TOWARD SOUND-BASED SYNTHESIS: THE FAR-FIELD CASE

Minh N. Do

University of Illinois at Urbana-Champaign
Department of Electrical and Computer Engineering
Coordinated Science Laboratory
Urbana, IL 61801
Email: minhdo@uiuc.edu

ABSTRACT

We consider the problem of synthesizing the sound at any desired position and time from the recording of a set of microphones. Similar to the image-based rendering approach for vision, we propose a sound-based synthesis approach for sound. In this approach, audio signals at new positions are interpolated directly from the recorded signals of nearby microphones. The key underlying problems for sound-based synthesis are sampling and reconstruction of the sound field. We provide a spectral analysis of the sound field under the far-field assumption. Based on this analysis, we derive the minimum sampling and optimal reconstruction for several common settings.

1. INTRODUCTION

Existing audiovisual recording systems use a single camera and microphone, and thus provide viewers with a passive viewing experience. We envision that, thanks to the continuing improvement in digital technology that now offers low-cost sensors and massive computing power, new systems employing multiple cameras and microphones, together with sophisticated processing algorithms delivering unprecedented immersive recording and viewing capabilities, are now feasible. We believe that sufficient sensing, networking, and computing power to practically address this vision already exists; the critical gap in achieving it lies in new signal processing theory and algorithms.

For visual signals, synthesizing new views of a scene directly from a set of acquired views is known as *image-based rendering* (IBR) [1]. In contrast with traditional model-based rendering, IBR requires very little or no geometrical models of the scene. IBR can be seen as an application of the sampling theory to the *plenoptic function* [2] that describes the light intensity passing through every viewpoint, in every direction, for all time, and for every wavelength. In this setting, acquired views from the cameras provide discrete samples of the plenoptic function, and the synthesized view is reconstructed from the continuous plenoptic function at a given point.

Similar to image-based rendering, we propose an approach to synthesize the sound at any position directly from the recording of a set of microphones, without knowing or recovering all the sources that generate the sound field. We refer to this approach as *sound-based synthesis*. As for IBR, the key issues in sound-based synthesis is how to *model*, *sample*, and *reconstruct* a sound field.

The sampling and reconstruction questions of a sound field has been first studied by Ajdler and Vetterli [3]. In that paper, the authors introduce and study the sampling and reconstruction problems of the *plenacoustic function*, which characterizes the impulse responses of a sound field, in particular in a room. Knowing this function, the actual sound at a desired position can be obtained via the convolution with the source signal. Due to the effects of the reflections on the walls and reverberations, spectral analysis of the impulse responses of a room can be quite complicated.

In this paper we consider directly the sound field signal, which is defined as what would be heard or recorded at any position and time. Addressing the sampling problem of a sound field would lead to the reconstruction of actual sound at any position. We consider the the far-field case, for example in a stadium or an open field, where certain assumptions significantly simplify the spectral analysis. Section 2 formulates the problem and introduces our model of a sound field. Section 3 provides a spectral analysis of the sound field under the far-field assumption. Based on this analysis, minimal sampling and optimal reconstruction are derived for several cases in Section 4. We conclude with some discussions in Section 5

This work was supported by an NSF ITR grant CCR-0312432.

2. PROBLEM FORMULATION

We define the sound field signal $r(\boldsymbol{p}, t)$ as the signal value that would be heard or recorded at the position $\boldsymbol{p} = (x, y, z)^T$ and time t. The sound field signal $r(\boldsymbol{p}, t)$ is the sum of each source signal delayed and attenuated according to the received position $\boldsymbol{p}$. We parameterize the emitting signal from the i-th source by $s_i(t)$ as the signal that would be received at the reference position $\boldsymbol{p} = \mathbf{0}$. Thus, the sound field signal at arbitrary position $\boldsymbol{p}$ is

$$r(\boldsymbol{p}, t) = \sum_i a_i(\boldsymbol{p})\ s_i(t - \tau_i(\boldsymbol{p})), \tag{1}$$

where $a_i(\boldsymbol{p})$ and $\tau_i(\boldsymbol{p})$ is the attenuation and delay, respectively, for the source signal $s_i(t)$ at the position $\boldsymbol{p}$. We make no assumption about the number of sources, which can be dynamic and infinite. These sources include both actual sound sources (e.g. speakers) as well as virtual sources (e.g. due to reflection).

