REDUCTION OF SPATIAL SAMPLING REQUIREMENT IN SOUND-BASED SYNTHESIS

Cac Nguyen, Robert L. Morrison, Jr., and Minh N. Do

Department of Electrical and Computer Engineering
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Email: {tnguyen8, rlmorris, minhdo}@uiuc.edu

ABSTRACT

We study the problem of synthesizing the sound field at arbitrary locations and times from the recordings of an array of audio sensors. Given prior estimates of the locations and frequencies of the sound sources, such as those obtained using adaptive source localization, we characterize the spatio-temporal support of the sound field spectrum. This characterization allows the spatial sampling requirements to be reduced in comparison to when no prior estimates of the sources are utilized. We derive an adaptive interpolation kernel, based on the estimated spectral support, to reconstruct the sound-field function using measurements from sensors on a coarse spatial-sampling grid. Simulation results demonstrate the gain achieved in reduced sampling requirements by using the proposed adaptive interpolation approach.

1. INTRODUCTION

Current audiovisual recording systems offer a limited sensory experience. The advent of compact low-cost sensors and unprecedented computing power has motivated the concept of immersive recording systems enabling dynamic playback. Such systems allow the audio and visual field to be reconstructed at arbitrary positions in space based on recordings from an array of sensors. Thus, the users are able to see and hear events at the positions and perspectives of their choices, navigating through a virtual world by means of a joystick or motion sensors.

In this paper we study the reconstruction of the continuous *sound-field function*, which is defined as the sound signal at any position and time, using samples of the sound field taken at an array of microphones. The problem of sampling and reconstructing the sound field in this manner was first considered by Ajdler and Vetterli [1]. In our previous analysis [2], we determined the spatio-temporal sampling requirements for reconstructing the sound field for the *far-field* case where the sound sources are located at a great distance from the sensor array. In determining the sampling requirements, we made no prior assumptions about the locations or frequencies of the sound sources. As such, the analysis in the previous work provides a worst-case estimate of the sampling requirements when no prior information is available.

The sensor spacing requirements based on Nyquist sampling using previous analyses [1, 2] are prohibitively small; microphone spacings about 5 cm are required for alias-free reconstruction of the sound field in practical scenarios. As the sources are closer to the microphone array (the *near-field* case), the required density of the sensors based on Nyquist analysis even becomes greater. Recently, Gallo et al. [3] developed an efficient scheme for sound-field synthesis by decomposing recording signals into time-frequency atoms and estimating the position in space from which these atoms were emitted.

The goal of this work is to reduce the spatial sampling requirements using prior knowledge about the sound sources. Specifically, we assume prior estimates of the source positions (i.e., the sources are located within a particular region of space with a known displacement and angular orientation with respect to the sensor array), and the range of temporal frequencies produced by the sources, are available. Such prior knowledge can be obtained in practice using adaptive source localization techniques. Using this information, we characterize the corresponding spatio-temporal spectrum, where the spectral support is greatly reduced in comparison to when no knowledge of the sources is utilized. From the spectral characterization, we derive an adaptive interpolation kernel that allows the continuous sound field to be reconstructed from a lower density of sensors. As the relative positions of the sources change, the interpolation kernel adapts so that the alias-free sound field can be reconstructed using a fixed low density sensor grid. Our method is thus can serve as a bridge and combine the two opposite approaches in sound-field synthesis: the one solely bases on interpolation and the other relies on exact source localization.

The organization of the paper is as follows. In Section 2, we formulate the problem and introduce a model for the sound field. Section 3 presents an analysis of the sound field spectrum in the case where we focus on one-dimensional sensor arrays and the sources being contained within a particular region in space. We derive expressions for the spectral support that are used to determine reduced sensor spacing requirements. We propose an adaptive shearing-based interpolation scheme to implement the reduction of sensor density requirement in Section 4. This section also includes experimental results using the proposed sound-field reconstruction approach.

2. PROBLEM SETUP

In the sound field of which we do not have knowledge, like the number of sources and the sources' location, we use an array of microphones to record sound signal. From the recorded data, we want to reconstruct the sound field at arbitrary positions and times. In this paper we consider linear array with identical spacing. We also assume a free field recording environment, where the response of obstructing elements such as walls are ignored. This assump-

This work was supported by NSF grants CCF-0312432 and CCF-0430877.

tion is relevant in practice when considering the sound-field in a stadium or open field. A linear medium is also assumed, where the sound field at each point in space is the superposition of the delayed and attenuated source signals.

