24th INTERNATIONAL CONGRESS ON SOUND AND VIBRATION
23–27 July 2017, London

A SPATIAL AUDIO CODEC SYSTEM FOR ANALYSIS AND SYNTHESIS OF SOUND FIELDS

Y.H. Chiang, Y. Li, Y.Y. Lo, and Mingsian R. Bai

University of National Tsing Hua, School of Power Mechanical Engineering, Hsinchu, Taiwan
email: msbai63@gmail.com

A unified framework is proposed for analysis and synthesis of spatial sound fields. In the sound field analysis (SFA) phase, an unbaffled 24-element circular microphone array (CMA) is utilized to "encode" the sound field based on the plane-wave decomposition, whereas in the sound field synthesis (SFS) phase a 32-element rectangular loudspeaker array is employed to "decode" the target sound field using pressure matching technique. Depending on the sparsity of the sound sources, the SFA stage can be implemented in two ways. For the sparse-source scenario, a two-stage algorithm is utilized to estimate the source bearings using the minimum power distortionless response (MPDR) and the associated amplitudes of plane waves using the Tikhonov regularization (TIKR) algorithm. Alternatively, a one-stage algorithm based on compressive sensing (CS) algorithm can be used. The SFA technique for the nonsparse-source scenario is useful in establishing the room response model, as required in the pressure matching step of the SFS phase. Three SFS approaches with optimal regularization parameters are compared in terms of localization performance and audio quality. The integrated acoustic array system is validated with localization performance. Experimental results are presented.

Keywords: sound field synthesis, circular microphone array

1. Introduction

Recording and reproducing spatial sound fields not only has fundamental significance in audio signal processing but also play an important role in applications of home theatre , videoconferencing, virtual reality, augmented reality, auralization, etc. In this paper, an integrated acoustic array system is proposed for analysis and synthesis of a reverberant sound field. Recently, an integrated array system was developed on the basis of freefield model for spatial audio recording and reproduction [1]. This paper extended the previous work to reverberant environment where a live room is fitted with reflective walls.

Figure 1 shows the generic framework of an integrated spatial audio system. In analogy to the analysis and synthesis filter bank in signal processing, the proposed technique is comprised of a microphone array for sound field analysis (SFA) and a loudspeaker array for sound field synthesis (SFS). The SFA phase is carried out in the context of a source localization and separation process. The bearings of sound sources are calculated by using direction of arrival (DOA) estimators [2], [3]. Next, inverse solution methods such as Tikhonov regularization (TIKR) [4], compressive sensing (CS) [5] and focal underdetermined system solver (FOCUSS) [6] techniques are exploited to extract the amplitude signals. As a result, the sound field of interest is "encoded" into the bearings and amplitudes of plane-wave components. In the SFS phase, a loudspeaker array is employed to "decode" the sound field previously encoded in the SFA phase through a pressure matching approach. The SFS procedure is carried out for multiple frequency bands. In this paper, three SFS approaches are introduced and compared in terms of localization performance and audio quality.

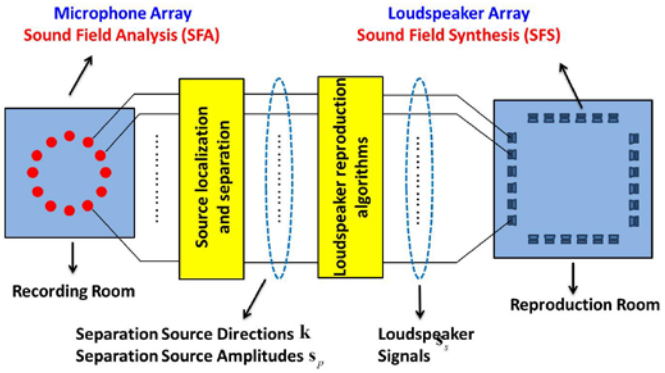


Figure 1: The generic framework of an integrated spatial audio codec system.

2. Sound field analysis (SFA)

SFA is achieved by localization and separation algorithms devised in context of sparse-source source scenario. By the term "sparse," we mean that the sources are spatially localized and fewer in number than the microphones.

