Proceedings of the Institute of Acoustics

SOUND IMAGE CONTROL FOR INTERNET

Michiaki UCHIYAMA and Miklo TOHYAMA

Kogakuin University, Department of Information Science and Engineering,
2665-1, Nakano-machi, Hachioji-shi, Tokyo, 192-0015 JAPAN

1. Introduction

In this paper, we describe a sound image projection system (SIPS) and a blind source separation technique for SIPS that is related to 3D acoustic space collaboration through a computer network. As a reproduction system, SIPS requires head related transfer functions (HRTFs) if it is to successfully create 3D acoustic spaces or 3D acoustic collaboration spaces like a virtual mall. When a moving source is reproduced, the system needs all HRTFs of all directions; however, it seems impractical to measure and save a large set of HRTFs. There has been some research concerning reducing the size of data base by preserving only a set of each certain discrete-angle HRTFs and taking a linear interpolation for any angle in the middle. However, these papers mainly focus on the interpolation technique itself, and few papers describe the differences in the accuracy of each angle and the accuracy of the interpolated shade-side HRTFs. Therefore, we consider the angular characteristics of the accuracy by linear interpolation; in particular, we discuss the accuracy of shade-side HRTFs and describe a method for generating binaural information based on the cross ambiguity function model (CAF).

Sound zoom-in/out and distance hearing control are quite important techniques for acoustic space collaboration. Generally, the sound zoom-in/out and distance hearing have been treated as a reproduction problem, on the assumption that independent sources are used. The use of independent sources is required in order to localize two or more sound sources effectively. As one example, however, take the creation of 3D sound contents by using arbitrary sources from other contents mixed with several sources. It is no longer a reproduction problem ; we should recognize it as a receiving problem. In this article, we suggest that the blind source separation (BSS) technique works effectively for SIPS, and we investigate a selective receiving method by using principal component analysis (PCA) with 2 point microphones.

2. Sound Image Projection System (SIPS)

A 3D acoustic collaboration space is a linked acoustic virtual reality connecting two distant users. Figure 1 shows the concept of the collaboration space. The basic theory of SIPS for two loudspeakers uses a set of HRTFs. As Fig. 2 shows, the basic technique uses two filters X and Y, to create a phantom source at any location. These filters satisfy the following simultaneous equations:

$$\begin{cases} Z_L = XH_L + YG_L \\ Z_R = XH_R + YG_R \end{cases}, \tag{1}$$

where Z_L and Z_R denote the transfer function between the phantom source and left (right) outer ear entrance at the listener's head position, and G and H are HRTFs for the left and right channel loudspeakers respectively. X and Y in Fig. 2 are filters for the SIPS. These filters are obtained by using of following solutions of the simultaneous equations (1):

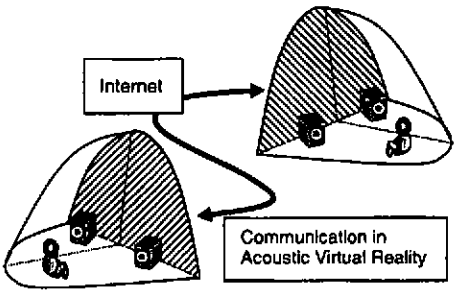


Figure 1 3D acoustic space collaboration through the network

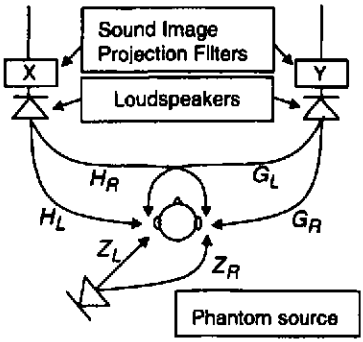


Figure 2 Two-channel sound image projection system (SIPS)

