A DEREVERBERATION ALGORITHM FOR SPHERICAL MICROPHONE ARRAYS USING COMPRESSED SENSING TECHNIQUES

Ping Kun Tony Wu, Nicolas Epain, Craig Jin

Computing and Audio Research Laboratory (CARLab)
School of Electrical and Information Engineering
The University of Sydney, NSW 2006, Australia

ABSTRACT

In this paper, we present a novel multichannel dereverberation algorithm that enhances a target signal in a reverberant environment. The proposed algorithm is designed for a spherical microphone array and formulated in the spherical harmonic domain. The algorithm employs sparse recovery, a compressed sensing technique, to estimate the position of the target signal and its early reflections. Room impulse responses are obtained according to the estimations and the MINT (the multiple-input/output inverse-filtering theorem) is used to calculate the inverse filters. The performance of the proposed method is evaluated using computer simulation and our results indicate the effectiveness of the proposed dereverberation algorithm.

Index Terms— Compressed Sensing, Dereverberation, Spherical microphone arrays

1. INTRODUCTION

Reverberation poses a major challenge to acoustic signal processing problems. It degrades speech quality and speech intelligibility, especially critical for non-native listeners and hearing impairment listeners, and also causes many acoustic algorithms to perform poorly. Speech dereverberation is an acoustic signal processing technique that aims to extract the original target signal from the reverberant microphone signal(s) in order to improve the quality and intelligibility of speech for various applications. Generally speaking, speech dereverberation algorithms are commonly classified into one of the following three categories: (1) beamforming-based approach: The observed signals received at microphones are weighted and summed, so as to form a beam in the direction of the target signal and attenuate other signals such as reverberation and noise from other directions; (2) model-based approach: The microphone signals are modified so as to better represent some features of the clean target signal according to an *a priori* model of the speech waveform or spectrum; and (3) an inverse filtering approach: The room impulse responses (RIRs) are estimated blindly using the microphone signals and then used to design inverse filters that compensates for the effect of the reverberation.

In this paper, we present a speech dereverberation algorithm that is based on the application of sparse recovery, a compressed sensing (CS) technique. Compressed sensing is a novel sensing paradigm that can be employed to find the sparse inverse solutions for an under-determined system. More information about the CS technique can be found in [1]. We show that the positions of the target signal and some of its early reflections can be estimated using the CS technique and the target speech signal can be enhanced with the aid of these estimations. The proposed algorithm is designed for a spherical microphone array and is formulated in the spherical harmonic domain. Spherical microphone arrays provide a promising tool for the spatial analysis of complex sound fields. As well, working in the spherical harmonic domain has several advantages including scalability and the ability to rotate the sound scene by a simple matrix operation. It is also shown in [2] that a convolutive mixture in time domain can be transformed into a instantaneous mixture in the spherical harmonic domain, which provides significant advantages for source localization and separation [3]. We present a simulation experiment in order to demonstrate the effectiveness of the proposed algorithm and we also compare the performance of the proposed algorithm with other algorithms.

2. METHOD

Consider a general multi-microphone system with L microphone signals which we model as:

$$x_i(t) = s(t) \otimes g_i(t),\ i = 1, 2, \ldots, L\,, \tag{1}$$

where $x_i(t)$ is the signal at the i-th microphone, $s(t)$ is the target signal, $g_i(t)$ is the impulse response describing the room transfer function from the target signal location to the i-th microphone and $\otimes$ represents the convolution operation. Spherical harmonic analysis, as used in higher order ambisonics (HOA) [4], is a powerful tool for describing the spatial properties of sound fields. The spherical harmonic expansion of a sound field can be obtained from a spherical microphone array [2, 5]. The spherical harmonic expansion, $\mathbf{B}$, of a sound

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

field corresponding to a set of plane waves can be expressed as a simple matrix product:

$$\mathbf{B} = \mathbf{YS}\,, \tag{2}$$

where, when using a time window of length K, we have:

$$\mathbf{S} = \begin{bmatrix} s_1(t) & s_1(t+1) & \dots & s_1(t+K-1) \\ s_2(t) & s_2(t+1) & \dots & s_2(t+K-1) \\ \vdots & \vdots & \vdots & \vdots \\ s_P(t) & s_P(t+1) & \dots & s_P(t+K-1) \end{bmatrix} ,$$
$$\mathbf{B} = [\mathbf{b}(t), \mathbf{b}(t+1), \dots, \mathbf{b}(t+K-1)] \ ,$$
$$\mathbf{b}(t) = \left[B_0^0(t), B_1^{-1}(t), \dots, B_m^n(t)\right]^T \ ,$$
$$\mathbf{Y} = [\mathbf{y}(\theta_1, \phi_1), \mathbf{y}(\theta_2, \phi_2), \dots, \mathbf{y}(\theta_P, \phi_P)] \ ,$$
$$\mathbf{y}(\theta_p, \phi_p) = \left[Y_0^0(\theta_p, \phi_p), Y_1^{-1}(\theta_p, \phi_p), \dots, Y_m^n(\theta_p, \phi_p)\right]^T \ ,$$
$$m \in [0, 1, \dots, M]\,, \qquad n \in [-m, \dots, m]\,.$$