2.1 Sparse Source Scenario

Suppose that sources are located far enough from a microphone array that wave fronts impinging on the array can be model as plane waves. Consider an *M*-element microphone array, where the sound pressures received at the microphones can be written as the array data vector

$$\mathbf{p}(\omega) = \mathbf{A}_p(\omega)\mathbf{s}_p(\omega) + \mathbf{v}(\omega), \tag{1}$$

where $\mathbf{s}_p(\omega) = \left[s_1(\omega) \ \cdots \ s_N(\omega) \right]^T$, denotes the Fourier transform of source signal vector at the angular frequency ω, $\mathbf{A}_p(\omega) = \left[\mathbf{a}(\theta_1) \ \cdots \ \mathbf{a}(\theta_N) \right]$ is the steering matrix [2], [3] and $\mathbf{v}(\omega)$ is the additive noise vector that is uncorrelated with the source signals. SFA stage can be performed as source localization and separation by using either of the two following approaches.

2.1.1 One-Stage CS Algorithms

CS techniques [5] are exploited to locate sources and separate their amplitudes in a single shot. Normally, we select a sufficient number of equally spaced angles as a dictionary to construct the sensing matrix $\mathbf{A}_p(\omega)$ such that the problem becomes underdetermined. The problem can be solved by using convex optimization algorithm (CVX) [7]. The angular grid must be fine enough to minimize basis mismatch. However, CVX can be very time-consuming. A more efficient method FOCUSS can be used.

2.1.2 Two-Stage MPDR and TIKR Algorithm

For the sparse-source scenario, we may also localize and separate the source signals in two steps. First, we determine the source directions using the minimum power distortionless response (MPDR) [2], [3] beamformer and separate their source signals using the TIKR algorithm.

The second stage of SFA is based on the solution of an inverse problem. We assume that the number of microphones, *M*, is greater than the number of sources, *N* such that the problem becomes overdetermined. Commonly used least-squares methods for solving this problem are pseudoinverse and the TIKR method. In this paper, TIKR method is used because it generally yields better audio

quality than pseudoinverse. The TIKR method is formulated as the following regulated least-squares problem [8]

$$\min_{\mathbf{s}_p(\omega)}\left(\left\|\mathbf{A}_p(\omega)\mathbf{s}_p(\omega)-\mathbf{p}(\omega)\right\|_2^2+\beta^2\left\|\mathbf{s}_p(\omega)\right\|_2^2\right), \tag{2}$$

where $\|\bullet\|_2$ denotes the vector 2-norm and β is a regularization parameter. The optimal solution can be shown to be

$$\hat{\mathbf{s}}_p(\omega)=\left(\mathbf{A}_p^H(\omega)\mathbf{A}_p(\omega)+\beta^2\mathbf{I}\right)^{-1}\mathbf{A}_p^H(\omega)\mathbf{p}(\omega). \tag{3}$$

The parameter β is selected to be 1% of the maximum singular value of the matrix $\mathbf{A}_p$ at low frequencies. Signals separated by the CS algorithm, the TIKR algorithm, and the FOCUSS algorithm are compared in terms of voice quality quantified with the perceptual evaluation of speech quality (PESQ) test [8, 10]. The mean opinion score (MOS) of PESQ ranges from 1 to 5, representing the speech quality from "very bad" to "excellent."

3. Sound Field synthesis (SFS)

Having parameterized the sound field in terms of bearing and amplitudes in the preceding SFA stage, the coded sound field can be decoded in the SFS stage, with the aid of a loudspeaker array.

The sound pressure reproduced by the loudspeaker array is matched to the target field at a large number of pre-selected control points *K* distributed in the area surrounded by the loudspeaker array. The pressure matching procedure can be described as the following optimization problem:

$$\min_{\mathbf{s}_s(\omega)}\left\|\mathbf{B}(\omega)\mathbf{s}_p(\omega)-\mathbf{H}(\omega)\mathbf{s}_s(\omega)\right\|, \tag{4}$$