$$\begin{cases} X = \dfrac{Z_L G_R - Z_R G_L}{H_L G_R - G_L H_R} \equiv \dfrac{X_N}{D} = \dfrac{X_N}{D_{MIN} D_{AP}} \\ Y = \dfrac{Z_R H_L - Z_L H_R}{H_L G_R - G_L H_R} \equiv \dfrac{Y_N}{D} = \dfrac{Y_N}{D_{MIN} D_{AP}} \end{cases} \quad (2)$$

where $D_{MIN} D_{AP}$ denotes the product of minimum-phase and all-pass components in the denominator. However, these inverse filters become generally non-causal because a non-minimum phase component is included in the denominator. We then define $\hat{X}$ and $\hat{Y}$ for the minimum-phase components of the denominator as follows:

$$\begin{aligned} \hat{X} &= \frac{X_N}{D_{MIN}} \\ \hat{Y} &= \frac{Y_N}{D_{MIN}} \end{aligned} \quad (3)$$

When we substitute X, Y in equation (1) by $\hat{X}, \hat{Y}$ in (3), the result is:

$$\begin{aligned} \hat{X} H_L + \hat{Y} H_L &= \frac{X_N}{D_{MIN}} H_L + \frac{Y_N}{D_{MIN}} G_L = Z_L D_{AP} \\ \hat{X} H_R + \hat{Y} H_R &= \frac{X_N}{D_{MIN}} H_R + \frac{Y_N}{D_{MIN}} G_R = Z_R D_{AP} . \end{aligned} \quad (4)$$

Because the D_{AP} is a transfer function that only has phase delay, the power spectra of the phantom source transfer functions (Z_L, Z_R) are completely preserved. From the above equations, binaural information of (4), which works quite well, is the same as (1). Therefore, the minimum-phase inverse filters given in equation (3) as the basis to realize SIPS.

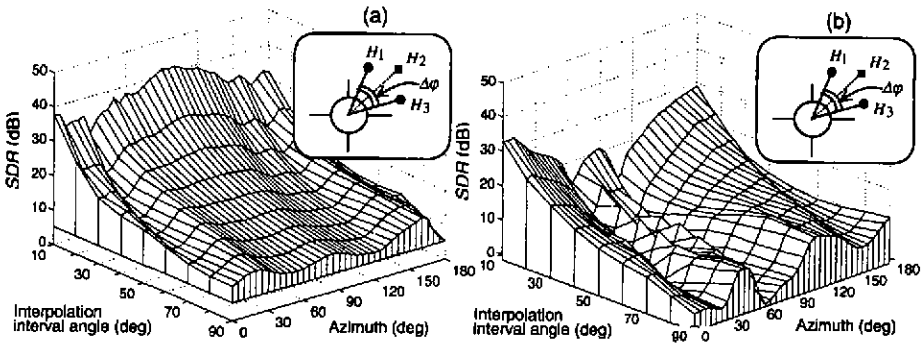


Figure 3 *SDR* of linear interpolated HRTFs (a): sunny-side HRTFs (b): shade-side HRTFs

3. Linear Interpolation of HRTFs

Because the sound pressure signal that reaches the human ear is evaluated in decibels, we use linear interpolation of logarithmic magnitude and the phase characteristics of certain discrete angle HRTFs for any angle in between. When we define that H_1 and H_3 is a set of HRTFs that has an opening angle $\Delta\varphi$, and H_2 is an HRTF in between; then, H_2 can be estimated as follows:

$$\ln\left|\hat{H}_2\right| = \alpha \ln|H_1| + (1-\alpha)\ln|H_3|. \qquad (5)$$

We evaluated the accuracy of estimated HRTF by:

$$SDR = 10\log_{10}\frac{\sum_{n=1}^{N}\{h_2(n)\}^2}{\sum_{n=1}^{N}\{h_2(n)-\hat{h}_2(n)\}^2} \qquad (6)$$

$$D_f\left(|H_2|\right) = \sqrt{\frac{1}{N}\sum_{N}\left(20\log_{10}\frac{|H_2(\omega_N)|}{|\hat{H}_2(\omega_N)|}\right)^2}, \qquad (7)$$

where Eq. (6) is defined in the time domain and Eq. (7) is for the frequency domain, respectively.

