling error plus noise. In practice, we have access only to the noise $v[n]$ since we do not know the modeling error. So, we set $\sigma = \beta\sigma_v$ for some $\beta > 1$, where σ_v is the noise standard deviation.

It still remains to be said how to estimate the channel from the weights. With N denoting the channel length, define the $KN \times 1$ vector:

$$\mathbf{h}_{T,\theta} = \begin{bmatrix} \text{sinc}\left(0 - \frac{T+\tau(z_1,\theta)}{T_s}\right) \\ \vdots \\ \text{sinc}\left(N-1 - \frac{T+\tau(z_1,\theta)}{T_s}\right) \\ \vdots \\ \text{sinc}\left(0 - \frac{T+\tau(z_K,\theta)}{T_s}\right) \\ \vdots \\ \text{sinc}\left(N-1 - \frac{T+\tau(z_K,\theta)}{T_s}\right) \end{bmatrix} \quad (8)$$

Note that $\mathbf{h}_{T,\theta}$ is the equivalent of $\mathbf{b}_{T,\theta}$ for the impulse response, i.e., it is a basis vector for the impulse response corresponding to the basis vector $\mathbf{b}_{T,\theta}$ of (4). Let the matrix $\mathbf{H} = [\mathbf{h}_{T_1,\theta_1}, \mathbf{h}_{T_2,\theta_1}, \dots]$ be the corresponding basis matrix. Then the vector $\mathbf{H}\hat{\mathbf{w}}$ contains the K impulse response estimates stacked up in order. These impulse responses are the final estimates produced by our sparse, array-based method.

A. Practical Considerations: Using FFTs for Computation

The matrix \mathbf{B} which is needed to solve (7) is very large—the number of rows is K times the length of the received calibration signal and the number of columns is the number of basis functions. For a typical set of parameters, the matrix has dimension 70000×110000 . This may be too large to represent in memory and for computation.

However, in order to solve (7) using a solver such as the one described in [26], we need only implement $\mathbf{B}\mathbf{u}$ and $\mathbf{B}^T\mathbf{v}$ for arbitrary vectors \mathbf{u} and \mathbf{v} of appropriate dimension. Now, the first element of the product $\mathbf{B}\mathbf{u}$, denoted $(\mathbf{B}\mathbf{u})(0)$ is given by $(\mathbf{B}\mathbf{u})(0) = \sum_{\theta} \sum_T u_{T,\theta} \mathbf{b}_{T,\theta}(0)$, where $u_{T,\theta}$ is the component of \mathbf{u} which multiplies $\mathbf{b}_{T,\theta}$ in the matrix product. Then, $(\mathbf{B}\mathbf{u})(0) = \sum_{\theta} \sum_T u_{T,\theta} l(0 - T - \tau(z_1, \theta))$. Note that the inner sum is the first element of a convolution, where \mathbf{u} and $l(\cdot)$ are treated as series in T .

Following this line of reasoning, we can compute all the elements of the inner sum over T by computing a single convolution. It is thus possible to efficiently implement the matrix multiplication operation $\mathbf{B}\mathbf{u}$ with one FFT per θ (we can pre-compute the FFT of the basis function, so only one is needed). A similar formula can be derived for $\mathbf{B}^T\mathbf{v}$ which also needs only one FFT per θ .

B. Reflectors with Frequency-Dependent Fading

In general, the attenuation coefficient of a reflector is frequency dependent. Nonetheless, the modeling error as a result of the flat fading assumption is expected to be relatively small. To understand this, note that a reflection that arrives with frequency distortion is, in the time-domain, equivalent to the signal passed through some filter. The impulse response of that filter may be decomposed into the sum of a number of impulses. Hence, the received signal can still be represented by the sum of the basis functions $\mathbf{b}_{T,\theta}$ in our framework.

Physically speaking, this is equivalent to modeling each single reflection as a group of plane waves arriving at the array, rather than a single plane wave. In the basis framework, a single reflection is a collection of multiple basis functions rather than a single basis function. While this makes the solution less sparse, in practice, the matrix \mathbf{B} over-represents the received signal so much that sparsity is still a meaningful constraint.

IV. RESULTS

In this section, we compare the performances of the sparse array-based echo canceller and of the matched filter, with both synthetic and real data. The measure of performance is the average echo return loss enhancement (ERLE) over all the microphones, estimated using:

$$\text{ERLE} = 10 \log_{10} \left(\frac{\frac{1}{KL_x} \sum_{k=1}^K \sum_{n=1}^{L_x} |x_k[n]|^2 - \sigma_v^2}{\frac{1}{KL_x} \sum_{k=1}^K \sum_{n=1}^{L_x} |y_k[n]|^2 - \sigma_v^2} \right) \quad (9)$$

where σ_v^2 is the variance of the noise. For synthetic data the noise is known, and we subtract it out from $x_k[n]$ and $y_k[n]$ before computing the ERLE. For actual data, a noise-only segment (i.e., in which the loudspeakers are not played) is used to compute σ_v^2 .