where $\mathbf{s}_p(\omega)=\left[s_1(\omega)\ \ \cdots\ \ s_P(\omega)\right]^T$ is the amplitude vector of *P*th primary plane-wave components, $\mathbf{s}_s(\omega)=\left[s_1(\omega)\ \ \cdots\ \ s_L(\omega)\right]^T$ denotes the amplitude vector of the input signals to the *L* secondary loudspeaker sources, $\mathbf{H}(\omega)\in\mathbb{C}^{K\times L}$ denotes the room response matrix, and $\mathbf{b}_d=\left[e^{-j\mathbf{k}_d\cdot\mathbf{y}_1}\ \ \cdots\ \ e^{-j\mathbf{k}_d\cdot\mathbf{y}_K}\right]^T$ being the steering vector for the *d*th primary plane-wave component to the *n*th control point, $\mathbf{y}_n, n=1,\ldots,K$. $\mathbf{B}(\omega)=\left[\mathbf{b}_1\ \ \cdots\ \ \mathbf{b}_d\right]\in\mathbb{C}^{K\times P}$ is the steering matrix from the plane-wave components obtained in the preceding SFA stage to the control points. Therefore, the optimal solution of the problem above can be written as

$$\mathbf{s}_s(\omega)=\mathbf{H}^{\#}(\omega)\mathbf{B}(\omega)\mathbf{s}_p(\omega), \tag{5}$$

where "#" symbolizes some kind of inverse operation on the matrix $\mathbf{H}(\omega)$. The following will mention three approaches in SFS stages by using different algorithms to calculate the input signal amplitudes to the secondary sources.

4. Room response measurement and interpolation

In general, the sound field in an enclosure is composed of direct field and reflections. For large rooms with reflective boundary, the sound field is so diffuse and reverberant that the aforementioned sparse-source condition no longer holds. In this paper, an experimental approach that takes advantage of the preceding nonsparse-source SFA technique is employed to establish the room response model $\mathbf{H}(\omega)$.

The measuring procedure requires an *M*-element microphone array sitting at the center of the loudspeaker array in the reproduction room. The pressure vector of the microphones due to the *l*th loudspeaker is

$$\mathbf{p}(\omega,l)=\mathbf{h}_{mic}(\omega,l)s(\omega,l), \tag{6}$$

where $\mathbf{h}_{mic}(\omega,l)=\left[h_1(\omega,l)\quad\cdots\quad h_M(\omega,l)\right]^T$ denotes the frequency response vector associated with the *l*th loudspeaker, $l=1,\ldots,L$, with L being the number of loudspeakers, and $s(\omega,l)$ denotes the input signal to the *l*th loudspeaker. Thus, the frequency response vector $\mathbf{h}_{mic}(\omega,l)$ can be regarded as the pressure vector received at *M* microphones due to unit-amplitude source input, i.e., $s(\omega,l)=1$.

Two key steps are involve in the room response modeling. The first step is to find the coefficient vector of plane-wave decomposition. By omitting the additive noise for simplicity, the frequency response vector $\mathbf{h}_{mic}(\omega,l)$ are coded into plane-wave components with uniform angular spacing

$$\mathbf{h}_{mic}(\omega,l)=\mathbf{A}_{mic}(\omega)\mathbf{c}(\omega,l), \tag{7}$$

where $\mathbf{c}(\omega,l)$ denotes the coefficient vector of plane-wave components, $\mathbf{A}_{mic}(\omega)$ is the steering matrix from the plane waves to microphones and is given by

$$\mathbf{A}_{mic}(\omega)=\begin{bmatrix} e^{-j\mathbf{k}_1\mathbf{r}_1} & \cdots & e^{-j\mathbf{k}_N\mathbf{r}_1} \\ \vdots & \ddots & \vdots \\ e^{-j\mathbf{k}_1\mathbf{r}_M} & \cdots & e^{-j\mathbf{k}_N\mathbf{r}_M} \end{bmatrix}, \tag{8}$$

where $\mathbf{k}_n, n=1,\ldots,N$, with *N* being the number of plane waves, denotes wave number vector of the *n*th plane-wave components, and $\mathbf{r}_m, m=1,\ldots,M$ is the microphone position vector. Applying plane-wave decomposition leads to the optimization problem:

$$\min_{\mathbf{c}(\omega,l)}\left\|\mathbf{h}_{mic}(\omega,l)-\mathbf{A}_{mic}(\omega)\mathbf{c}(\omega,l)\right\|. \tag{9}$$

It follows that the optimal solution of the coefficient vector can be written as:

$$\mathbf{c}(\omega,l)=\mathbf{A}_{mic}^{\#}(\omega)\mathbf{h}(\omega,l),\quad l=1,\ldots,L, \tag{10}$$

where "#" symbolizes some kind of inverse operation of the matrix $\mathbf{A}_{mic}(\omega)$. Here, the one-stage TIKR algorithm is used to calculate the coefficient vector.