$\mathbf{Y}$ is a $(M+1)^2 \times P$ spherical harmonic matrix, truncated to order M, with column p providing the spherical harmonic expansion for a plane-wave source located in the direction (θ_p, ϕ_p), P is the number of entries in the dictionary of possible plane-wave source directions and is typically chosen much larger than $(M+1)^2$, $\mathbf{S}$ is a $P \times K$ matrix of plane-wave signals (the p-th row of $\mathbf{S}$ is non-zero if there is a signal in the direction (θ_p, ϕ_p)) and $\mathbf{B}$ is a $(M+1)^2 \times K$ matrix containing the HOA signals. In this work, we have chosen $P = 642$, $M = 2$, $K = 2048$. The HOA signals, $\mathbf{B}$, are band-pass filtered so that they contain only the frequencies where the encoding is accurate and can be considered as instantaneous mixtures. The frequency band for accurate HOA encoding is limited by measurement noise that is amplified by the encoding filters at low frequencies and spatial aliasing that pollutes the HOA signals at high frequencies. The frequency range for the band pass filter is 300 to 3500 Hz and the spherical microphone we use consists of two concentric arrays of 12 omnidirectional microphones. There are 12 microphones located on the surface of a rigid sphere with a radius of 3 cm; the other 12 microphones are located on the surface of a open sphere with a radius of 15 cm. Please refer to [2] for a detailed description of the processing.

Equation (2) is an under-determined system. In general, there is an infinite number of solutions and the inverse problem is ill-posed. Our approach to this ill-posed problem is to impose sparsity on the solution $\mathbf{S}$, so that the resulting sound field is explained by a minimum number of plane wave sources. Mathematically we formulate the sparse recovery problem as:

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{S}\|_{1,2} \\ &\text{subject to} \quad \mathbf{B} = \mathbf{YS}\,, \end{aligned} \tag{3}$$

where $\|\mathbf{S}\|_{1,2}$ is the $l_{1,2}$ norm of $\mathbf{S}$ and is defined as:

$$\|\mathbf{S}\|_{1,2} = \sum_{p=1}^{P} \sqrt{\sum_{k=1}^{K} s_p(t+k-1)^2}\,.$$

The computational cost of this optimization problem is high because of the large size of the inverse problem. In order to reduce the computational complexity and the sensitivity to noise, we use a SVD (Singular Value Decomposition) method [6]. The idea is to decompose $\mathbf{B}$ into the signal and noise subspaces and discard the noise subspace. This can be represented mathematically as:

$$\mathbf{B} = \mathbf{ULV}^T\,. \tag{4}$$

As only the first $(M+1)^2$ singular values are non-zero, we express $\mathbf{B}$ as:

$$\mathbf{B} = \mathbf{U\Lambda\Psi}^T\,, \tag{5}$$

where $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are obtained by keeping the first $(M+1)^2$ columns of $\mathbf{L}$ and $\mathbf{V}$, respectively. We now express $\mathbf{S}$ in the subspace defined by $\mathbf{\Psi}$, *i.e.* we define $\mathbf{\Omega}$ such that:

$$\mathbf{S} = \mathbf{\Omega\Psi}^T\,. \tag{6}$$

Using (5) and (6), the equality constraint in (3) becomes:

$$\mathbf{Y\Omega} = \mathbf{U\Lambda}\,, \tag{7}$$

yielding the following convex optimization problem:

$$\begin{aligned} &\text{minimize} \quad ||\mathbf{\Omega}||_{1,2} \\ &\text{subject to} \quad \mathbf{Y\Omega} = \mathbf{U\Lambda}\,. \end{aligned} \tag{8}$$

CVX, a Matlab package for specifying and solving convex optimization problem [7], is used for solving the optimization problem (8).

We make the hypothesis that, for every K-long time window, the directions of the target signal and its first J reflections correspond to the $J+1$ most energetic rows in $\mathbf{\Omega}$. We track these directions across the entire signal and choose those indices that occur most frequently over time. Note that in this work we have set $J = 3$.