Figure 3 shows the signal to deviation ratio (*SDR*), and Fig. 4 depicts the spectrum distortion (D_f). Figure 3(a) and Fig. 4(a) correspond to linear interpolation on sunny-side HRTFs. Figure. 3(b) and Fig. 4 (b) correspond to shade-side HRTFs ($\alpha = 0.5$). We can see from Fig. 3 that the accuracy of the linear interpolation on sunny-side HRTFs is kept high in each direction even if the interpolation interval becomes wide. However, the accuracy on shade-side HRTFs decays rapidly as interpolation interval becomes wide. In particular, the accuracy corresponding to the directions from 30° to 120° is significantly reduced even though the interpolation interval is narrow. When the shade-side HRTFs decays the accuracy much, the binaural information in the direction of the ear axis becomes affected.

Figure 4 is to discuss about spectrum information which is important for front-back perception. We can see from Fig. 4 (a), that the spectrum distortion (D_f) is low from 30° to 120° of the sunny-side;

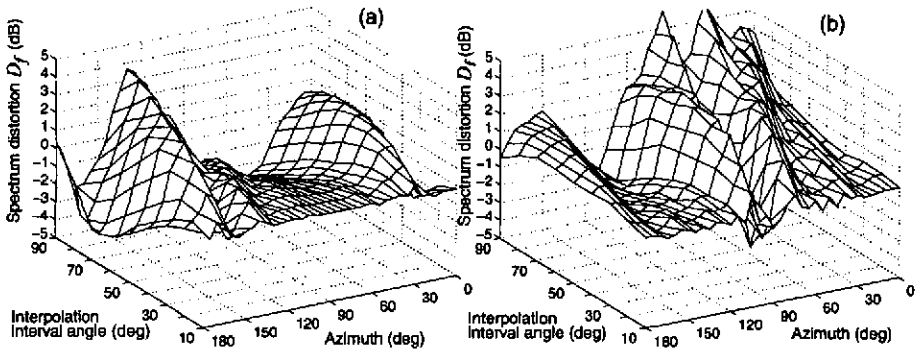


Figure 4. Spectrum distortion (D_f) of linear interpolated HRTFs (a): sunny-side HRTFs (b): shade-side HRTFs

however, in the regions 0°-30° and 120°-180°, D_f significantly increases as the interpolation interval becomes wide. On the other hand, the shade-side HRTFs have a noticeable spectrum distortion over a wide range of directions. Because the level and phase difference information is important for perception of left-right direction, we can judge that the interpolation interval in the directions from 60° to 120° can be comparatively wide. On the other hand, because it is necessary to transmit the spectrum information with few distortions for front-back perception, the interpolation interval should be narrow. However, the accuracy of shade-side HRTFs by *SDR* and D_f decays significantly for front-back perception. The accuracy of the shade-side HRTFs must therefore be improved.

4. Generation of binaural information based on cross ambiguity function

As we described in the preceding section, the interpolation accuracy of the shade-side HRTF was not as high as it was desired. Therefore, we will investigate the generation of shade-side HRTFs based on cross ambiguity function (CAF), and we will attempt to improve the level of the binaural information. Moreover, the amount of necessary HRTFs for the data base is expected to be reduced greatly if the accuracy of shade-side HRTFs created from sunny-side HRTFs by using the CAF model becomes sufficiently high.