The second step is to construct the room response matrix by using the SFS procedure. Based on the preceding plane-wave decomposition, the frequency response vector at the pre-selected *K* control points ($K \gg L$) due to the *l*th loudspeaker source can be expressed as

$$\mathbf{h}_{ctr}(\omega,l)=\mathbf{A}_{ctr}(\omega)\mathbf{c}(\omega,l),\quad l=1,\ldots,L\ll K \tag{11}$$

where the steering matrix from plane waves to the control points

$$\mathbf{A}_{ctr}(\omega)=\begin{bmatrix} e^{-j\mathbf{k}_1\mathbf{r}_1} & \cdots & e^{-j\mathbf{k}_N\mathbf{r}_1} \\ \vdots & \ddots & \vdots \\ e^{-j\mathbf{k}_1\mathbf{r}_K} & \cdots & e^{-j\mathbf{k}_N\mathbf{r}_K} \end{bmatrix}, \tag{12}$$

$\mathbf{r}_k, k = 1,\ldots,K$ are the position vectors of control points. Lastly, assemble Eq. (10) for all L loudspeakers to obtain the complete room response matrix

$$\mathbf{H}(\omega) = \mathbf{A}_{ctr}(\omega)\mathbf{C}(\omega), \tag{13}$$

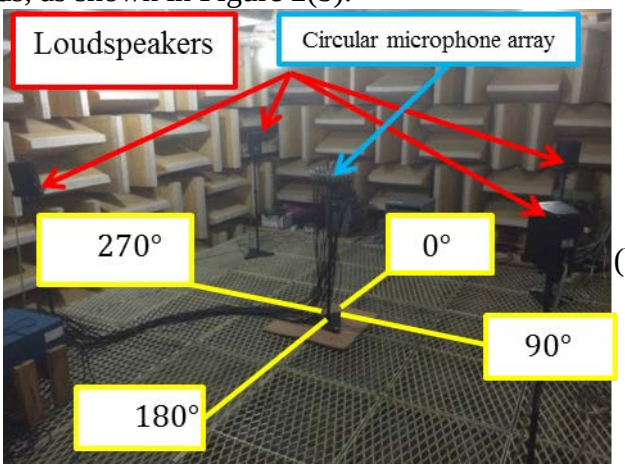
where $\mathbf{C}(\omega) = \left[\mathbf{c}(\omega,1) \quad \cdots \quad \mathbf{c}(\omega,L)\right]$ denotes the coefficient matrix from plane waves to the control points.

5. Experimental validations

Three inverse problems needing to be solved in implementing the audio codec system are in the SFA stage of Eq.(3), in the room response modelling of Eq.(10), and in the SFS stage of Eq.(5). These three inverse solution stages are ill-posed to various degrees at low frequencies. It is the most crucial to regularize the room response model. Large regularization parameter will make the inverse problem more robust. We select 10 as the parameter β for regularization in the TIKR method.

Experiments are conducted to validate the proposed audio analysis and synthesis system. In the SFA stage, a 24-element circular microphone array with 0.12m radius is utilized to capture and parameterize the sound field in an $5.4\text{m}\times3.5\text{m}\times2\text{m}$ anechoic chamber (the recording room), as shown in Figure 2(a). In the SFS stage, a rectangular, $1.62\text{m}\times1.62\text{m}$, 32-loudspeaker array is employed to reproduce in a $3.6\text{m}\times3.6\text{m}\times2\text{m}$ live room (the reproduction room) the sound field previously encoded in the SFA stage. The walls of the room are lined with acoustically reflective boards, as shown in Figure 2(b).

(a)



(b)

Figure 2: (a) The experimental arrangement for the SFA in an $5.4\text{m}\times3.5\text{m}\times2\text{m}$ anechoic room, (b) The experimental arrangement for the SFS in a $3.6\text{m}\times3.6\text{m}\times2\text{m}$ live room fitted with reflective walls.

