# The Sound Wave Ray-Space

Mehrdad Panahpour Tehrani[1], Yasushi Hirano[1],Toshiaki Fujii[2], Shoji Kajita[3], Kazuya Takeda[4], Masayuki Tanimoto[2], and Kenji Mase[3]
[1,3]*Information Technology Center, Nagoya University*
[2]*Department of Electrical Engineering and Computer Science, Nagoya University*
[4]*Center for Integrated Acoustic Information Research, Nagoya University*
[1]*{mehrdad, hirano}@itc.nagoya-u.ac.jp*
[2]*{fujii, tanimoto}@nuee.nagoya-u.ac.jp*
[3,4]*{kajita, kazuya.takeda, mase}@nagoya-u.jp*

## Abstract

*This paper addresses the problem of 3D sound representation without sound source localization and proposes a theory based on the ray-space representation of light rays, which is independent of object's specifications. An array of beam-formed microphone-arrays (MAs), are set and each MA generates a sound-image (SImage) by scanning the viewing range of a camera in the same location. SImage has the same size of an image and contains of blocks of sound wave with duration of one image-frame. Captured SImages with the array of MAs generate the sound wave ray-space. To make a dense SImage ray-space, we propose to use the geometry compensation of corresponding images in the location of each MA. By a dense sound ray-space, any virtual SImage, which corresponds to an arbitrary listening-point, can be generated. The listening-point sound is generated by averaging the sound wave in each pixel or group of pixel of the virtual SImage.*

## 1. Introduction

Sound can be recorded, computed and replayed by directed speakers, using the well-known sound processing methods, efficiently. Several approaches tried to generate arbitrary listening-point generation of sound; however there are few effective model such as Head Related Transfer Function (HRTF) [1] and representation of the sound sources in 3D space to have an efficient processing. Meanwhile, images are rendered by computer graphics algorithms and have become more attractive and more efficient and image synthesis hardware has come to existence, such as Free viewpoint TV (FTV) [2]. The free viewpoint systems should have a correct correspondence of sound and images in an arbitrary viewpoint. Therefore, a representation method of sound sources in 3D space using computer graphics and image processing techniques is necessary. Many representation methods have already been proposed. These methods are categorized into image based rendering (IBR) [3,4], model based rendering (MBR) [5] methods and their combinations [6]. Ray-space representation method is an IBR method and independent of object specifications. Based on the ray-space representation of light rays, we propose a theory to represent the 3D sound wave. Due to different analogy of light rays and sound wave, it is hard to describe the sound wave in ray based representation. The light ray (wavelength approximately 400nm to 700nm) belongs to the partial case and travels on line, whereas the sound waves (wavelength approximately 15m to 0.015nm) are propagated and attenuated. Considering the attenuated wave in space as a sound ray traveled on line gives the opportunity to treat the sound wave as sound ray. According to aforementioned discussion the proposed method is suitable for sound wave due to its similarity to the light ray. Because the sound is also a wave that originates at sources, travels in space and scatters on surfaces as light does. Therefore, if the sound waves are simulated by light, we can represent them by a domain, where their rendering can be done easily. So every sound-source and reflected-sound is interpreted as a light-source and reflected-light, respectively. The intensity of the sound wave is proportional to the intensity of the light ray. By generating a Sound Image (SImage) from the location of the observer, we can describe the sound field of listener's environment. It corresponds to a camera viewing range.

*0-7803-9332-5/05/$20.00 ©2005 IEEE*

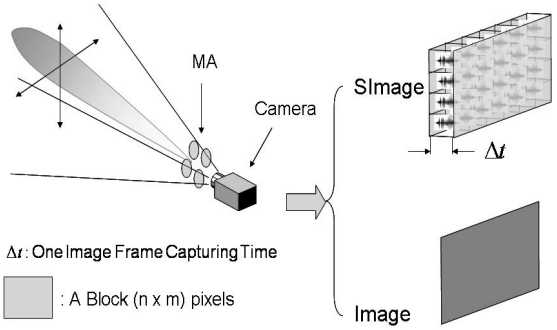


Figure 1. Capturing an Image and an SImage with a camera and a MA

Each SImage is generated by using an array of microphone (MA). Captured SImages generates the sound ray-space and will be dense by geometry compensation [7] of corresponding camera viewing images. Hence, any virtual SImage, which corresponds to an arbitrary listening-point can be generated, and it perfectly corresponds to its virtual viewpoint image. The listening-point sound is generated by averaging the sound wave in each pixel or group of pixel of the virtual SImage.