Once the directions of the target signal and its reflections are obtained, an un-mixing matrix $\mathbf{Y}'^{-1}$ is estimated as:

$$\mathbf{Y}'^{-1} = \text{pinv}(\mathbf{Y}')\,, \tag{9}$$

where

$$\mathbf{Y}' = [\mathbf{y}(\theta_{p_0}, \phi_{p_0}), \mathbf{y}(\theta_{p_1}, \phi_{p_1}), \dots \mathbf{y}(\theta_{p_J}, \phi_{p_J})] \ ,$$

$(\theta_{p_j}, \phi_{p_j})$ is the j-th identified direction and pinv($\mathbf{X}$) represents the pseudo-inverse of $\mathbf{X}$. We then apply the un-mixing matrix to the band-pass filtered HOA signals to estimate the target signal and its reflections. In order to identify the target signal, $\hat{\mathrm{s}}_{\mathrm{tar}}(t)$, we first choose the signal with the most energy. We then compute the time delays, q_j, between this signal and the other signals, referred to as $\hat{\mathrm{s}}_{\mathrm{ref},j}(t)$, using a cross-correlation method:

$$q_j = \underset{x}{\operatorname{argmax}} \left(\frac{1}{L} \sum_{t=0}^{L-x-1} \hat{\mathrm{s}}_{\mathrm{tar}}(t+x)\, \hat{\mathrm{s}}_{\mathrm{ref},j}(t) \right) , \qquad (10)$$

If we come across a signal with negative time delay, we then select this signal as the new target signal and continue. The attenuation parameter for each reflection is determined as the ratio of the norms of the extracted reflection and target signals:

$$\alpha_j = \frac{\|\hat{\mathrm{s}}_{\mathrm{ref},j}(t)\|}{\|\hat{\mathrm{s}}_{\mathrm{tar}}(t)\|} . \qquad (11)$$

Given the directions of the target signal and its reflections, and also the relative amplitude and delays of these signals, we estimate the RIRs for each microphone in the array. We then apply the MINT [8] technique to these RIRs in order to calculate multichannel inverse filters that can be used to estimate the clean target signal.

3. EXPERIMENT

We evaluated the proposed dereverberation algorithm with a spherical microphone array in three different reverberant rooms using computer simulation. All three rooms have the same size of 14 × 10 × 3 m (W × L × H), but with different reverberant properties determined by the absorption coefficients. The target source and spherical microphone array are located at (5, 4, 1) m and (4, 4, 1) m respectively relative to the corner of the room. The experimental setup is shown in Fig. 1 and the reverberation time versus frequency for each room are shown in Fig. 2. The RIRs were obtained using MCRoomSim, a multichannel room acoustics simulator that is suitable for a spherical microphone array simulation [9]. These RIRs were used to filter male voice recordings to create test signals that simulate spherical microphone array recordings in a reverberant environment. These signals were approximately 4 seconds in duration with all of the audio processed at a sampling rate of 16 kHz.

We compared the proposed algorithm with the ICA-based approach [2] and the MUSIC (MUltiple SIgnal Classification) algorithm [10] combined with the MINT technique. The performance of these algorithms were evaluated using the Perceptual Evaluation of Speech Quality (PESQ) measure as described in the ITU-T recommendation P.862 [11] and the Segmental Signal-to-Reverberation Ratio (SegSRR) [12]. The output of the PESQ is a measure of the subjective assessment quality of the degraded signal and is rated as a value between 0 (unacceptable) and 4.5 (excellent). The SegSRR is a signal-based measure which is obtained by segmenting a signal into N_{seg} smaller time frames (typically a duration of 20-40 ms), calculating the normalized SRR (NSRR) for each frame and averaging all of these NSRR values. This can be expressed as:

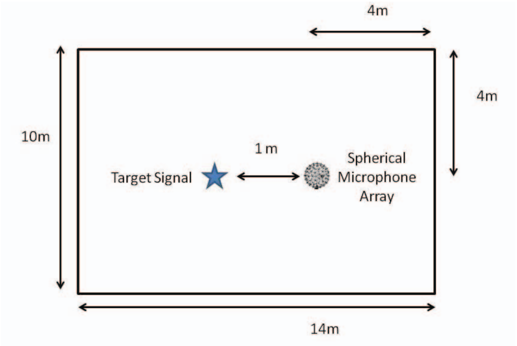

Fig. 1. The geometry of the simulation experimental setup is shown.