5.1 Two-Source Example

In the experiment of SFA, a loudspeaker source positioned at the angle $\theta = 60°, 240°$ plays two 10-sec speech clips. After recording the source by CMA, we used three algorithms to extract the source signals. First, we applied the two-stage MPDR and TIKR algorithms. The MPDR spectrum is plotted as a function of angle and frequency in Figure 4(a). The resulting frequency-averaged and normalized MPDR spectrum is shown in Figure 4(b), which peaks at the angle $60°, 240°$ as desired. The results show that the source is accurately localized using MPDR. Next, the source signals are extracted using the TIKR algorithm. We also applied the one-stage CS-CVX algorithms and the one-stage FOCUSS algorithms which are exploited to locate sources and separate their amplitudes in a single shot. The result is summarized in TABLE I.

The experimental extracted signals by different algorithms have attained MOS=1.56, 3.11 and 2.84 respectively, in the PESQ test. The TIKR and CS algorithms perform comparably well in signal separation with satisfactory voice quality than the FOCUSS algorithms. However, the CS is computational expensive than the TIKR algorithm.

The SFS stage is conducted for three different methods. The coherence between the loudspeaker and the microphone is poor below 200 Hz, so the signals below 200 Hz are not processed. Method 1, band-limited processing is applied from 200 Hz to the spatial aliasing frequency 952 Hz in the SFS stage. In this frequency range, pressure matching is performed on the basis of the room response model. Below 200 Hz, unprocessed audio signals are fed directly to the loudspeakers. Above 952 Hz, simple vector panning [12] approach is adopted. Optimal regularization parameter β that achieves the highest MOS in room response modelling is calculated using the golden section search algorithm (GSS) [11] is 0.0008634.

Method 2, instead of vector panning method, we use delay and sum algorithm [2] to process the signals above 952 Hz. Method 3, we use pressure matching to obtain the signals above 200 Hz. Utilizing different regularization parameter in pressure matching will cause different results in localization performance and audio quality.

Figure 5(a) and Figure 5(c) show the MPDR spectrum and the normalized MPDR spectrum obtained using Method 3 for β = 0.01 and 10, respectively. The small regularization parameter β yields better localization performance than the large β. These two signals are compared to the clean signal via the PESQ test. The results show that the large β ensures satisfactory voice quality, whereas small β leads to impaired voice quality.

Three methods localization results are show in Figures 6(a)-(f).The MPDR spectrum are plotted as a function of angle and frequency in Figures 6(a), (c) and (f).The resulting frequency-averaged and normalized MPDR spectrum is shown in Figures 6(b), (d) and (f).

(a) (b)

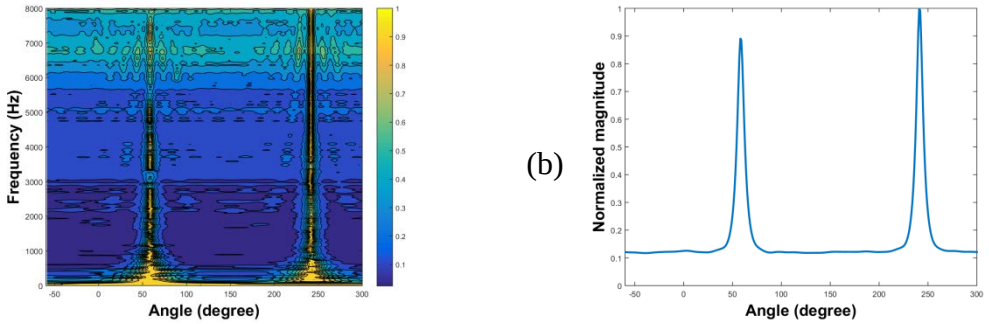
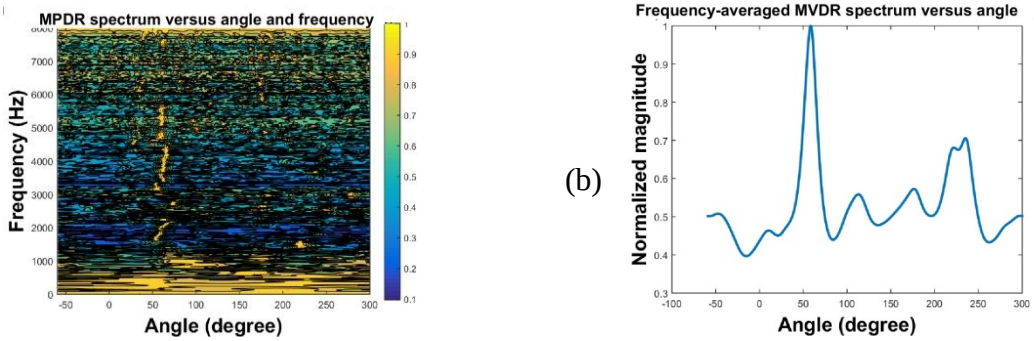
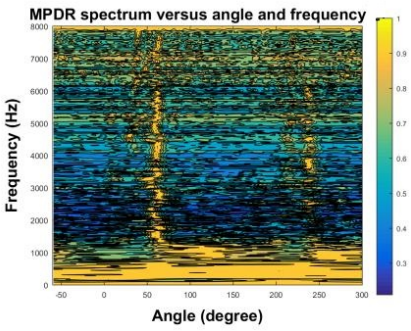