## 2. SImage

Sound is 1D signal in time domain. In order to process sound signal like image, a 2Dx1D signal of sound wave is generated in space and time domains. We capture a sound wave with duration of a frame for each pixel or group of pixel. After capturing the sound wave for all location of an image, we can generate the SImage, as shown in Figure 1.

In order to generate an SImage, the light equivalent of the sound is determined for every pixel or pixel-group in the SImage. The sound wave for duration of a frame is captured, which is proportional to a pixel location. To generate an SImage, we have to scan the viewing range of an image corresponded to the SImage. The problem of capturing a sound wave for each pixel or group of pixel of a SImage can be solved by beam-forming of a MA. Note that the resolution of the SImage depends on the scanning accuracy or the beam width of the MA. The larger number of microphone in an array, the higher resolution of SImage can be obtained. By placing the microphones in two dimensions, we can generate 2Dx1D SImage.

## 3. Ray-Space of SImages

Ray-space was originally proposed as a common data format for 3D image communication [3]. A similar idea has been proposed in computer graphics field for generating photo-realistic images into

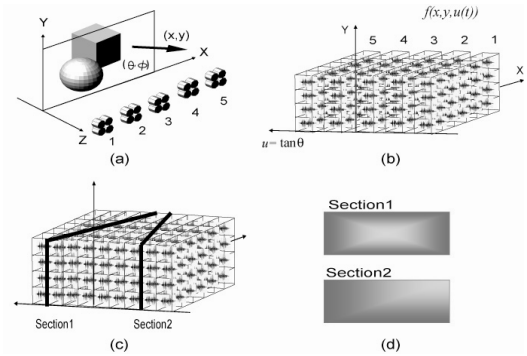


Figure 2. Ray-space method of SImages
(a) Ray space recording (b) Recorded ray-space
(c) SImage generation (d) Generated SImage

computer generated virtual world [8]. It is one of the "image-based-rendering techniques", and is called Lumigraph [4]. It has been widely used to create photo-realistic virtual worlds. Both of them are based on the idea that a view image from an arbitrary viewpoint can be generated from a collection of real view images. Ray space method describes 3D spatial information as the information of the ray, which transmits in the space.

Due to the similarity of light rays and sound wave rays, we proposed the sound wave representation based on ray-space. Figure 2 shows an example of the definition of ray-space. Let $(x,y,z)$ be three space coordinates, and $(\theta,\varphi)$ be the parameters of direction. $x$ and $y$ are the intersection of the sound ray with $XY$-plane. $\theta$ and $\varphi$ are the angles of the sound rays passing through $XY$-plane with $Z$-axis in horizontal and vertical directions, respectively as shown in Figure 2(a). In free space, a sound ray going through the space is uniquely designated by the intersection $(x,y)$ with plane $z = 0$ and its direction $(\theta,\varphi)$. These sound ray parameters construct a 4D space. In fact, this is a 4D subspace of 5D ray-space $(x,y,z,\theta,\varphi)$ of SImages. In this sound ray-parameter, we define a function $f$ whose value corresponds to an intensity of the specified sound ray in a given time ($t$). Thus, all the intensity data of rays can be expressed by equation (1). This ray parameter is called the ray-space.

$$f(x,y,\theta,\varphi)\ ,\ -\pi \le \theta < \pi \text{ and } -\pi/2 \le \varphi < \pi/2 \qquad (1)$$

The sound rays that pass through the specific plane can be captured using a MA, as it shown in section 2. When we set a MA at $(x_0,y_0,z_0)$, the intensity data of sound rays are given by equation (2).

$$f(x,y,z,\theta,\varphi),\quad x = x_0, y = y_0, z = z_0 \qquad (2)$$

For simplicity, we consider only 2D subspace $f(x,\theta)$ of 5D SImage ray-space, which the vertical parallax ($\varphi$) and vertical position ($y$) are neglected. We call this