Figure 5 shows a model based on CAF. As a time delay and an amplitude modulation are caused between both sides of HRTF, we define the signals as follows:

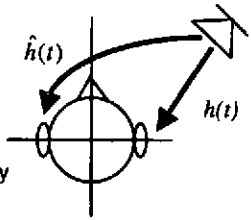
$$\begin{aligned} h_R(t) &= h(t) \\ h_L(t) &= \rho_{RL} h\left((t-\tau_{RL})/\alpha_{RL}\right) \end{aligned} \quad , \tag{8}$$

where τ_{RL}, α_{RL}, ρ_{RL} denote, respectively, interaural time delay, time scaling of waveform, interaural level difference, and $h(t)$ denotes an impulse response. These model is similar to dichotic hearing. We define a cross ambiguity function as follows;

$$A_{h_R h_L}(\tau_{RL}, \alpha_{RL}, \rho_{RL}) = \rho_{RL} \int_{-\infty}^{+\infty} h_L h_R^* \left(\frac{t-\tau_{RL}}{\alpha_{RL}} \right) dt \quad . \tag{9}$$

$$h_R(t) = h(t)$$

$$\hat{h}_L(t) = r_{RL} h\left(\frac{t}{\alpha_{RL}} - t_{RL}\right)$$



τ_{RL} : Interaural time delay
α_{RL} : Time-scaling factor
ρ_{RL} : Magnitude-scaling factor

Figure 5. A model of cross ambiguity function

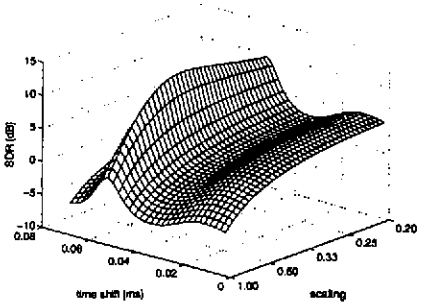


Figure 6. An SDR accuracy of shade-side ear's HRTF based on cross ambiguity function

There are some methods for achieving time axis scaling related to α_{RL}, for example, the method using multirate sampling and chirp-z transform. Our methods use high sampling effect by zero padding on frequency axis and low-pass filters. Figure 6 shows the change in accuracy, as evaluated by *SDR*, of shade-side HRTF at 150° by using CAF. The estimated *SDR* that is maximized by CAF is about 13 dB. When the linear interpolation interval equals 50°, generated HRTF by CAF is almost as accurate as the linear interpolation. Figure 7 shows the true and estimated impulse responses of HRTFs at 30°. The trace in Fig. 7(b) and the solid line in Fig. 7(c) show shade-side impulse response generated from sunny-side HRTF to be maximized by a CAF. Fig. 7(c) is an enlargement of the true and the estimated impulse responses. We can see in this curve the shade-side HRTF can almost be estimated from a sunny-side HRTF, and we confirm that it is possible to generate a binaural information from 1-ch HRTFs.

5. Selective receiving technique

The use of several independent sources is required in order to localize a phantom source effectively. We should recognize that this is not a reproduction problem, and it is instead a reception problem. We suggest that the blind source separation (BSS) technique works effectively for SIPS, and we use the principal component analysis (PCA) frame by frame for the received signals from two microphones. Moreover, we investigate a possibility of selective receiving with on-line processing.

We assume that $\mathbf{X}$ and $\mathbf{Y}$ denote the vectors corresponding to each of the two signals received by the respective microphones. For these vectors, if the mean value is subtracted, then the data matrix has the form:

$$B = \begin{pmatrix} \mathbf{X}^T \\ \mathbf{Y}^T \end{pmatrix}. \qquad (10)$$

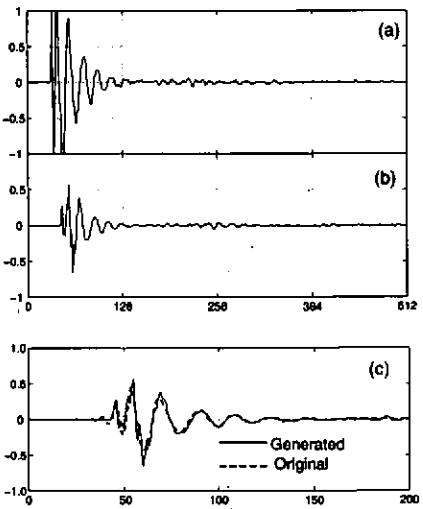


Figure 7. Comparison of head-related-impulse-response (a): sunny-side (b): Shade-side (c): generated by CAF