The attenuation for sound is typically equal to the inverse of square of the distance from the source. In that case, we have

$$a_i(\boldsymbol{p}) = \frac{\|\boldsymbol{p}_i\|_2^2}{\|\boldsymbol{p}_i - \boldsymbol{p}\|_2^2}, \tag{2}$$

where $\boldsymbol{p}_i$ is the position of the source i. In the far-field case, where the sound sources are far away and we are only interested in a relatively small region near the reference position, then $\|\boldsymbol{p}_i\| \gg \|\boldsymbol{p}\|$. In that case, we have

$$a_i(\boldsymbol{p}) \approx 1 \tag{3}$$

The time delay is equal to the sound propagation distance divided by the speed of sound λ (typically $\lambda = 343m/s$). Generally, as for the attenuation in (2), $\tau_i(\boldsymbol{p})$ is a non-linear function of $\boldsymbol{p}$. However, in the far-field case where the sound sources are far away, we can consider the wavefronts to be parallel planes. Let $\boldsymbol{u}_i$ denote the unit normal vector of the wavefront planes of the source i. Then the signed distance of wavefront propagation of the source i from the reference position $\mathbf{0}$ to the position $\boldsymbol{p}$ is $\langle \boldsymbol{u}_i, \boldsymbol{p} \rangle = \boldsymbol{u}_i^T \boldsymbol{p}$; see Figure 1. In that case, $\tau_i(\boldsymbol{p})$ is a linear function

$$\tau_i(\boldsymbol{p}) = \frac{\boldsymbol{u}_i^T \boldsymbol{p}}{\lambda}. \tag{4}$$

Our input is the recording samples from a set of microphones at fixed positions. Hence, effectively we have discrete samples of the sound field function $r(\boldsymbol{p}, t)$. Our task is to synthesize from these samples the sound that would be recorded (or heard) at any desired position.

The "conventional" approach to this problem is to locate all the sound sources in the field, to recover those sources, and to reconstruct the sound at any position. Sound source recovery is a nontrivial problem, especially in dynamic environments. A popular method for source recovery is *adaptive beamforming*. Recent research using modern adaptive beamforming techniques [4] has shown promising results for sound source recovery in controlled environments with a few sources and perfect knowledge of the source positions. However, such method only works when the number of sources is small, for example in a room.

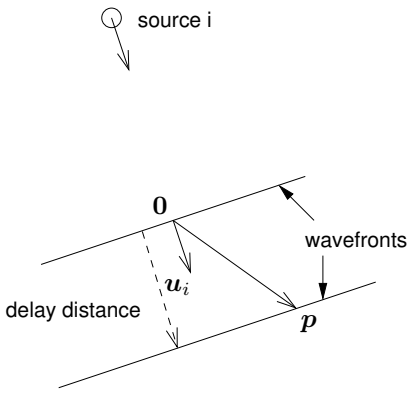

Fig. 1. Delay distance of wavefront propagation in the far-field case.

In large dynamic environments with many sources, such as a sport event, there is little hope for recovering individual sound source. In such cases, an attractive alternative is to directly synthesize the sound signals at desired positions, without explicitly recovering the sound sources. This sound-based synthesis approach for sound is similar to the image-based rendering for vision. The key underlying problems for the sound-based synthesis are sampling and reconstruction of the sound field function $r(\boldsymbol{p}, t)$.

3. SPECTRAL ANALYSIS

To address the sampling and reconstruction of the sound field function $r(\boldsymbol{p}, t)$ we need to analyze its spectral support. The similar analysis was carried for the plenoptic function [5, 6] and plenacoustic function [3].

The Fourier transform of the sound field function $r(\boldsymbol{p}, t)$ can be written as

$$\begin{aligned} R(\boldsymbol{f}_{\boldsymbol{p}}, f_t) &= \int\int r(\boldsymbol{p}, t) e^{-j2\pi(\boldsymbol{f}_{\boldsymbol{p}}^T \boldsymbol{p} + f_t t)} d\boldsymbol{p} dt \\ &= \sum_i \int\int s_i(t - \tau_i(\boldsymbol{p})) e^{-j2\pi f_t t} dt \\ &\qquad a_i(\boldsymbol{p}) e^{-j2\pi \boldsymbol{f}_{\boldsymbol{p}}^T \boldsymbol{p}} d\boldsymbol{p} \\ &= \sum_i S_i(f_t) \int a_i(\boldsymbol{p}) e^{-j2\pi(\boldsymbol{f}_{\boldsymbol{p}}^T \boldsymbol{p} - f_t \tau_i(\boldsymbol{p}))} d\boldsymbol{p}, \end{aligned} \tag{5}$$

where $S_i(f_t) = \int s_i(t)e^{-j2\pi f_t t}dt$ is the Fourier transform of the source signal $s_i(t)$.

In the far-field case, substituting (4) into (5) we get

$$R(\boldsymbol{f_p}, f_t) = \sum_i S_i(f_t) A_i \left(\boldsymbol{f_p} - \frac{f_t \boldsymbol{u}_i}{\lambda} \right), \qquad (6)$$

where $A_i(\boldsymbol{f_p}) = \int a_i(\boldsymbol{p})e^{-j2\pi \boldsymbol{f_p}^T \boldsymbol{p}}d\boldsymbol{p}$ is the Fourier transform of the attenuation function $a_i(\boldsymbol{p})$.

