ForumAcusticum
2 0 0 5

Sound Field Auralization System in Free Listening Positions Using Wave Field Synthesis and Head Related Transfer Functions

Toshiyuki Kimura, Wataru Mizuno
Graduate School of Information Science, Nagoya University, 464-8601 Nagoya, Chikusa-ku, Furo-cho, Japan,
e-mail: **kimura@sp.m.is.nagoya-u.ac.jp**,**mizuno@sp.m.is.nagoya-u.ac.jp**

Takanori Nishino
Center for Information Media Studies, Nagoya University, 464-8601 Nagoya, Chikusa-ku, Furo-cho, Japan,
e-mail: **nishino@media.nagoya-u.ac.jp**

Katsunobu Itou, Kazuya Takeda
Graduate School of Information Science, Nagoya University, 464-8601 Nagoya, Chikusa-ku, Furo-cho, Japan,
e-mail: **itou@is.nagoya-u.ac.jp**,**kazuya.takeda@nagoya-u.jp**

The free viewpoint television (FTV) system, which consists of a lot of cameras, is developed in order to visualize the three-dimensional object in free viewpoints. In this study, the free listening-point auralization system is proposed in order to append the sound information to FTV system and appreciate a musical player in free positions. The proposed method employs wave field synthesis and head related transfer functions. Microphones are arranged at the same position of cameras placed around actual sound sources and channel signals are recorded by microphones. Image source signals are estimated from channel signals according to the principle that the wave field of actual sound sources is synthesized by image source signals. Binaural signals of listening positions are calculated by convolving head related transfer functions to image source signals. Since the proposed method doesn't require the information of actual sound sources (e.g. position, number), this method can be applied to the case of moving sound sources. The proposed method was evaluated in two-dimensional plane. It was considered based on both objective and subjective results that, for the proposed method, an image sources increase may cause more accurate sound field reproduction.

1 Introduction

Visual display techniques and sound field auralization techniques are being developed to enable the construction of more realistic communication systems. In particular, the Free Viewpoint Television (FTV) system has been developed [1] as the "ultimate 3D TV" and proposed to the Motion Pictures Experts Group [2]. The configuration of the FTV system is shown in Figure 1. Images

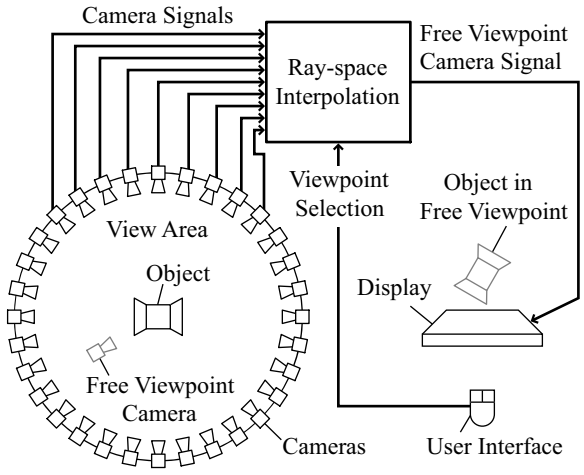

Figure 1: Free Viewpoint Television (FTV) system

of an object are captured by cameras placed around the object. The user selects a viewpoint using an interface, the signal from the camera at that viewpoint is synthesized based on ray-space interpolation, and the image is displayed. As a result, the user can view the object freely from virtually any viewpoint.

The aim of this study is to develop a more realistic television system that enables the user to enjoy a musical performance from virtually any position by adding sound information to FTV system. This was done by introducing a sound field auralization system in which the listening position varies with the user's selected position. This auralization system uses wave field synthesis [3] and head-related transfer functions (HRTFs) [4]. This paper describes the auralization system and the experiments conducted to evaluate its performance and presents some of the results.