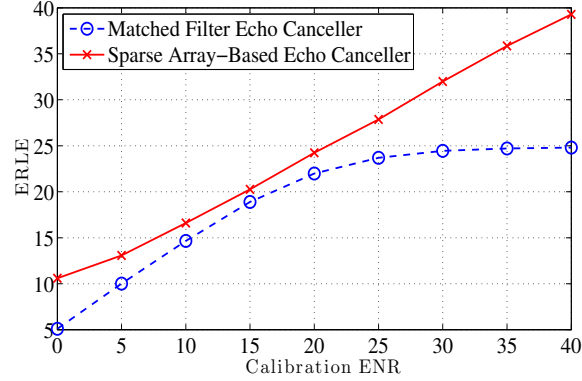


Fig. 4. Results with Synthetic Data

The ERLE is computed at a number of received Echo-to-Noise Ratios (ENR). The ENR represents the total power of the received echo signal during the *calibration* period (referenced to the power of interfering signals present with the loudspeakers silent). The filters computed during the calibration period are then used to cancel the loudspeaker signals during a *testing* period (in which all loudspeakers are simultaneously played) to compute the ERLE. For both synthetic and real data, the length of signals in the calibration period are 1 second per loudspeaker, and the testing period is 15 seconds long. The sampling frequency is 16kHz.

We emphasize that the results are comparing two different initial estimates using the test signal, so they are not comparable to those obtained with adaptive filters. Moreover, the testing periods are relatively short compared to conventionally chosen times, which may also affect the results.

A. Synthetic Data

To generate the synthetic signals, a rectangular room with attenuating but flat-fading walls was assumed. A single loudspeaker (assumed to be a point source) whose signal was to be canceled was placed at an arbitrarily chosen location in the room. Interference was generated by another point source placed at a different arbitrarily chosen location. The output of the interfering source was white Gaussian noise whose level was scaled to the required ENR.

The reverberation time of the simulated room was 50ms, for a filter length of $N = 800$ samples. A 4-element linear microphone array was simulated. The vectors $\mathbf{b}_{\theta,T}$ were computed for θ and T spaced at 5° and 1 sample time, respectively.

The ERLE with the synthetic data is shown in Fig.4. The results demonstrate that the ERLE is markedly

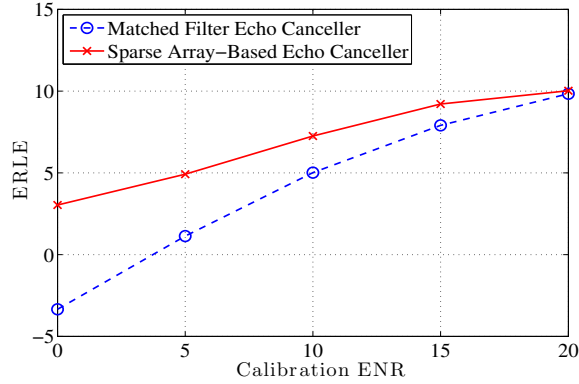


Fig. 5. Results with Experimental Data

higher when the filters are computed with the array-based method than with the matched filter. A few points are worth mentioning:

- 1) When the noise is small (high ENR), the performance of the matched filter saturates. This happens because the autocorrelation of a (finite) white signal is not a perfect impulse, and this residual autocorrelation acts as noise. The array-based method is not subject to this effect and its performance keeps improving.
- 2) For low ENR, the array-based method does a better job of rejecting the noise. In this case, the limiting factor with the matched filter is the cross-correlation between the signal $l[n]$ and the noise.
- 3) The least improvement seems to occur in the regime where the noise and the residual cross-correlation are at similar levels. Although the reasons for this effect are not clear, we conjecture the algorithm used to solve (7) uses the provided slack σ to reduce the coefficient estimates rather than allocating it to reject the noise.

B. Experimental Data

The experimental data was collected in a room with a 5.1 loudspeaker system and a Microsoft Kinect device, which has a 4-microphone array [27]. The calibration tones were the ones used on the original Kinect. The sound track of a game was played at the loudspeakers during the testing period. The ENR of the collected signals was about 20dB, and lower ENRs were achieved by corrupting these signals with the collected noise.

The results with an assumed impulse response length $N = 1600$ samples are shown in Fig.5. Results for both longer and shorter lengths were collected, but they follow the same trends and are thus omitted

here. The figure shows that, again, there is an advantage to using the array-based method to compute the filters. The performance does not improve as much as for the case of synthetic data, but this is to be expected as the synthetic data exactly fits the assumed model. Nonetheless, the array-based echo canceller outperforms the matched filter. At very low ENRs, the matched filter actually adds echo ($ERLE < 0\text{dB}$). While the matched filter would not be operated in such a regime, it is encouraging that the array-based method still cancels some of the echo.