Figure 4: Localization results in the SFA simulation for two speech sources located at 60° and 240°.
(a) MPDR spectrum plotted versus angle and frequency, (b) frequency-averaged and normalized MPDR spectrum.

(a) (b)



(c)



(d)

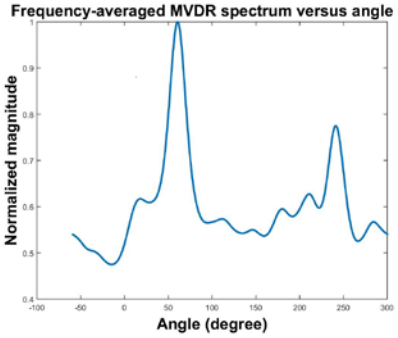
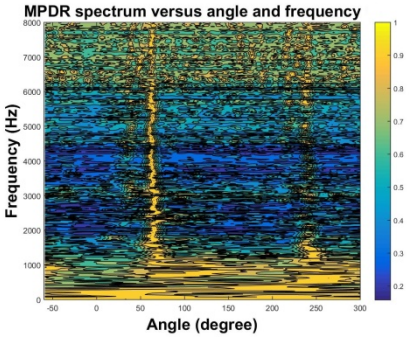
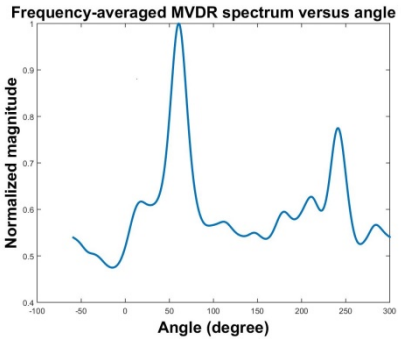


Figure 5: Localization results in the SFS experiment by method 3 with different regularization parameter. β =10 (a)-(b) and β =0.01 (c)-(d).

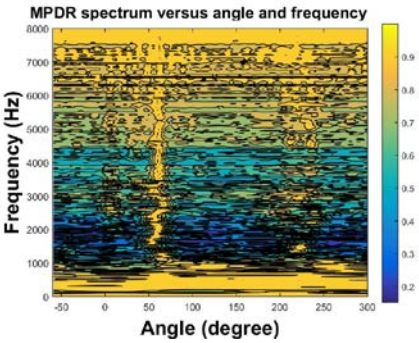
(a)



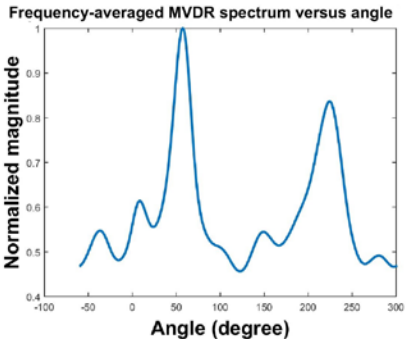
(b)



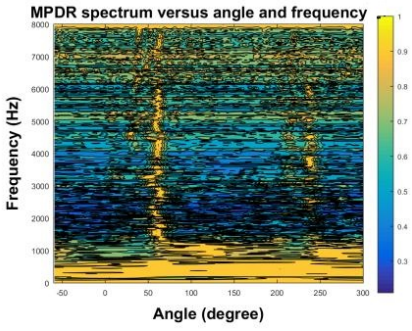
(c)