The key observation is that because of (3), function $A_i(\boldsymbol{f_p})$ is concentrated around $\boldsymbol{f_p} = \boldsymbol{0}$. Indeed, if $a_i(\boldsymbol{p}) = 1$ then $A_i(\boldsymbol{f_p}) = \delta(\boldsymbol{f_p})$. Thus, from (6) the spectral support of the sound field function $r(\boldsymbol{p}, t)$ is around the region

$$\boldsymbol{f_p} = \frac{f_t \boldsymbol{u}_i}{\lambda}, \quad \text{where} \quad S_i(f_t) \neq 0. \qquad (7)$$

This region is in turn specified by the spectral supports of the source signals $s_i(t)$ and the ranges of $\boldsymbol{u}_i$.

To gain more insight about this spectral support region, we consider the following two particular cases. In the first case, we consider the sound field function along the line $\boldsymbol{p} = (x, 0, 0)^T$. This corresponds to the common case where the microphones are placed a long a line. The sound sources are supposed to be on the plane $z = 0$. The unit normal vectors $\boldsymbol{u}_i$ of the wavefront planes for the source i can be parameterized by

$$\boldsymbol{u}_i = (\sin\theta_i, -\cos\theta_i, 0)^T, \qquad (8)$$

where θ_i is commonly referred to as the direction of arrival (DOA); see Figure 2

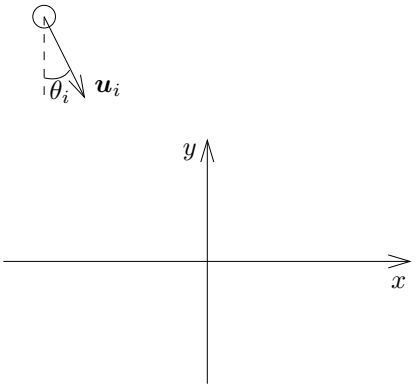

Fig. 2. Parameterize the unit normal vector of the wavefront planes via the direction of arrival.

The equation (7) specifying the spectral support of $r(x, t)$ becomes

$$f_x = \frac{f_t \sin\theta_i}{\lambda}. \qquad (9)$$

Thus, if we suppose that each source signals $s_i(t)$ is bandlimited to $f_t^{\max}$ and the range of DOA's is $|\theta_i| \leq \theta^{\max}$,

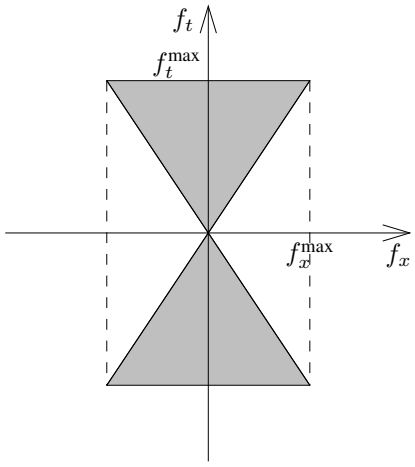

Fig. 3. The spectral support region of the sound field function $r(x, t)$, where $f_x^{\max} = f_t^{\max} \sin\theta^{\max}/\lambda$.

then the spectral support in (9) of $r(x, t)$ will be a bow-tie shaped region as shown in Figure 3.

In the second case, we extend the spatial domain to the whole plane $\boldsymbol{p} = (x, y, 0)^T$. Then the bow-tie shaped region in Figure 3 becomes part of a cone in the 3-D space (f_x, f_y, f_t). If we furthermore restrict the spectral support of the source signal $s_i(t)$ to $f_t^{\min} \leq |f_t| \leq f_t^{\max}$ then the projection of the spectral support region of the function $r(x, y, t)$ onto the plane (f_x, f_y) is shown in Figure 4.

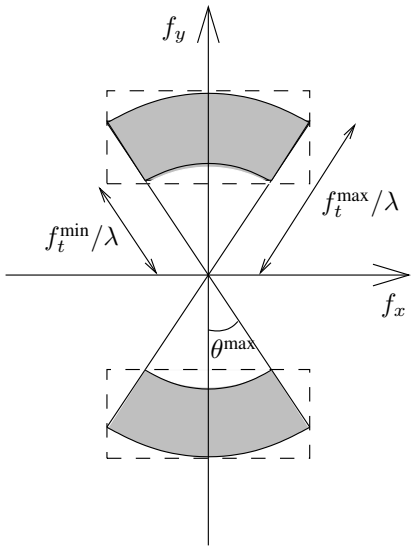

Fig. 4. The spatial spectral support region of the sound field function $r(x, y, t)$ where the spectral support of $s_i(t)$ is $f_t^{\min} \leq |f_t| \leq f_t^{\max}$.