V. CONCLUSIONS

In this paper, we proposed an algorithm that used array information to compute the RTFs with which to initialize AEC filters, using a generic reflection model to exploit the structure in the signals received at the array. We pose the problem of computing RTFs as a least squares optimization with a sparsity constraint. The results clearly indicate the value of utilizing array information in the computation of the initial filters.

A number of questions remain. The premise of the paper was to incorporate the array information into the initial estimate and perform adaptation on each filter individually. It would be interesting to compare this to jointly adapting all the channels. It may be possible to simplify the adaptation by utilizing the fact that the initial estimates are highly structured. Finally, other means to exploit structure may be relevant. For instance, it may be possible to localize the reflectors in the room and utilize this information to further improve performance. Indeed, further exploiting the signal structure induced by the array geometry is a fertile area of research.

REFERENCES

- [1] T. Gustafsson, J. Vance, H. Pota, B. Rao, and M. Trivedi, "Estimation of acoustical room transfer functions," in *Decision and Control, 2000. Proceedings of the 39th IEEE Conference on*, vol. 5. IEEE, 2000, pp. 5184–5189.
- [2] F. Ribeiro, C. Zhang, D. A. Florêncio, and D. E. Ba, "Using reverberation to improve range and elevation discrimination for small array sound source localization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1781–1792, 2010.
- [3] I. Dokmanic and M. Vetterli, "Room helps: acoustic localization with finite elements," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2617–2620.
- [4] T. Betlehem and T. D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, 2005.
- [5] M.-S. Song, C. Zhang, D. Florencio, and H.-G. Kang, "An interactive 3-d audio system with loudspeakers," *Multimedia, IEEE Transactions on*, vol. 13, no. 5, pp. 844–855, 2011.
- [6] 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 Transactions on*, vol. 1, no. 1, pp. 101–108, 1993.

- [7] J. Benesty and S. L. Gay, "An improved PNLMS algorithm," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2. IEEE, 2002, pp. II-1881.
- [8] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic echo cancellation filters—an overview," *Signal Processing*, vol. 80, no. 9, pp. 1697–1719, 2000.
- [9] D. L. Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 5, pp. 508–518, 2000.
- [10] A. S. Chhetri, J. W. Stokes, and D. A. Florêncio, "Acoustic echo cancellation for high noise environments," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 905–908.
- [11] H. Buchner and W. Kellermann, "Acoustic echo cancellation for two and more reproduction channels," in *Conf. Rec. IEEE IWAENC*, 2001, pp. 99–102.
- [12] A. Moonen, R. Van Den Berg, M. Bekooij, H. Bhullar, and J. Van Meerbergen, "A multi-core architecture for in-car digital entertainment," in *Proc. of GSPx Conference*, 2005.
- [13] D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Active acoustic echo cancellation in spatial soundfield reproduction," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 620–624.
- [14] H. Buchner, S. Spors, and W. Kellermann, "Wave-domain adaptive filtering: Acoustic echo cancellation for full-duplex systems based on wave-field synthesis," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 4. IEEE, 2004, pp. 117–120.
- [15] B. Debail and A. Gilloire, "Microphone array design with improved acoustic echo rejection," in *Acoustic Echo and Noise Control, IEEE International Workshop on*, 2001, pp. 55–58.
- [16] J. Benesty, P. Duhamel, and Y. Grenier, "A multichannel affine projection algorithm with applications to multichannel acoustic echo cancellation."
- [17] W. L. Kellermann, "Acoustic echo cancellation for beamforming microphone arrays," in *Microphone Arrays*. Springer, 2001, pp. 281–306.
- [18] M. Dahl and I. Claesson, "Acoustic noise and echo cancelling with microphone array," *Vehicular Technology, IEEE Transactions on*, vol. 48, no. 5, pp. 1518–1526, 1999.
- [19] G. Turin, "An introduction to matched filters," *Information Theory, IRE Transactions on*, vol. 6, no. 3, pp. 311–329, 1960.
- [20] M. H. Hayes, *Statistical digital signal processing and modeling*. Wiley, 2009.
- [21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, p. 943, 1979.
- [22] R. Mignot, L. Daudet, and F. Ollivier, "Room reverberation reconstruction: Interpolation of the early part using compressed sensing," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2013.
- [23] F. Ribeiro, D. Florêncio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1449–1460, 2012.
- [24] J. Benesty, Y. A. Huang, J. Chen, and P. A. Naylor, "Adaptive algorithms for the identification of sparse impulse responses," *Selected methods for acoustic echo and noise control*, vol. 5, pp. 125–153, 2006.
- [25] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *Information Theory, IEEE Transactions on*, vol. 52, no. 1, pp. 6–18, 2006.
- [26] E. Van Den Berg and M. P. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

- [27] S. Selfon, ““XBox, Listen”: Driving gameplay with Kinect audio input and speech recognition,” Presented at Game Developers Conference, 2011, Available: <http://www.microsoft.com/en-us/download/details.aspx?id=26102>.