2D plane of the ray-space data as "Epipolar Plane SImage (EPSI)", which is the cut of ray-space data of Figure 2(b), parallel to *xu*-plane for a given *y*. In another word, EPSI is the intersection of epiploar plane with the corresponded camera plane in stereovision in the location of MA, which is well-known in computer graphic field as Epipolar Plane Image (EPI). Therefore, if there is an array of MAs in which all have a common epipolar plane, the intersection of the epipolar plane with all MA planes can be shown in one SImage, which is EPSI. Let *X*, *Z* be real-space coordinates, and *x*, *u* be the ray-space coordinates as shown in Figure 2(b). The sound rays that pass through a point *(X,Z)* in the real-space form a line in the ray-space is given by equation (3).

$$X = x + uZ, \quad u = \tan\theta \tag{3}$$

Note that symbol *u* is a function of time. Therefore, the ray-space data of SImages is *f(x,y,u(t))*. It gives the most important and interesting characteristic of the 4D ray-space representation that a view SImage corresponds to a cross section of the ray-space data.

Therefore, the SImage acquisition/display process can be considered the recording/extracting process of the ray-space data along the locus. In the acquisition process, we sample the sound ray data in the real space and record them as the ray-space data along the locus as shown in Figure 2(c). In the display process, we cut the ray-space along the locus and extract the section SImage of the ray-space as shown in Figure 2(d). However, in the case of not-dense ray-space data, interpolation is needed. It generates the missing sound rays between two EPSI lines. The EPSI lines are in (*x*) direction for a given *u* as 1D subspace of 5D ray-space data. The interpolation task should be done on 2D SImage ray-space data or EPSI. The interpolation generates an EPSI line between two EPSI lines. In the next section the ray-space interpolation of SImage is explained under EPSI constraint.

# 4. Interpolation Algorithm

## 4.1. Image Interpolation

The ray-space interpolation technique of images is based on the adaptive filtering interpolation [7,8]. This technique was designed to replace a simple linear interpolation technique, which is suitable for input images with viewpoints parallel to the same plane. To make the viewpoint parallel, lens distortion removal and rectification [9] algorithms are needed.

The interpolation starts with the preparation of a set of filters corresponds to the best matched or disparity.

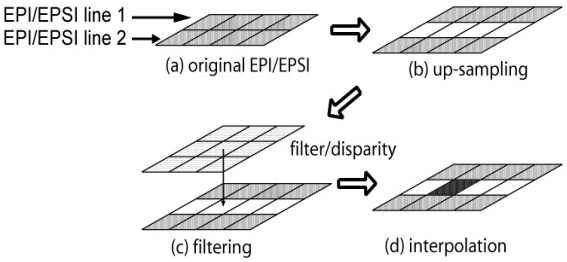

Figure 3. Interpolation under EPI/EPSI constraint

The input images are converted into 2D ray-space data or EPI after performing lens distortion and rectification algorithms. Each EPI is up-sampled. Then the best match direction or filter for interpolation is found. The best filter corresponds to the disparity map. The disparity map gives the distance (in pixels) between two corresponded pixels of two Images. Note that the matching algorithm [10] is searching for the best matched area or the best disparity for a given location of the "pixel to be interpolated" between two EPI lines.

Then, the "pixel to be interpolated" value is calculated by averaging the matched pixel values of two EPI lines, which is on the line passing the "pixel-to be interpolated". Those two pixels in two EPI lines are "the best-corresponded pixel" pair for interpolation in the assigned maximum disparity. The maximum disparity depends on the capturing interval. The main advantage of this method is the approximate geometric models are not required prior any interpolation. Moreover, the scene complexity does not have any effect on interpolation speed.

## 4.2. SImage Interpolation

In previous section the image interpolation is explained. Here, we would like to use the same scheme for SImages. To generate middle EPSI line, we have to look for the corresponded pixels or blocks of sound wave between two EPSI lines. The best match or the best disparity map can be obtained by each of the following ways.

**(a)** Performing block matching scheme on SImages or EPIS lines, as it is done for images or EPI lines, and generating a disparity map using SImages. In this method, after finding the disparity map, interpolation can be performed by averaging the found corresponded pixels or blocks of sound wave. Note that the best match for a given location is found by a minimization solution of sound wave level difference in two SImages or EPSI, for duration of a frame, in a given maximum disparity.

**(b)** Using the disparity map obtained by the images captured by the cameras in the location of MA. In this

method, the intermediate SImage or EPSI is generated by averaging the found corresponded pixel or block of sound waves. The corresponded location in two SImages or EPSI lines are found by the disparity map of the captured images or EPI lines by cameras in the same location of MAs.