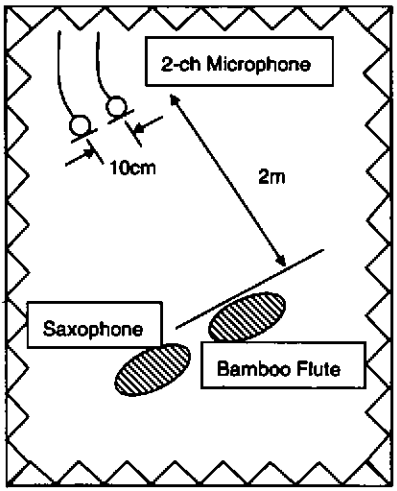


Figure 8. Selective receiving method using two point microphones

The covariance matrix becomes a square matrix as follows:

$$S = \frac{1}{n-1} BB^{\mathrm{T}}, \qquad (11)$$

where n denotes the data length. When we project a column vector of the covariance matrix on the eigenvectors, we can obtain a coordinate transform that will enhance the desired signal. Figure 9 shows a scatter diagram; Figure 9(b) is a result of coordinate transformation by PCA. However, if the cross correlation between the received signals becomes low, the difference between the two eigenvalues becomes small, and the louder signal (which must be projected on the principal axis) leaks out to another orthogonal axis. As a result, PCA has the characteristic that the signal separation does not work well.

Figure 8 shows an example of recording environment using two microphones. First of all, we investigated a possibility of BSS using PCA under the ideal condition that the time delay of the two channel signals was known beforehand. Figure 10 shows a preliminary result of the extraction of a voice buried in white noise. If the time delay is completely estimated, then PCA works well, as shown in fig. 10(c) and (d), and there is no difference between frame processing and batch processing. These results suggest the possibility of on-line processing by using PCA with each frame.

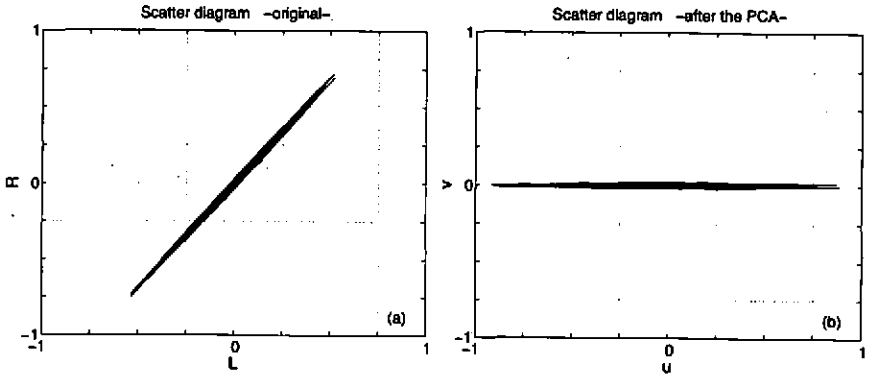


Figure 9. Scatter diagram (a): two-channel recorded signals (b): after the PCA processing

Next, we simulated an ideal case for a practical situation where that the time delay cannot be completely estimated from a recorded original signal. In order to accurately estimate the time delay we examined the cross correlation function using a high sampling based on a zero padding on the frequency axis. The purpose was the improvement of the signal separation extent by estimating the true time delay hidden in the intervals of discrete signals. Then we created two receiving signals which were mixtures of a bamboo flute buried in saxophone with the accurately estimated time delay. The PCA results can be seen in Fig. 11. The source separations are lightly deteriorated but we were able to improve the results by a post processing such as subtraction between two separated signals.

Finally we attempted the BSS under a synchronous recording, which corresponds to a practical situation. Here, it was difficult to get a complete estimate of the time delay. Consequently we confirmed that the results deteriorated if the time delay could not be estimated completely. The fact that the source cannot be considered as a point source is thought to make the problem more complicated than that of the movement of the source.