(d)



(f)



(g)

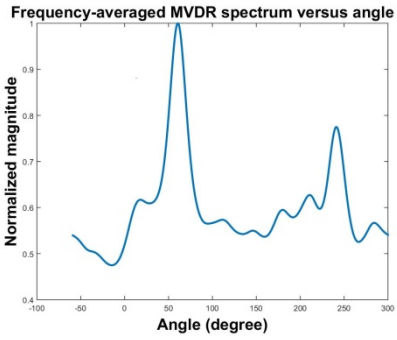


Figure 6: Localization results in the SFS experiment by three different approaches. (a)-(b) Method 1, (c)-(d)Method 2, and (e)-(f)Method3.

Table 1: The MOS of PESQ for the source signal extracted using three methods (experiment)

Signal	$60°$ (male speech)	$240°$ (female speech)	CPU time (sec)
Clean Signal	4.5	4.5	
One-stage FOCUSS	1.56	1.61	275
One-stage CS	3.11	2.99	27588
Two-stage TIKR	2.84	2.79	8

6. Conclusions

Three inverse problems are required to be solved in implementing the audio codec system. Because of the ill-posedness at low frequencies, especially in the room response modelling stage, choosing an appropriate regularization parameter β is crucial in the inverse solution. Therefore, room response modelling generally requires a larger regularization parameter than those used in the SFA and SFS stages. There is a tradeoff between localization performance and voice quality in the synthesized sound field. In general, large β results in a small solution norm with good voice quality, whereas small β yields small residual norm with good localization performance.

In the analysis stage, the one-stage CS algorithm is computationally more expensive than the two-stage TIKR algorithm. In the synthesis stage, method 1 performs well in localization, but not as well in the reproduced voice quality. As compared with method 3, method 2 leads to the reproduced signals with boosted high-frequency content above 952 Hz, and localization is poor. Method 3 has performed the best in terms of voice quality and localization performance.

RFERENCES

1 M. R. Bai, Y. H. Hua, C. H. Kuo, and Y. H. Hsieh, “An Integrated Analysis- Synthesis Array System for Spatial Sound Fields,” J. Acoust. Soc. Am., vol. 137, no. 3, pp. 1366-1376, Mar. 2015.

2 M. R. Bai, J.-G. Ih, and J. Benesty, Acoustic Array Systems, Wiley-IEEE Press, 2013, 1st edition, Singapore, Chaps. 3-4, 2013.

3 J. Capon, “High-Resolution Frequency-Wavenumber Spectrum Analysis,” Proc. IEEE, vol. 57, no. 8, pp.1408-1418, Aug. 1969.

4 C.W. Groetsch, “The theory of Tikhonov regularization for Fredholm equations of the first kind,” Pitman Advanced Pub. Program, Boston (1984).

5 E. J. Candes and M. B. Wakin, “An Introduction to Compressive Sampling,” IEEE SP Mag., vol. 25, no. 2, pp. 21-30, Mar. 2008.

6 I.F. Gorodnitsky and B.D. Rao “Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm,” Proc. IEEE, vol. 45, no. 3, Mar. 1997.

7 M. R. Bai and C. C. Chen, “Application of Convex Optimization to Acoustical Array Signal Processing,” J. Sound Vib., vol. 332, no. 25, pp. 6596-6616, Dec. 2013.

8 M. Bertero, T.A. Poggio and V. Torre Ill-Posed Problems in Early Vision, Proc. IEEE, 76(8), 869-889, (1988).

9 ITU-T Recommendation, “Perceptual evaluation of speech quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs,” ITU-T Recommendation P.862, Feb. 2001.

10 ITU-R Recommendation, “Subjective Assessment of Sound Quality,” ITU-R BS.562-3. pp. 1-8. Dec. 2003.

11 R. P. Brent, “Algorithms for Minimization without Derivatives,” Prentice-Hall, Inc., Englewood Cliffs, New Jersey, pp. 48-75, 1973.

12 Y. H. Kim and J. W. Choi, “Sound Visualization and Manipulation,” Wiley, Singapore, 2013.