### 4.3. Arbitrary Listening-point Generation

After having a virtual SImage (Figure 2(d)), the listening-point sound is generated by averaging the sound wave in each pixel or group of pixel of the virtual SImage.

## 5. Discussion

There are two challenging issues in the proposed theory and algorithm for 3D sound wave representation and rendering methods.

The first one is the sampling rate. Maximum allowable disparity or sampling rate of SImage should be investigated as it has been done in plenoptic sampling theorem [11] for multi-view images. Due to static characteristic of the sound wave for small changes in space, the sampling rate or the interval of MAs to capture SImages is larger than that of images.

The second one is the disparity used for interpolation. In section 4.2 two methods are proposed. Method (a) is easy to use and in the case of just generating free listening-point, the system does not need to be equipped by cameras, however the accuracy of the generation is lower than method (b). On the other hand, method (b) can perform better than method (a). In addition, the free viewpoint/listening-point generation of video/audio can be synchronized. However, the method (b) is more complex. Note that installing a camera and a MA in same location is hardly possible. Nevertheless, close alignment of a camera and a MA will give quite good result due to static character of sound waves in space.

## 6. Conclusion

This paper proposed a theory to represent the 3D sound field using ray-space method, and an algorithm to generate free listening-point. The proposed methods are independent of sound sources location. In addition, the proposed theory can solve the problem of 3D media integration. The proposed methods are currently being developed on a practical system.

In our future research, we will work on sampling theory for capturing SImages, and will perform an efficient integration of 3D audio/video.

## 7. Acknowlegdement

This work has been supported by the SCOPE Fund project, Ministry of Internal Affairs and Communication, Japan (ref. No.: 041306003).

## 8. References

[1] F.L. Wightman and D.J. Kistler, "A Model of HRTFs Based on Principal Component Analysis and Minimum-phase Reconstruction," *Journal of the Acoustical Society of America*, vol. 91(3), pp. 1637- 1647, 1992.

[2] P. Na Bangchang, T. Fujii, M. Tanimoto, “Experimental System of free viewpoint television”, *Proc. SPIE*, Santa Clara, CA, USA, vol. 5006-66, pp. 554-563, Jan 2003.

[3] T. Fujii, T. Kimoto, and M. Tanimoto, “A new flexible acquisition system of ray-space data for arbitrary objects”, *IEEE Trans. On Circuit and Systems for Video Technology*, vol. 10, no. 2, pp. 218-224, March 2000.

[4] M. Levoy, and P. Hanrahan, “Light field rendering”, *ACM SIGGRAPH ’96*, pp. 31-42, 1996.

[5] T. Matsuyama, X. Wu, T. Takai, T. Wada, “Real-time dynamic 3-D object shape reconstruction and high-fidelity texture mapping for 3-D video”, *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 3, 2004.

[6] W. C. Chen, J. Y. Bouguet, M. H. Chu, R. Grzeszczuk, “Light Field Mapping: Efficient Representation and Hardware Rendering of Surface Light Fields”, *ACM Trans. on Graphics*, vol. 21, no. 3, pp. 447-456, 2002.

[7] M. Droese, T. Fujii, M. Tanimoto, “Ray-Space Interpolation Constraining Smooth Disparities Based on Loopy Belief Propagation”, *Proc. of 11th International Workshop on Systems, Signals and Image Processing ambient multimedia (IWSSIP)*, Poland, pp. 247-250, 2004.

[8] R. Szeliski, “Video mosaics for virtual environments”, *IEEE Computer Graphic Application*, vol. 16, pp. 22-30, March 1996.

[9] F. Fuseillo, E. Trucco, and A. Verri, “A compact algorithm for rectification of stereo pairs”, *Machine Vision and Applications*, vol. 12, pp. 16-22, 2000

[10] M. P. Tehrani, T. Fujii, M. Tanimoto, “Offset Block Matching of Multi-view Images for Ray-Space Interpolation”, T*he journal of the institute of Image information and Television Engineers (ITE)*, vol. 58, no. 4, pp. 86-94, April 2004.

[11] J.X. Chai, S.C. Chan, H.Y. Shum, X. Tong, “Plenoptic Sampling”, *ACM SIGGRAPH*, pp. 307-318, 2000.