2 Sound Field Auralization System

2.1 Overview

The configuration of the proposed system is shown in Figure 2. Sound signals are recorded by microphones places at the same positions as the cameras. The image source signals are estimated by convolving these signals

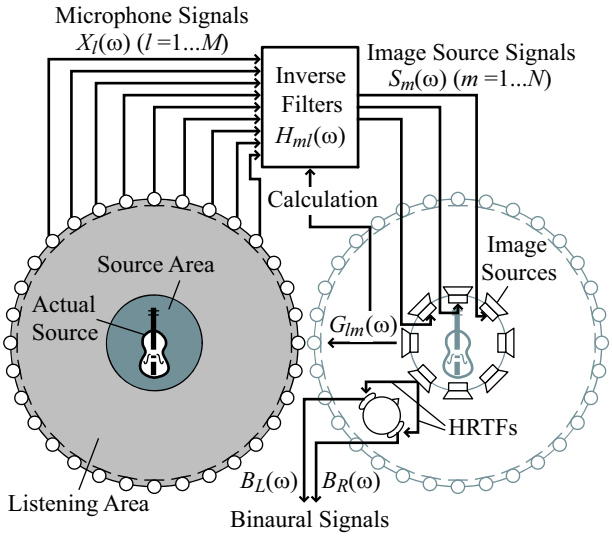

Figure 2: Configuration of the proposed system

with the inverse transfer functions calculated from the room transfer functions between the image sources and microphones. This is done based on the assumption that the microphones record the wave front synthesized by the image sources, which are placed at the boundary of the source area, based on Huygens principle. Binaural signals are then synthesized by convolving the image source signals with the HRTFs between the image sources and the listening position selected by the user. As a result, the user can freely enjoy the sound from virtually any listening position. Since only the HRTFs must be varied to change the listening position, information about the actual sound sources (e.g. position, number) is not required. The proposed system can thus be used when the sound sources are moving, such as in a theatrical performance.

2.2 Estimation of Image Source Signals

Given the assumption described in Section 2.1, the microphone signals are synthesized by convolving the image source signals with the room transfer functions using

$$X_l(\omega) = \sum_{k=1}^{N} G_{lk}(\omega) S_k(\omega) \quad (l = 1...M), \qquad (1)$$

where $X_l(\omega)$ is the lth microphone signal, $S_k(\omega)$ is the kth image source signal, $G_{lk}(\omega)$ is the room transfer function from the kth image source to the lth microphone, and M and N are the numbers of microphones and image sources. The image source signals are estimated from the microphone signals using

$$S'_m(\omega) = \sum_{l=1}^{M} H_{ml}(\omega) X_l(\omega) \quad (m = 1...N), \qquad (2)$$

where $S'_m(\omega)$ is the mth estimated image source signal and $H_{ml}(\omega)$ is the inverse transfer function from the lth microphone to the mth image source.

The inverse transfer functions are calculated as follows. Substituting Equation (1) into Equation (2), we get

$$\sum_{l=1}^{M} H_{ml}(\omega) G_{lk}(\omega) = \begin{cases} S'_m(\omega)/S_m(\omega) & (m=k) \\ 0 & (m \neq k) \end{cases}. \qquad (3)$$

If the image source signals are estimated, $S'_m(\omega)/S_m(\omega)$ should be 1. However, the $H_{ml}(\omega)$ does not satisfy the causality when $S'_m(\omega)/S_m(\omega)$=1 since $G_{lk}(\omega)$ has an initial delay. Therefore, the delay for n_0 samples is set to $S'_m(\omega)/S_m(\omega)$ in order to calculate the $H_{ml}(\omega)$ satisfying the causality:

$$S'_m(\omega)/S_m(\omega) = e^{-j\omega \frac{n_0}{F_s}} \quad (m = 1...N), \qquad (4)$$

where F_s is the sampling frequency. Thus, from Equation (3), a matrix equation is obtained:

$$\mathbf{G}(\omega)\mathbf{H}(\omega) = \mathbf{D}(\omega), \qquad (5)$$

where

$$\mathbf{G}(\omega) = \begin{pmatrix} G_{11}(\omega) & \dots & G_{M1}(\omega) \\ \vdots & \ddots & \vdots \\ G_{1N}(\omega) & \dots & G_{MN}(\omega) \end{pmatrix} \qquad (6)$$

$$\mathbf{H}(\omega) = \begin{pmatrix} H_{11}(\omega) & \dots & H_{N1}(\omega) \\ \vdots & \ddots & \vdots \\ H_{1M}(\omega) & \dots & H_{NM}(\omega) \end{pmatrix} \qquad (7)$$

$$\mathbf{D}(\omega) = \begin{pmatrix} e^{-j\omega \frac{n_0}{F_s}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & e^{-j\omega \frac{n_0}{F_s}} \end{pmatrix}. \qquad (8)$$

Inverse transfer functions are then calculated from Equation (5):

$$\mathbf{H}(\omega) = \mathbf{G}^{+}(\omega)\mathbf{D}(\omega), \qquad (9)$$

where $\mathbf{G}^{+}(\omega)$ is the Moore-Penrose pseudo inverse matrix of $\mathbf{G}(\omega)$. This matrix can be calculated using singular value decomposition [5].