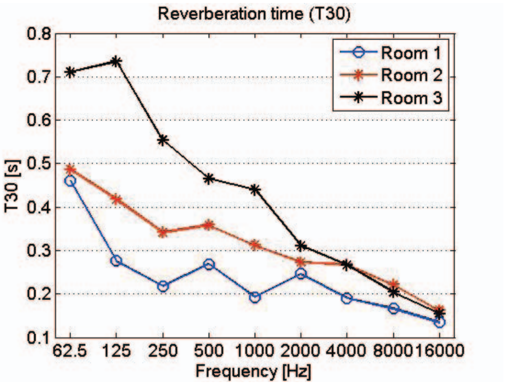

Fig. 2. The reverberation time for each simulated shoebox room.

$$\mathrm{SegSRR} = \frac{1}{N_{\mathrm{seg}}} \sum_{l=0}^{N_{\mathrm{seg}}-1} \mathrm{NSRR}(l) , \qquad (12)$$

where the NSRR is calculated as:

$$\mathrm{NSRR} = 10 \log_{10}\left(\frac{\|\gamma s_d\|_2^2}{\|\hat{s} - \gamma s_d\|_2^2}\right) \mathrm{dB} , \qquad (13)$$

s_d is the target signal after propagation through the direct path, $\hat{s}$ is the estimated target signal and γ is a scalar which is obtained using the least square minimization:

$$\gamma = \underset{x}{\operatorname{argmin}} \|\hat{s} - x s_d\|_2^2 . \qquad (14)$$

4. RESULTS AND DISCUSSION

The performance of the proposed algorithm compared with the other algorithms is shown in Tables 1-3. We refer to our algorithm as a sparse recovery (SR) dereverberation algorithm. As a reference, the PESQ measure and SegSRR are also calculated for the raw microphone signals. According to the results, we see that the proposed algorithm outperforms

Table 1. The PESQ score and SegSRR for various speech enhancement algorithms in Room 1

Method	PESQ	SegSRR
Raw Microphone Signal	2.56	5.46
ICA approach	2.80	5.90
MUSIC+MINT	2.90	9.29
SR dereverberation algorithm	**3.03**	**11.47**

Table 2. The PESQ score and SegSRR for various speech enhancement algorithms in Room 2

Method	PESQ	SegSRR
Raw Microphone Signal	2.27	1.83
ICA approach	2.61	5.15
MUSIC+MINT	2.69	6.36
SR dereverberation algorithm	**2.76**	**6.93**

Table 3. The PESQ score and SegSRR for various speech enhancement algorithms in Room 3

Method	PESQ	SegSRR
Raw Microphone Signal	2.03	0.48
ICA approach	2.47	5.05
MUSIC+MINT	2.51	5.27
SR dereverberation algorithm	**2.53**	**5.42**

the other techniques in all three simulated shoebox rooms. This is expected since (1) the ICA-based approach is unable to enhance the target signal efficiently at low and high frequencies due to the HOA encoding error [2] and (2) the MUSIC technique can only effectively detect the position of the target signal which makes this approach similar to beamforming-based approaches.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a dereverberation algorithm that employs sparse recovery, a CS technique. The proposed algorithm estimates the position of the target signal and its reflections. According to the results from the simulation experiment, the proposed algorithm effectively reduces the reverberation and outperforms other algorithms. The proposed method has been simulated using a single-source scenario. However, in reality, there are usually multiple sources in a reverberant environment. Thus the performance of the proposed method in a multi-source scenario will be evaluated in future work. Moreover, a psychoacoustic listening test will be designed and conducted to evaluate the subjective performance of the proposed algorithm.

References

[1] E. Candes and M. Wakin, “An introduction to compressive sampling,” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[2] N. Epain and C. Jin, “Independent component analysis using spherical microphone arrays,” *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 91–102, 2012.

[3] Haohai Sun, H. Teutsch, E. Mabande, and W. Kellermann, “Robust localization of multiple sources in reverberant environments using eb-esprit with spherical microphone arrays,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011.

[4] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*, Ph.D. thesis, Université Paris 6, Paris, France, 2000.

[5] N. Epain, C. Jin, and A. van Schaik, “The application of compressive sampling to the analysis and synthesis of spatial sound fields,” in *Audio Engineering Society Convention 127*, 2009.

[6] A. Wabnitz, N. Epain, McEwan A., and C. Jin, “Upscaling ambisonic sound scenes using compressed sensing techniques,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011.

[7] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming,” http://cvxr.com/cvx, April 2011.

[8] M. Miyoshi and Y. Kaneda, “Inverse filtering of room acoustic,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.

[9] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik, “Room acoustics simulation for multichannel microphone arrays,” in *Proceedings of the International Symposium on Room Acoustics*, 2010.

[10] R. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276 – 280, 1986.

[11] “Perceptual evaluation of speech quality (PESQ) and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codec,” ITU-T Recommendation P. 862, 2001.

[12] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*, p. 43, Springer, 2010.