We focus on the case of a one-dimensional array as shown in Figure 1 where the sound field is recorded along the x-axis (i.e. $y = 0$) and a number of sources at positions (x_i, y_i) are considered. An observer at coordinate $(x, 0)$ and time t experiences the sound field

$$r_i(x,t) = \frac{1}{d_i^2(x)}\, s_i\left(t - \frac{d_i(x)}{c}\right) \tag{1}$$

from a source i, where $s_i(t)$ is the source signal at (x_i, y_i), $d_i(x)$ is the Euclidean distance between the source and observer: $d_i(x) = \sqrt{(x-x_i)^2 + y_i^2}$, and $c = 343$ m/s is the speed of sound in air at 20 degrees Celsius.

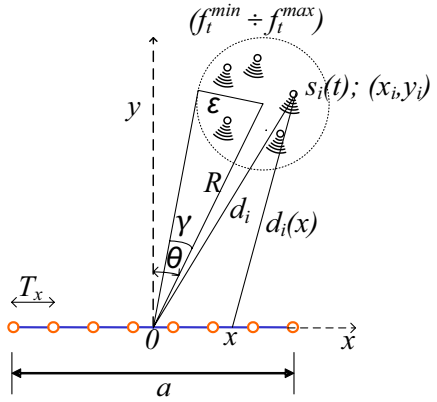


Fig. 1. Illustration of a one-dimensional array of microphones. A source signal $s_i(t)$ is emitted from the coordinate (x_i, y_i).

We consider P (could be infinite) sources and an array of finite aperture a. The observed sound field along the x-axis can be expressed as

$$r(x,t) = \begin{cases} \sum_{i=1}^{P} r_i(x,t) & |x| \le \frac{a}{2} \\ 0 & \text{elsewhere.} \end{cases} \tag{2}$$

The general sound-synthesis problem can be formulated as follows. Given recording of the sound field $r[m,n] = r(mT_x, nT_t)$, $(m,n \in \mathbb{Z})$, where T_x is the spatial sampling interval (in x dimension) and T_t is the temporal sampling interval (in t dimension), we wish to reconstruct the continuous sound-field $r(x,t)$ over the aperture $|x| \le a/2$. The restriction of finite aperture $|x| \le a/2$ in (2) is applicable in practice as we normally interpolate the sound field using only nearby microphones; e.g. in clusters.

The objective of this work is to use prior estimates of the sound source positions, and their range of temporal frequencies, to decrease the spatial sampling requirements. We assume that we know the source positions with some degree of uncertainty as shown in Figure 1. This can be formalized by considering a ball of radius ϵ where all of the sources are located. Here, we have knowledge of the uncertainty ϵ, as well as the distance R and angle θ from the center of the sensor array to the center of the uncertainty ball. In addition, we know the minimum and maximum temporal frequencies, f_t^{min} and f_t^{max}, respectively. Given these parameters, and a fixed low-density sensor array, the goal is to design the interpolation kernel to recover the alias-free continuous sound field.

3. SPECTRAL ANALYSIS

Given the model in (1) and (2) the spectrum of the sound field $r(x,t)$ is determined by taking a 2-D Fourier transform with respect to x and t as

$$\begin{aligned} R(f_x, f_t) &= \sum_{i=1}^{P} \int_{-a/2}^{a/2} \int_{-\infty}^{\infty} r_i(x,t) e^{-j2\pi(f_x x + f_t t)} dt dx, \\ &= \sum_{i=1}^{P} S_i(f_t) \int_{-a/2}^{a/2} \frac{1}{d_i^2(x)} e^{-j2\pi(f_x x - f_t d_i(x) c^{-1})} dx, \end{aligned}$$

where $S_i(f_t) \overset{\text{def}}{=} \int_{-\infty}^{\infty} s_i(t) e^{-j2\pi f_t t}$ is the 1-D Fourier transform of the source signal $s_i(t)$.

