A HYBRID SPEAKER ARRAY-HEADPHONE SYSTEM FOR IMMERSIVE 3D AUDIO REPRODUCTION

Rishabh Ranjan and Woon-Seng Gan

School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore
{rishabh001, ewsgan}@ntu.edu.sg

ABSTRACT

Spatial sound systems aim at rendering realistic sound experience to the listeners with uniform sound fields in the entire listening area. Today with the advancement of multichannel surround sound techniques, such systems are being practically realized, especially, at theatres, lecture halls, auditoriums, etc. Current practices, which are most widely used as home theatre systems, are based on multichannel stereophony, like 5.1, 10.2 and higher surround channel system. These systems require multiple loudspeakers to be placed in fixed configuration but often constrained by the room size. Sound reproduction systems like wave field synthesis (WFS) based on principle of natural propagation of sound waves, can create replica of true sound field uniformly over an extended listening area. However, WFS based systems too require hundreds of densely spaced loudspeakers enclosing the listener area and thus, difficult to realize in homes. In this paper, we introduce a new hybrid system by combining the WFS and binaural synthesis over headphones (based on active noise control techniques) to reduce the need of installing loudspeakers everywhere in a living room.

Index Terms— WFS, Binaural synthesis, ANC

1. INTRODUCTION

Audio rendering systems have evolved significantly over the last couple of decades. Different sound technologies have been developed and being commercialized in the consumer market. Two channel stereophony based systems have advanced into multichannel set ups creating an immersive experience for the listeners. However, such systems require complex loudspeaker placements, constrained by the room size. Listeners' movements are constrained as the best impression is only achieved in the center of listening area, which is far from ideal. Another technology, which is also widely used for private listening is the binaural synthesis over headphones. But conventional headphones suffer from the problem of in-head localization and front-back confusions. Sound field synthesis based approach like wave field synthesis (WFS) [1, 2], higher order ambisonics (HOA) [3] using loudspeaker arrays are known to overcome the problems of conventional listening systems. They exhibit homogenous sound field over extended listening area, while sounds can be perceived to come from anywhere in the virtual space around you. Although, theoretically they provide high fidelity sound field, but at the cost of umpteen densely spaced loudspeakers and that is why, such systems are seen only in places like cinemas or concert halls.

Figure 1 Proposed hybrid system in a home scenario

In this work, a new hybrid system is proposed to overcome the practical and physical limitations of the conventional reproduction techniques. The proposed system combines WFS and binaural synthesis over headphones playback to reduce the need of installing many loudspeakers in a home entertainment setups. WFS is used to drive a frontal loudspeaker array, which is positioned on the top of the TV screen, to provide strong frontal localization cues. The rear and side auditory cues are presented over headphones via a virtual WFS technique to complete the entire 360 degree auditory scene presentation as shown in Figure 1. A techniques based on the active noise control (ANC) concept is introduced to ensure that virtual WFS over headphones perform as close to the physical speaker (sides and rear) arrays that would be required in a full-fledge WFS system. Measurement results on dummy head showed that the proposed system performs very close to an enclosed array setup driven by WFS.

This paper is organized as follows. Section 2 outlines some of the recent approach in rendering WFS effect over headphones. The proposed hybrid system is introduced in Section 3, which describes the ANC techniques for binaural synthesis over headphones using Virtual WFS. The experimental results with key findings are reported in the Section 4, and concluded in Section 5.

2. RELATED WORK

There have been some works in recent past to combine WFS and binaural synthesis. Menzel et. al. [4, 5] presented a novel system called "Binaural Sky" to reproduce a virtual headphone using binaural room synthesis. Authors used a pair of focused sources reproduction, using overhead WFS circular array, in front and close to the listener. These focused sources act as the transaural loudspeakers and they can be easily moved around by adjusting the driving signals. In [6], a

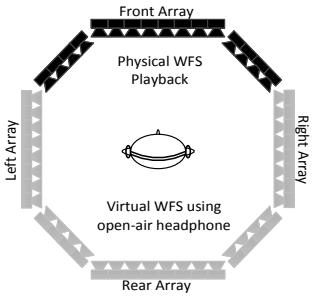

Figure 2. Proposed hybrid system structure (■ : physical speakers; ■ : virtual speakers)

simulation of wave field synthesis is presented for perceptual quality analysis of different complex set ups using virtual WFS with the help of headphone playback. The main objective of the above work was to analyze the WFS through binaural playback. Through subjective tests, it was shown that azimuths were accurately estimated by the subjects. In this paper, we combine the physical WFS frontal array playback with binaural playback for side and rear auditory image over headphones to provide a complete 3D sound reproduction.