4. SAMPLING AND RECONSTRUCTION

With the spectral analysis from the previous section, we are now ready to solve the minimal sampling and optimal reconstruction problems for sound-based synthesis. Sampling the sound field function $r(\boldsymbol{p}, t)$ on a lattice generates replicated spectra. To avoid aliasing and to ensure perfect reconstruction, the sampling lattice has to be dense enough so that these replicas do not overlap.

For the first case, where the sound field is restricted to the x-axis and the spectral support of $r(x, t)$ is given in Figure 3, optimal sampling is achieved using the quincunx lattice [7, 3, 6]. However, with quincunx sampling, the interpolation function is nonseparable, and thus the reconstruction algorithm needs to use samples on both x and t axes. A suboptimal but much simpler scheme is to use rectangular sampling. In this case, since the interpolation function is separable (a tensor product of two sinc functions), at any instant of time we can synthesize the sound at any point along the x axis using the recorded samples at that time of the nearby microphones. Such synthesis algorithm especially suits real-time applications. Rectangular sampling considers the spectral support of $r(x, t)$ to be bounded by the dashed-line box in Figure 3, which leads to the following sampling requirement

$$\Delta x \leq \frac{1}{2f_x^{\max}} = \frac{\lambda}{2f_t^{\max}\sin\theta^{\max}}. \tag{10}$$

For example, when $f_t^{\max}$ = 5 kHz (typical human speech) and $\theta^{\max} = 30^\circ$, the maximal sampling interval Δx for putting the microphones along the x-axis line is equal to 6.9 cm. In practice, we might need slightly smaller sampling interval to account for the spread of $A_i(\boldsymbol{f_p})$.

For the second case, where sound field is extended to the whole plane $z = 0$ and the spatial spectral support of $r(x, y, t)$ is given in Figure 4, simple geometry leads to the following bounds on the spatial frequencies

$$|f_x| \leq \frac{f_t^{\max}\sin\theta^{\max}}{\lambda} \tag{11}$$

$$\frac{f_t^{\min}\cos\theta^{\max}}{\lambda} \leq |f_y| \leq \frac{f_t^{\max}}{\lambda} \tag{12}$$

These bounds lead to the following requirements on spatial sampling intervals

$$\Delta x \leq \frac{\lambda}{2f_t^{\max}\sin\theta^{\max}} \tag{13}$$

$$\Delta y \leq \frac{\lambda}{2(f_t^{\max} - f_t^{\min}\cos\theta^{\max})} \tag{14}$$

For example, when $f_t^{\min}$ = 4 kHz, $f_t^{\max}$ = 5 kHz, and $\theta^{\max} = 30^\circ$, the maximal spatial sampling intervals are Δx = 6.9 cm and Δy = 11.2 cm.

5. DISCUSSION

We have demonstrated the feasibility of the sound-based synthesis approach in synthesizing the sound at any desired position and time from the recording of a set of microphones. The advantage of this approach is it works for *any number of sound sources*, without knowing their positions and recovering their signals. The resulting synthesizing algorithm is based on simple interpolation and can be done in real-time. The price to pay is a large number of required microphones. Our spectral analysis of the sound field function solves the minimum sampling and optimal reconstruction problems of the sound field in the far-field case. While the required number of microphones might appear to be too large, it might become practical in the near future thanks to the continuing improvement in producing low-cost microphones. Moreover, results from our spectral analysis suggest possible ways of increasing the spatial sampling intervals (i.e. reducing the number of microphones) by adaptively localizing the bandwidth of the source signals or the directions of arrival in time.

Acknowledgement. The author thanks Prof. Douglas Jones for fruitful discussion about sound field reconstruction.

6. REFERENCES

[1] L. McMillan and G. Bishop, "Plenoptic modeling: an image-based rendering system," in *Proc. SIGGRAPH*, 1995, pp. 39–46.

[2] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*, M. Landy and J. A. Movshon, Eds., pp. 3–20. MIT Press, 1991.

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

[4] D. L. Jones, "Four-dimensional sound source recovery from arbitrary acoustic array," in *Proc. IEEE Int. Conf. on Multimedia & Expo*, Baltimore, 2003.

[5] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *Proc. SIGGRAPH*, 2000, pp. 307–318.

[6] C. Zhang and T. Chen, "Spectral analysis for sampling image-based rendering data," *IEEE Trans. Circ. and Syst. for Video Tech.*, 2003, to appear.

[7] E. Dubois, "The sampling and reconstruction of time-varying imagery with application in video systems," *Proc. IEEE*, vol. 73, no. 4, pp. 502–522, April 1985.