Extending the far-field case analysis in [2] we consider sources that are not at infinite distance but at distances d_i of several times the length of the array a. In this case, within the aperture $|x| \le a/2$ we can use a first-order Taylor approximation

$$d_i(x) = \sqrt{(x-x_i)^2 + y_i^2} \approx d_i - \frac{x_i}{d_i} x = d_i - x \sin\theta_i, \tag{3}$$

where d_i is the distance and θ_i is the angle from the i source to the origin. Then Fourier transform $R(f_x, f_t)$ can be written as

$$R(f_x, f_t) \approx \sum_{i=1}^{P} S_i(f_t) e^{-j2\pi f_t d_i c^{-1}} K_i(f_x + f_t c^{-1} \sin\theta_i), \tag{4}$$

with the kernel

$$\begin{aligned} K_i(f) &= \int_{-a/2}^{a/2} \frac{1}{d_i^2(x)} e^{-j2\pi f x} dx \\ &= \int_{-\infty}^{\infty} \text{rect}\left(\frac{x}{a}\right) \times \frac{1}{(x-x_i)^2 + y_i^2} e^{-j2\pi f x} dx \\ &= a\,\text{sinc}(af) \ *_f \ \frac{\pi}{y_i} e^{-j2\pi f x_i - 2\pi y_i |f|}, \end{aligned} \tag{5}$$

where the last equality is due to the Laurentzian pair of Fourier transform [4]. The kernel magnitude $|K(f)|$ decays like $1/|f|$ and has the essential support in interval $|f| \le \frac{2.5}{a} + \frac{1}{y_i}$) (the ratio of the cutoff value to the maximal value is about 10 percent). Using this, from (4) we see that the spectral support of $r(x,t)$ is confined on the following region

$$\mathcal{F}_r = \bigcup_i \left\{ (f_x, f_t) \ : \ |f_x + f_t c^{-1} \sin\theta_i| \le \frac{2.5}{a} + \frac{1}{y_i} \right\}. \tag{6}$$

Note that the related result for the far-field setting in [2] can be seen as a special case of (6) with $y_i = \infty$ and $a = \infty$.

If we have no prior knowledge of the source locations, then we have to consider all possible values of y_i and θ_i for the union in (6), and in general this leads to a "bow-tie" frequency region as shown in [1, 2]. When when some estimate of the locations and frequencies of the sources is available then (6) can be used

to obtain a smaller spectral support of $r(x,t)$ and thus reduce the spatial sampling requirement.

In particular, we consider a case where all sources are located within a ball of radius ϵ, as shown in Figure 1. Here the ball is located at a distance R from the origin and oriented at angle θ with respect to the y-axis. It follows that angles of the sources θ_i are in the range $[\theta-\gamma, \theta+\gamma]$, where $\sin\gamma = \epsilon/R$. Consequently, the spatial frequency support along f_x dimension of $r(x,t)$ for a single temporal frequency $f_t \geq 0$ is

$$-f_t c^{-1}\sin(\theta+\gamma) - \frac{2.5}{a} - \frac{1}{R\cos\theta} \leq f_x \leq -f_t c^{-1}\sin(\theta-\gamma) + \frac{2.5}{a} + \frac{1}{R\cos\theta}. \quad (7)$$

The spectral support region for $f_t \leq 0$ is symmetric across the origin.

As an example, we consider a scenario with human speech sources with a range of temporal frequencies up to 4 kHz in a ball region with $\epsilon = 2$ m, $R = 10$ m, $\theta = \pi/6$, and the array length $a = 2$ m. The actual sound field spectrum shown in Figure 2 matches with theoretical estimate from (7).

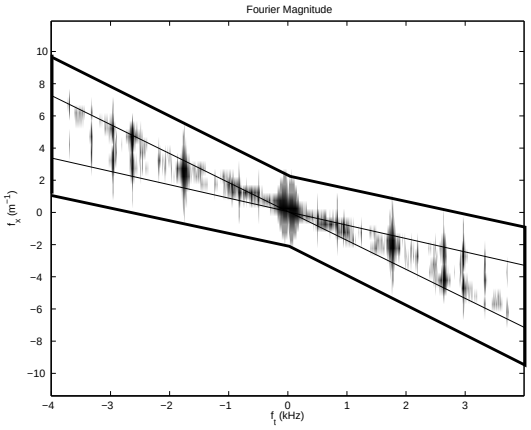


Fig. 2. Actual spectrum of a sound field with sources in a ball region. The theoretical estimate (7) of the spectral support is in the sheared bow-tie region with thick boundary.