3. PROPOSED HYBRID SYSTEM

The proposed hybrid system consists of a frontal loudspeaker array mounted on a visual display along with an open-air headphone worn by the listener to perceive both physical frontal sound and virtual side and rear sound image. Figure 2 shows the structure of the proposed hybrid system. An open and inverted U-shaped loudspeaker array is considered for the frontal projection, while symmetric virtual WFS array for side and rear is used for binaural reproduction over headphones. The main goal here is to retain the spatial and temporal characteristics of WFS at the listener position by using a combination of two reproduction techniques. WFS rendering is being used in the back-end to compute all the loudspeaker signals (including physical as well as virtual ones) at the same time. Overall system block diagram is shown in Figure 3 and can be divided into three processing stages:

1) WFS renderer
2) Frontal auditory scene processing using WFS
3) Rear and side auditory scene processing using virtual WFS over headphones

In the following sub-sections, we will provide brief overview of the WFS renderer and present the three processing stages.

3.1 WFS Renderer

WFS renderer is the processing core of the proposed hybrid system, which computes the driving signals of all the loudspeakers comprising the WFS enclosed set up shown in Figure 2. WFS is fundamentally based on the Huygens' principle, which states that secondary sources can be used to synthesize the natural wave fronts of the primary sources [7]. WFS driving signals (loudspeaker signals) are derived by solving discretized Rayleigh Integral with sound field of a monopole point source and applying stationary phase approximation [1, 2, 8] as:

$$Q(n,w) = S(w)\sqrt{\frac{jw}{2\pi c}}\sqrt{\frac{z_2}{z_2+z_1}}\,\frac{z_2}{|\overrightarrow{r_1}|}\,\frac{e^{-j\frac{w}{c}|\overrightarrow{r_1}|}}{\sqrt{|\overrightarrow{r_1}|}}, \qquad (1)$$

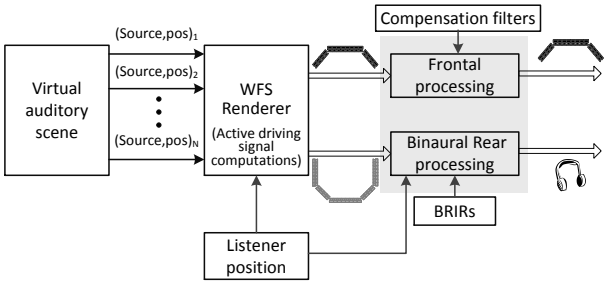

Figure 3. Overall hybrid system block diagram

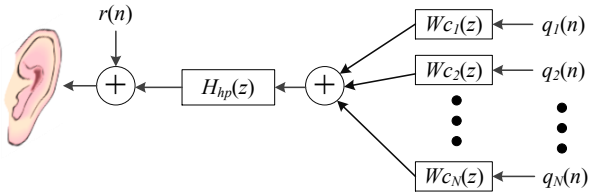

Figure 4. Headphone isolation compensation for frontal processing

where, $Q(n,w)$ is the driving signal of n^{th} loudspeaker for virtual source $S(w)$ behind the loudspeaker array (non-focused source). z_1 and z_2 are the distance of the source and the listener from array respectively. $\overrightarrow{r_1}$ is the position vector from source to the loudspeaker, n. Thus, driving signal at each loudspeaker can simply be calculated by summing the delayed and weighted contribution from filtered source(s) signal(s). An important property of WFS is that it can synthesize virtual source even in front of the loudspeaker array (focused source) creating an illusion of source around the listener. The only restriction is that listener cannot be positioned between array and source. Driving signals for a focused virtual source is defined similarly as (1), while reversing the phase components in (1) [8, 9]. Based on the position of source and listener, appropriate loudspeakers are activated for rendering using active secondary source criterion method by Spors [2, 10]. In the next subsection, we present the frontal auditory scene reproduction via physical loudspeaker array.