2.3 Synthesis of Binaural Signals

The binaural signals are synthesized as shown in Figure 3. Binaural signals $B_L(\omega)$ and $B_R(\omega)$ are synthesized from image source signals $S_m(\omega)$ using

$$B_L(\omega) = \sum_{m=1}^{N} q(\Delta_m) I_L(d_m, \phi_m, \omega) S_m(\omega) \qquad (10)$$

$$B_R(\omega) = \sum_{m=1}^{N} q(\Delta_m) I_R(d_m, \phi_m, \omega) S_m(\omega), \qquad (11)$$

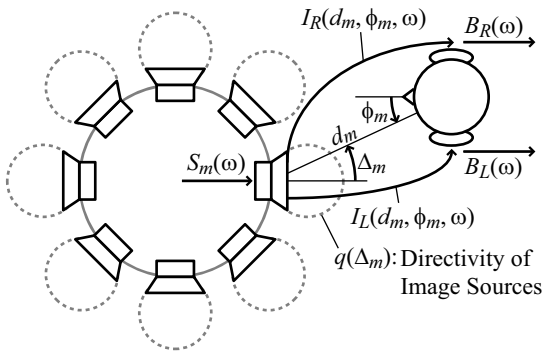

Figure 3: Synthesis of binaural signals

Table 1: Experimental conditions

Number of image sources	12, 18, 24, 36, 48
Number of microphones	N, N×2, N×3, N×4
Sampling frequency (F_s)	32 kHz
Sound velocity (c)	340 m/s

where d_m is the distance between the mth image source and the listening position, ϕ_m is the azimuth angle of the mth image source at the listening position, and $I_L(d_m, \phi_m, \omega)$ and $I_R(d_m, \phi_m, \omega)$ are the HRTFs for the left and right ears at distance d_m and azimuth angle ϕ_m for the sound source. The directivity function of the mth image source, $q(\Delta_m)$, is defined based on Δ_m (the azimuth angle of the listening position for the mth image source):

$$q(\Delta_m) = \begin{cases} \cos\Delta_m & (|\Delta_m| \leq 90^\circ) \\ 0 & (|\Delta_m| > 90^\circ) \end{cases}. \quad (12)$$

The synthesized binaural signals are presented to the user over headphones.

3 Evaluation of Performance

3.1 Conditions

The performance of the proposed system was evaluated experimentally. The signals and transfer functions were simulated using a PC based on the assumption that actual sources, microphones, image sources, and listeners were arranged in free space, as shown in Figure 4. One actual source was placed 0.6 m from the center at a 45° azimuth angle. The microphones and image sources were equally spaced on circles with radii of 2.1 and 0.8 m, respectively. The other experimental conditions are shown in Table 1. The $x_l(n)$ (lth microphone signal) was calculated using

$$x_l(n) = \frac{1}{d_{l0}} s_0[n - \text{round}\Big(\frac{d_{l0}F_s}{c}\Big)] \quad (l = 1...M), \quad (13)$$

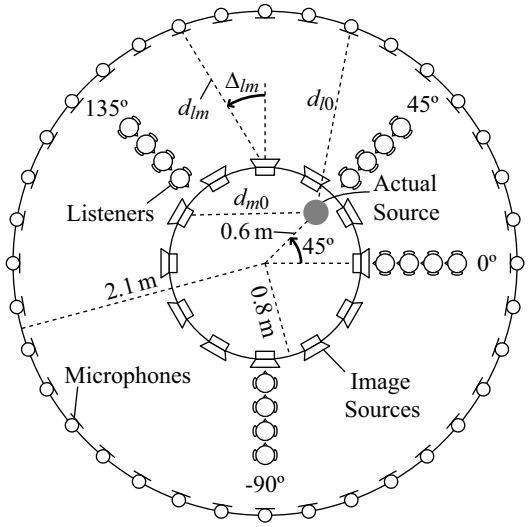

Figure 4: Experimental arrangement

Table 2: Calculation conditions

FFT frame length	2048 samples
Calculated bandwidth	250 Hz – 13333 Hz
Delay sample (n_0)	512 samples
ITF length	1024 samples

where $s_0(n)$ is the actual source signal and d_{l0} is the distance between the actual source and the lth microphone. A piano sound (sampling frequency of 32 kHz; duration of 5 s) was used as the actual source signal.