4. PROPOSED INTERPOLATION

In this section, we propose an approach for implementing the reduction in spatial sampling requirement using the spectrum analysis in the previous section. Normally, when we reconstruct the signal at an arbitrary position on the array, we separately interpolate recorded data from microphones in spatial and temporal dimensions with respect to rectangular sampling. More effective interpolation can be obtained by jointly consider spatial and temporal dimensions using multidimensional convolution, but at a much higher computational cost. Specifically, suppose that we want to reconstruct the signal at an arbitrary position $x^{(a)}$ and time t_0, then a "normal" interpolation would use the actual recordings at that time of the nearby microphone as

$$r(x^{(a)}, t_0) = \sum_m r(x^{(m)}, t_0)\ \mathrm{sinc}\left(\frac{x^{(a)} - x^{(m)}}{T_x}\right), \quad (8)$$

where T_x is the distance between two successive microphones, and $x^{(m)}$ is the position of the microphone m in the x-axis. The accuracy of the interpolation in (8) depends on the compactness is the spatial bandwidth (i.e. along the f_x dimension) of $r(x,t)$.

We propose an efficient and effective interpolation method bases on the idea of *shearing*. From the previous section we know that the spectrum support of $r(x,t)$ is in a sheared bow-tie region that is concentrated along the line $f_x + f_t c^{-1}\sin\theta$ (see Figure 2). To reduce the required spatial Nyquist frequency, we can shear the spectrum to the line $f_x = 0$ using the following transform in the frequency domain

$$\begin{cases} f_x' = f_x + f_t c^{-1}\sin\theta, \\ f_t' = f_t. \end{cases} \quad (9)$$

This transform is illustrated in Figure 3.

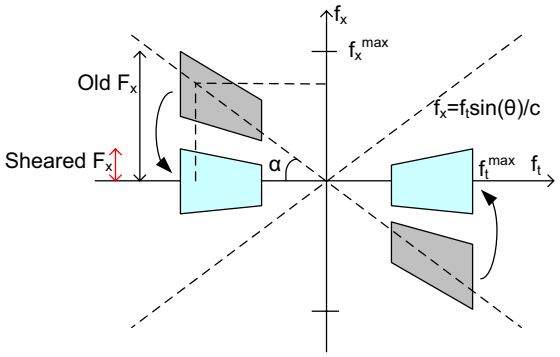


Fig. 3. Illustration of shearing in frequency domain. The sheared spectrum has reduced spatial bandwidth along f_x dimension.

In the spatial domain the signal $r(x',t')$ corresponding to the above sheared spectrum is determined by shearing the original data $r(x,t)$ with the transform [4]

$$\begin{cases} x' = x, \\ t' = t - xc^{-1}\sin\theta. \end{cases} \quad (10)$$

The sheared signal $r(x',t')$ has reduced spatial bandwidth, and thus can be used in interpolation (8) instead of $r(x,t)$ with greater accuracy. Then substituting the transform from (10) we obtain the following interpolation for the original signal $r(x,t)$

$$r(x^{(a)}, t_0) = \sum_m r\left(x^{(m)}, t_0 - \frac{(x^{(m)} - x^{(a)})\sin\theta}{c}\right) \times \mathrm{sinc}\left(\frac{x^{(a)} - x^{(m)}}{T_x}\right). \quad (11)$$

In practice, the sound field is also sampled on the temporal dimension t. The required temporal locations $t_0 - ((x^{(m)} - x^{(a)})\sin\theta)/c$ generally do not fall into discrete locations nT_t. In that case we simply need to apply a fractional delay filter [5] to the sampled signal from each microphone m.

The shearing interpolation (11) effectively uses data samples of the sound field $r(x,t)$ along the lines $t + xc^{-1}\sin\theta = \text{const}$. In

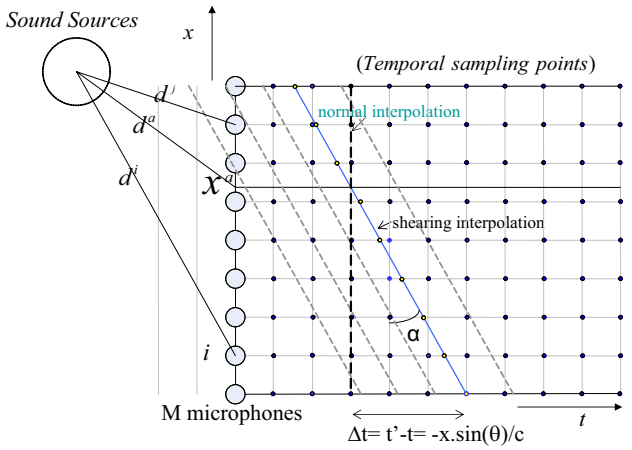


Fig. 4. Illustration of interpolation trajectories.

comparison, the normal interpolation (8) uses data samples along the lines $t = \text{const}$. Figure 4 illustrates these trajectories.