3.2 Frontal auditory scene processing

In the subsequent stage, based on which loudspeakers are active in the WFS enclosed array set up (Figure 2), either frontal playback using WFS or rear and side playback over headphones or both is rendered. Therefore, we must ensure that auditory scene synthesized at listeners' ears using the hybrid system retains the natural characteristics of WFS. Open-air headphones are used in the proposed hybrid system such that direct sounds from the loudspeakers pass through the headphone without much attenuation. However, due to the passive structure of the headphone, it acts as a low pass filter and high frequencies are attenuated depending on the transfer function of the headphone shell. Headphone isolation can be compensated by playing filtered driving signals through the headphone as shown in Figure 4. Driving signals computed by the WFS renderer are passed through compensation filters, $Wc_i(z)$ corresponding to each physical loudspeakers and played back over the headphones such that when added to the direct signal, $r(n)$, headphone become acoustically transparent. Compensation filters are estimated individually using the speakers' responses measured with and without headphone ($G_i(z)$ and $P_i(z)$, respectively) and employing

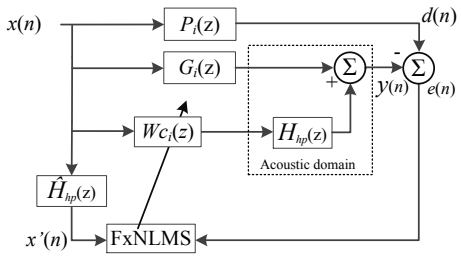

Figure 5. Headphone isolation compensation filters estimation

normalized version of filtered-x least mean square algorithm (FxNLMS), as shown in Figure 5. The main advantage of this approach is that the set of estimated compensation filters is valid for any virtual source positions rendered by WFS.

The WFS frontal array prototype is built using 16 speakers, with 8 in the middle and 4 each on the either side positioned at 45^o, as shown in Figure 2. Two 10-channel MOTU Ultralite-mk3 hybrid sound cards are used to drive the WFS frontal array as well as the headphones. Speaker responses were measured on Neumann KU 100 head with and without headphones using AKG 417 miniature microphones mounted near ear opening. Open-air headphones AKG K702 were used for the measurement process. To validate the headphone compensation approach, we estimated the WFS virtual source frequency response at listener position in centre of the listening area by playing the driving signals through the speaker array. Figure 6 shows the frequency response of a WFS virtual source measured with and without headphones along with the compensated response. Clearly, due to the headphone, attenuation of 15-20 dB is observed in the high frequency region above 1.5 kHz for the headphone modified frequency response. After applying headphone compensation filters and adding with the direct signal, the resultant compensated frequency response approaches the original frequency response measured without headphones, as shown in Figure 6.

3.3 Rear and Side auditory scene processing

For rear and side auditory scene processing, driving signals corresponding to the virtual loudspeakers are used to synthesize the binaural signals to be reproduced over headphones such that they cannot be distinguished from the sounds coming from the physical loudspeakers. Using the virtual WFS technique, we synthesize the binaural signals by convolving the driving signals computed by WFS renderer with the transfer functions of each of the virtual speakers to the listeners' ears and summing them together before playing through the open-air headphones. This transfer function measured in a room environment is known as the binaural room transfer function. In addition to the individualized binaural transfer functions, headphone transfer functions (HPTF), which modify the intended spectrum, are also unique to every individual, and must be equalized individually for accurate reproduction of binaural signals [11]. We consider an open-air headset structure with two pairs of binaural microphones (internal and external) positioned to compensate for individual HPTFs as well as repositioning of headphones. Internal microphones are positioned near the ear opening, while external microphones are fixed just outside the ear cup. Signals acquired by the external microphones contain all the

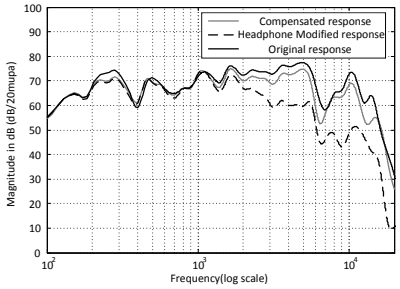

Figure 6. Headphone isolation compensation results for a WFS virtual source 1 m behind the array

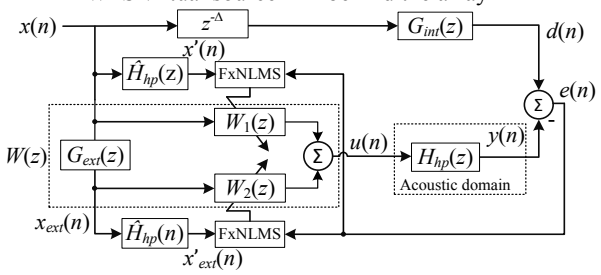

Figure 7. Proposed hybrid FxNLMS algorithm

spatial information except the pinnae and headphone shell reflections. A hybrid mechanism [12] is proposed using ANC technique based on simple combination of conventional and a modified FxNLMS algorithm, as shown in Figure 7. The modified FxNLMS further improves the convergence rate of conventional FxNLMS by using an additional filter in the form of $G_{ext}(z)$ in the secondary path but at the cost of higher steady state error. $G_{int}(z)$ and $G_{ext}(z)$ represent the speaker transfer functions, measured at internal and external microphones, respectively. An additional forward delay is introduced in the primary path to compensate for the overall delay of secondary path. The hybrid approach combines both the FxNLMS versions for optimized steady state error and faster convergence rate. $W(z)$ is the combined adaptive filter defined as:

$$W(z) = W_1(z) + G_{ext}(z)W_2(z), \tag{2}$$

where $W_1(z)$ and $W_2(z)$ are adaptive filters based on conventional and modified FxNLMS, respectively with their optimum solution defined as:

$$W_1^o(z) = \frac{G_{int}(z)z^{-\Delta}}{H_{hp}(z)} \quad \text{and} \quad W_2^o(z) = \frac{G_{int}(z)z^{-\Delta}}{G_{ext}(z)H_{hp}(z)}\,. \tag{3}$$

Therefore, spatial filter $G_{ext}(z)$ is the only difference between the two FxNLMS algorithms. This spatial filter assists the hybrid FxNLMS to converge faster and attain optimum steady state error at the same time. Figure 8 compares the hybrid FxNLMS performance with the other two approaches. Clearly, the hybrid FxNLMS is optimum in terms of both steady state error as well as convergence rate. As shown in Figure 8, hybrid FxNLMS is twice as fast as the conventional FxNLMS with an improvement of 20 dB of error reduction. A multichannel version of the hybrid FxNLMS algorithm is implemented in this work for rear and side processing of the proposed system in Figure 2. WFS driving signals of the virtual loudspeakers are taken as input signals and internal microphones as error sensors for the multichannel version. The main objective is to compensate for individual HPTF, while adapting to the desired WFS response and make the headphone acoustically transparent so as to sound as close to physical WFS playback.

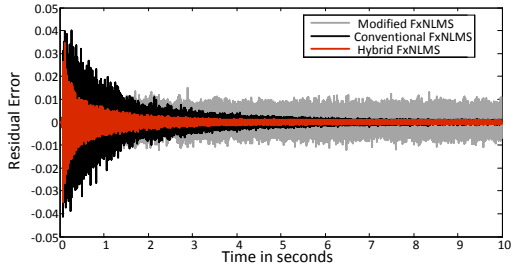

Figure 8. Comparison of residual error plots for hybrid FxNLMS with traditional and modified FxNLMS

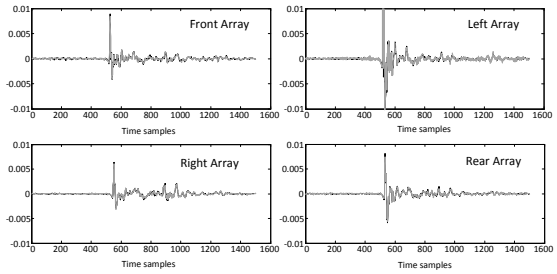

Figure 9. Impulse response plots: Real Vs Virtual WFS (Black: Measured IR Grey: Estimated IR)

Table 1. Spectral distortion scores (dB) for the virtual WFS over headphones

Source	Front Array		Left Array		Right Array		Rear Array	
	Simulated	**Measured**	**Simulated**	**Measured**	**Simulated**	**Measured**	**Simulated**	**Measured**
Far Center	0.18	1.90	0.35	1.79	0.67	2.02	0.40	1.82
Far Left	0.17	1.84	0.10	2.51	0.59	1.80	0.20	1.51
Far Right	0.32	1.71	0.60	1.68	0.45	2.04	0.27	1.73

4. EXPERIMENTAL RESULTS

To evaluate the performance of the hybrid system, virtual sounds reproduced over headphones must be seamlessly integrated with the sounds coming from physical WFS array. In other words, both the spectral and temporal features of the actual WFS playback must be retained in the virtual sound field reproduction as well. We evaluate the virtual WFS performance by recording response of sine sweep signal played through the loudspeakers. Measurements for virtual WFS via binaural synthesis are done using the WFS frontal array (Figure 2) by rotating the dummy head in steps of 90°. Therefore, there are 4 set of measurements corresponding to *front*, *left*, *right* and *rear* array as indicated in Figure 2. For all the 4 sets, three non-focused virtual sources, namely, *Far centre*, *Far left*, and *Far right*, were considered. Centre virtual source is positioned 1m behind the array such that only middle 8 loudspeakers are active. Left and right virtual sources are positioned respectively to left and right side of the array such that either left or right 4 loudspeakers are active along with the middle array. Binaural impulse responses corresponding to all the 12 virtual source positions (4 sets × 3 virtual sources) were measured using dummy head with the headset worn. Furthermore, individual binaural room impulse responses were measured for all the 16 loudspeakers and all the 4 sets on the dummy head. Spectral distortion (SD) score [13, 14] is used to objectively quantify the spectral error between frequency response of virtual source for virtual WFS and physical WFS:

$$SD = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(20log\frac{|H(f_k)|}{|\hat{H}(f_k)|}\right)^2} \quad \text{[dB]}, \tag{4}$$

where, $H(f)$ is the magnitude response of reference frequency response (either measured or simulated), $\hat{H}(f)$ is the estimated transfer function response, and K is the total number of frequency samples in the observed range (100 Hz – 16 kHz).