The $g_{lm}(n)$ (room transfer function from the mth image source to the lth microphone) was calculated using

$$g_{lm}(n) = \frac{q(\Delta_{lm})}{d_{lm}} \delta[n - \text{round}\Big(\frac{d_{lm}F_s}{c}\Big)] \quad (m = 1...N, l = 1...M), \quad (14)$$

where $\delta(n)$ is Dirac's delta function and d_{lm} is the distance between the mth image source and the lth microphone. The $q(\Delta_{lm})$ was defined as shown in Equation (12), where Δ_{lm} is the azimuth angle of the lth microphone in the mth image source. Inverse transfer functions $h_{ml}(n)$ were calculated using Equation (9). The calculation conditions are shown in Table 2. Image source signals $s_m(n)$ were estimated by convolving microphone signals $x_l(n)$ with inverse transfer functions $h_{ml}(n)$ using Equation (2).

3.2 Objective Evaluation

If the image source signals are correctly estimated, the wave front synthesized by the image sources should be the same as that synthesized by the actual sources. The estimation accuracy of the image source signals was thus

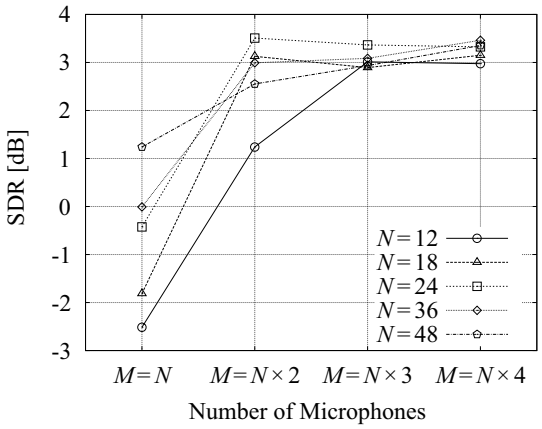

Figure 5: SDR results

evaluated using the signal to deviation ratio (SDR):

$$\mathrm{SDR[dB]} = 10\log_{10}\frac{\sum_{m=1}^{N}\sum_{n}\{s'_m(n-n_0)\}^2}{\sum_{m=1}^{N}\sum_{n}\{s'_m(n-n_0)-s_m(n)\}^2}, \quad (15)$$

where $s'_m(n)$ is the mth reference image source signal calculated using

$$s'_m(n) = \frac{1}{d_{m0}}s_0[n - \mathrm{round}\Big(\frac{d_{m0}F_s}{c}\Big)], \quad (16)$$

where d_{m0} is the distance between the actual source and the mth image source.

The SDR for various numbers of microphones is shown in Figure 5. It increased with the number up to three times the number of image sources and then leveled off. It remained constant even when the number of image sources increased when the number of microphones was more than three times the number of image sources. This means that the estimation accuracy of the image source signals is independent of the number of image sources if the number of microphones is more than three times the number of image sources. The number of microphones was thus set to four times the number of image sources.

3.3 Measurement of HRTFs

The listening positions were at 1, 1.2, 1.4, and 1.6 m with azimuth angles of 0, 45, 135, and -90° from the center, as shown in Figure 4. At these positions, the distance between the image sources and listening positions was very short (0.2–0.8 m). The distance between the sound source and the listener is more than 1 m in conventional HRTF databases [6, 7, 8, 9, 10, 11, 12]. Since HRTFs depend on the distance when the distance is less than 1 m, errors may arise if the HRTFs are modified to a distance

Figure 6: Piezoelectric dodecahedral loudspeaker

Table 3: Measurement conditions

Room temperature	24.0 °C
Background noise level	13.8 dB(A)
Sound pressure level	69.0 dB(A)
Sampling frequency	48 kHz
TSP signal length	32768 samples
HRTF length	512 samples

of less than 1 m. Thus, the HRTFs for the close distances were measured.