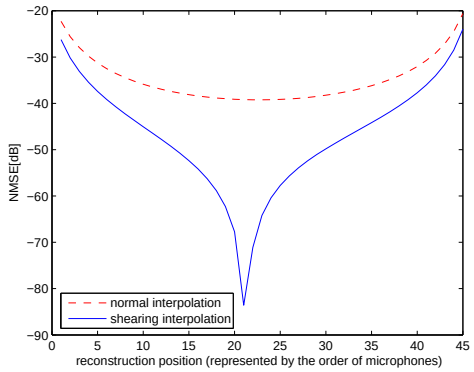
The sheared lines $t + xc^{-1} \sin\theta = \text{const}$ equalize the delays of far-field sources generating plane waves arriving at angle θ. Signals coming from the ball oriented at angle θ as depicted in Figure 1 have similar geometrical property with these plane waves. Intuitively, (11) interpolates along the sheared lines where the spatial variation of $r(x,t)$ is low, and thus accurately reconstructs the signal using a coarse spatial sampling grid.

Figure 5(a) shows normalized mean square errors (NMSEs) using the normal interpolation (8) and proposed shearing interpolation (11) for the simulated setting as described at the end of Section 3 (i.e. sources in a ball region with $\epsilon = 2$ m, $\theta = \pi/3$, $R = 10$ m, and $|f_t| \leq 4$ kHz.).

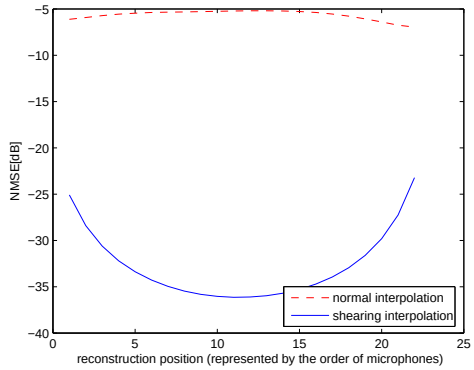
Figure 5(b) shows the interpolation NMSEs when the distance between microphones are twice increased. The special improvement here is that the NMSE performance of the shearing interpolation retains good if we increase the distance between microphone while the normal interpolation will get worse very fast. It is because the shearing interpolation is based on shearing transform so the resulting spatial bandwidth is reduced. With bigger intervals of sensors, Nyquist sampling requirement is no longer satisfied in normal interpolation but the shearing one. These results agree with our analysis showing that if we shear the spectrum to the center of f_x, then we can reduce the spatial sampling requirement.

5. CONCLUSION

This work provides an analysis of spatial bandwidth of recorded data from an array of microphones and the corresponding required sensor density for reconstructing signal at arbitrary locations. It also addresses the problem in which case can we reduce spatial sampling requirement and how to implement such reduction. The possibility of reducing sensor density in the case of limited angle of arrival suggests a potential of reconstructing a sound field with moving sources. For instance, within a small time duration, an estimate of the location of a moving source can be tracked by beamforming techniques. Based on this information, we can construct an adaptive interpolation kernel to reconstruct the sound field. As a result, the sensor density used for this interpolation is reduced



(a) Normal spacing of microphone, $T_x = 4.5$ cm.



(b) Double spacing of microphones, $T_x = 9$ cm.

Fig. 5. NMSE comparison of two interpolation methods. The setup is described in text as for Figure 2. The horizontal axes represent microphone positions along an aperture of 2 m.

compare to normal requirement. Our future work will also consider the problem in 2-D case in which use a 2-D microphone array to reconstruct the sound field.

6. REFERENCES

[1] T. Ajdler and M. Vetterli, "The plenacoustic function, sampling and reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, Hong Kong, 2003.

[2] M. N. Do, "Toward sound-based synthesis: the near-field case," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, Montreal, Canada, May 2004.

[3] E. Gallo, N. Tsingos, and G. Lemaitre, "3d-audio matting, postediting, and rerendering from field recordings," *EURASIP Journal on Advances in Signal Processing*, 2007.

[4] R. E. Blahut, *Theory of Remote Image Formation*, Cambridge University Press, 2004.

[5] T. I. Laakso, V. Välimäki, M Karjalainen, and U. K. Laine, "Splitting the unit delay – tools for fractional delay filter design," *IEEE Signal Proc. Mag.*, vol. 13, pp. 30–60, Jan. 1996.