5. Conclusion

In this paper, we described a sound image projection system and an application of blind source separation technology for 3D acoustic space collaboration through a computer network. We proposed a method of generating a binaural information based on the cross ambiguity function, which improves the accuracy of shade-side HRTFs and reduces the size of the data base. At present, although the accuracy of this method is almost the same as the results of linear interpolation, we can foresee a reduction in the size of the data base. We also investigated the two-point receiving PCA which can be defined to BSS for SIPS. We confirmed that our PCA method can separate two sources clearly when the time delay was completely estimated. In future studies, we must improve the estimation accuracy of shade-side HRTFs, and we should do preprocessing in which the time delay is completely presumed.

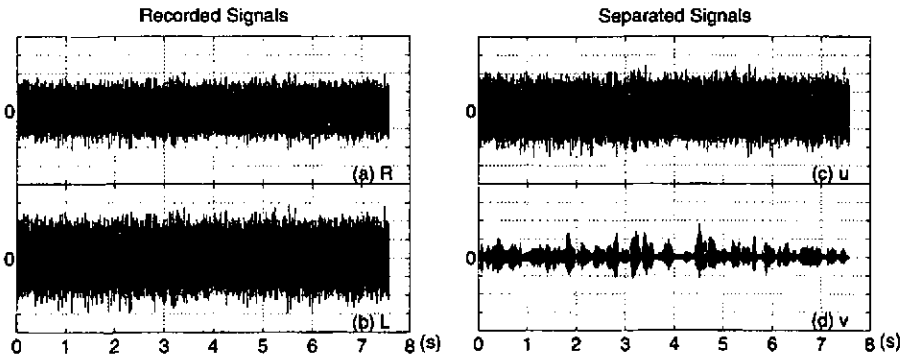


Figure 10. Temporal waveform at an ideal condition, (a),(b): original, (c),(d): after the PCA processing

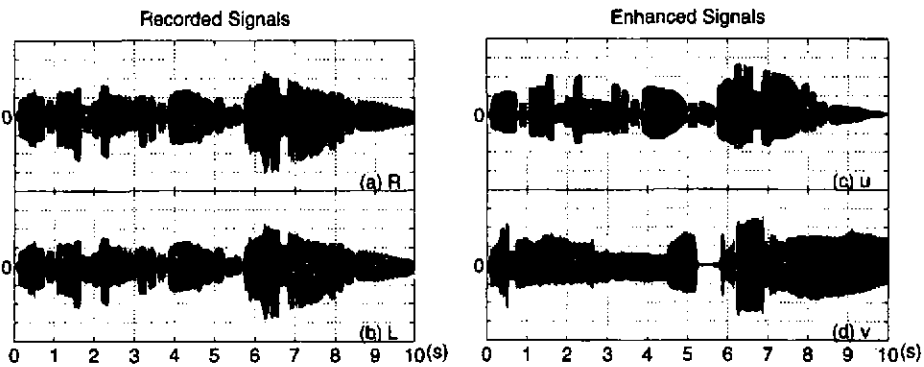


Figure 11. Temporal waveform under practical conditions, (a),(b): original, (c),(d): after the PCA processing

References

[1] M.Tohyama and T.Koike, "Fundamentals of acoustic signal processing", Academic Press, London, UK, 1998

[2] H.Moller et al., "Binaural technique: Do we need individual recording? ", J. Audio Eng. Soc., vol. 44, No. 6, 1996 June, pp451-469.

[3] K.W.Lo and B.G.Ferguson, "Method for computing the passive wideband cross ambiguity function", Proceedings of the 5th ICSV, vol. 4, pp.2213-2220.

[4] X.R.Cao and R.W.Liu, "General approach to blind source separation", IEEE Trans. Signal Processing, vol. 44, 1996 March, pp562-570.