A piezoelectric dodecahedral loudspeaker [13] (Figure 6) was used as the sound source because it can be regarded as a point source even if the distance is 0.2 m. The sound source and head and torso simulator (HATS) (B&K type 4128) were placed in a measurement room in which the reverberation time is very short, and condenser microphones (Sony ECM-77B) were attached to the HATS ears. The height of the sound source was the same as that of the ears. A time stretched pulse (TSP) signal [14] was used as the measurement signal. The measurement conditions are shown in Table 3. The sound pressure level was the value at the position of 1 m from the loudspeaker. The HRTFs were measured at distances of 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, and 1 m and azimuth angles of -179, -178, ..., -1, 0, 1, ..., 179, and 180°. The measured HRTFs were saved in a database and denoted as $i_L(d,\phi,n)$ and $i_R(d,\phi,n)$. Based on the calculation of d_m and ϕ_m described in Section 2.3, the HRTFs were modified:

$$i_L(d_m,\phi_m,n) = \frac{d_p}{d_m}i_L(d_p,\phi_p,n) \quad (17)$$

$$i_R(d_m,\phi_m,n) = \frac{d_p}{d_m}i_R(d_p,\phi_p,n), \quad (18)$$

where d_p and ϕ_p are the closest measurement position to d_m and ϕ_m, and $i_L(d_p,\phi_p,n)$ and $i_R(d_p,\phi_p,n)$ are the HRTFs measured at d_p and ϕ_p. Binaural signals $b_L(n)$ and $b_R(n)$ were synthesized from $i_L(d_m,\phi_m,n)$ and $i_R(d_m,\phi_m,n)$ as described in Section 2.3.

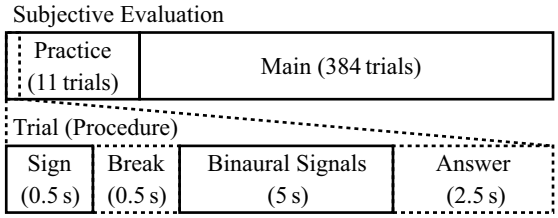

Figure 7: Experimental design

Table 4: Trial conditions

	Factor	Level
Practice (11)	= 1 conditions × 11 directions	Actual Source -75, -60, ..., 60, and 75°
Main (384)	= 6 conditions × 4 distances × 4 azimuths × 4 repetitions	N=12, 18, 24, 36, 48, and Actual Source 1.0, 1.2, 1.4, and 1.6 m 0, 45, 135, and -90°

3.4 Subjective Evaluation

If the binaural signals synthesized using the image source signals are the same as those synthesized using the actual source signals, the directional perception of the image sources should be the same as that of the actual sources. The accuracy of the directional perception was thus subjectively evaluated in a localization test.

The localization test was performed in a soundproof room. Each subject sat in the room and listened to the binaural signals through headphones (Audio-Technica ATH-A1000). The background noise level was 18.5 dB(A), and the sound pressure level of the headphones was 61.2 dB(A). The subjects were five male university students (22–24 years old) with normal hearing. The design of the test is illustrated in Figure 7. The practice trials and main trials were each presented randomly for each subject. The trial conditions are summarized in Table 4. Each subject was instructed to identify the direction of the sound after listening to the binaural signals and mark it on an answer sheet (shown in Figure 8). Answers could be given at intervals of 15°.

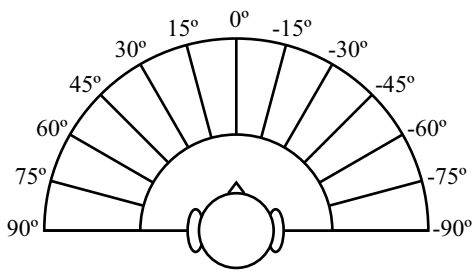

Figure 8: Answer sheet

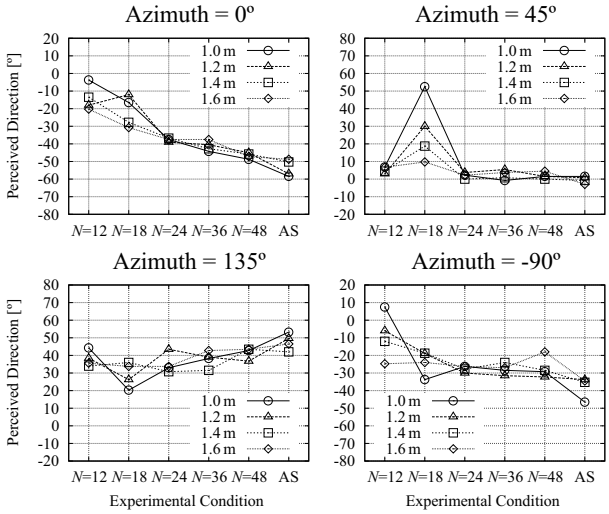

Figure 9: Localization results

As shown in Figure 9, as the number of image sources was increased, the perceived direction approached that of the actual source. This means that the directional perception with the proposed system is the same as that of the actual source if there is a sufficient number of image sources.

4 Conclusion

A sound field auralization system was described that adds sound information to the Free Viewpoint Television system. Image source signals are estimated from the sounds captured by microphones placed at the same positions as cameras in a two-dimensional sound field. Testing on a PC showed that image source signals can be estimated if the number of microphones is more than three times the number of image sources. The results of a localization test indicated that directional perception is reproduced if the number of image sources is large enough.

The next step is to evaluate the effectiveness of the system in an actual environment by measuring the room transfer functions. It also needs to be evaluated in a three-dimensional sound field.

References

[1] T. Fujii and M. Tanimoto, 'Free Viewpoint TV System Based on Ray-space Representation'. *Proc. SPIE ITCom* 2002, Boston, Vol. 4864, pp. 175-189 (2002)

[2] M. Tanimoto and T. Fujii, *'FTV - Free Viewpoint Television'*, M8595, ISO/IEC JTC1/SC29/WG11,

August 2002.

[3] A. J. Berkhout, D. de Vries, and P. Vogel, 'Acoustic Control by Wave Field Synthesis'. *Journal of Acoustical Society of America*, Vol. 93, No. 5, pp. 2764-2778 (1993)

[4] J. Blauert, *'Spatial Hearing'*, pp. 78, revised edition, MIT Press, Cambridge, Mass. (1997)

[5] J. Bauck and D. H. Cooper, 'Generalized Transaural Stereo and Applications', *Journal of Audio Engineering Society*, Vol. 44, No. 9, pp. 683-705 (1996)

[6] W. G. Gardner and K. D. Martin, 'HRTF Measurement of a KEMAR', *Journal of Acoustical Society of America*, Vol. 97, No. 6, pp. 3907-3908 (1995), http://sound.media.mit.edu/KEMAR.html

[7] T. Nishino, M. Ikeda, K. Takeda, and F. Itakura, 'Interpolating Head Related Transfer Functions," *Proc. WESTPRAC* VII, Kumamoto, pp. 293-296 (2000), http://www.itakura.nuee.nagoya-u.ac.jp/HRTF/

[8] Hearing Development Research Laboratory, University of Wisconsin, http://www.waisman.wisc.edu/hdrl/

[9] Advanced Acoustic Information Systems Laboratory, Tohoku University, http://www.ais.riec.tohoku.ac.jp/lab/db-hrtf/

[10] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, 'The CIPIC HRTF Database', *Proc. IEEE WASPAA* 2001, New Paltz, pp. 99-102 (2001), http://interface.cipic.ucdavis.edu/.

[11] Listen HRTF Database, Listen Project, http://recherche.ircam.fr/equipes/salles/listen/

[12] Shimada Laboratory, Nagaoka University of Technology, http://www.audio.nagaokaut.ac.jp/hrtf/

[13] Y. Tahara, K. Ishikawa, H. Kawamura, S. Sasaki, and M. Nakamura, 'Acoustic Scale Model Experiments Using a Piezoelectric Dodecahedral Speaker System', *Journal of Acoustical Society of Japan*, vol. 59, No. 10, pp. 614-621 (2003), in japanese.

[14] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, 'An Optimum Computer-generated Pulse Signal Suitable for the Measurement of Very Long Impulse Responses', *Journal of Acoustical Society of America*, vol. 97, No. 2, pp. 1119-1123 (1995)