Table 1 lists all the SD scores for the 3 virtual source positions and 4 sets measurement of loudspeaker array. Two reference transfer functions were used to compute the SD scores. Measured reference is the actual measurement on the dummy head, while playing WFS virtual source through the loudspeaker array. Simulated reference represents direct convolution of WFS driving signals with speaker impulse response and synthesized at the listeners' ears by summing them together. For estimated response, WFS driving signals convolved with the corresponding hybrid adaptive filters, summed together, played through the headset and recorded at dummy head's ears. Clearly, using the multichannel version of hybrid FxNLMS, the spectral distortion is less than 1 dB with the simulated reference which means signals synthesized using virtual WFS are exactly similar to the desired WFS response. However, with measured reference, slightly higher average SD scores were observed of around 2 dB, ensuring no perceptual difference between the physical and virtual WFS performance. This spectral deviation from the simulated frequency response might be due to the room reflections and non-linearity in the air. The temporal characteristics of virtual WFS are validated by comparing the measured impulse responses (IRs) of physical WFS virtual source with that of virtual WFS. Figure 9 shows the impulse responses of *Far centre* WFS virtual sources for all the 4 sets computed for left ear. Evidently, virtual WFS sources are reproduced accurately, with temporal features well matched with the measured IRs as shown. However, very little differences in magnitude are observed similar to as in spectral distortion method. Hence, it can be concluded that rear and side auditory scene can be reproduced coherently along with the physical WFS giving us an immersive sound experience.

5. CONCLUSION

In this work, we presented a hybrid system combining a physical WFS array reproducing the frontal auditory scene, while rear and side auditory scene is synthesized over headphones using virtual WFS. Open-air headphones, which are embedded with two pairs of microphones, are used to adapt to every individual, which is essential for accurate sound localization. With emphasis on the strong frontal localization cues, we use the physical WFS frontal array along with visual aid to provide listener an immersive sound experience when used in conjunction with open-air headset for surround sound. WFS renderer is used as the processing core to compute all the driving signals and therefore, provide seamless integration of physical WFS with the virtual WFS over headphones. Dummy head measurement results show that spatial and temporal characteristics are retained in the virtual sound field reproduction as well. A detailed subjective study is underway to validate the practicality of the proposed system.

6. REFERENCES

[1] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America,* vol. 93, p. 2764, 1993.

[2] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *124th AES Convention*, 2008, pp. 17-20.

[3] M. A. Gerzon, "Periphony - With-Height Sound Reproduction," *Journal of the Audio Engineering Society,* vol. 21, pp. 2-10, 1973.

[4] D. Menzel, H. Wittek, G. Theile, and H. Fastl, "The binaural sky: A virtual headphone for binaural room synthesis," in *International Tonmeister Symposium*, Bavaria, Germany, 2005.

[5] K. Laumann, G. Theile, and H. Fastl, "A virtual headphone based on wave field synthesis," *Journal of the Acoustical Society of America,* vol. 123, p. 3515, 2008.

[6] F. Völk, J. Konradl, and H. Fastl, "Simulation of wave field synthesis," *J Acoust Soc Am,* vol. 123, p. 3159, 2008.

[7] P. Vogel, "Application of wave field synthesis in room acoustics," PhD Thesis, Delft University of Technology, 1993.

[8] D. de Vries, "Wave Field Synthesis," in *AES Monograph*, New York, 2009.

[9] A. F. Franco, S. Merchel, L. Pesqueux, M. Rouaud, and M. O. Sorensen, "Sound Reproduction by Wave Field Synthesis," Aalborg University, Project Report2004.

[10] S. Spors, "Extension of an analytic secondary source selection criterion for wave field synthesis," in *Audio Engineering Society Convention 123*, 2007.

[11] F. Brinkmann and A. Lindau, "On the effect of individual headphone compensation in binaural synthesis," *Fortschritte der Akustik: Tagungsband d. 36. DAGA,* pp. 1055-1056, 2010.

[12] R. Ranjan and W.-S. Gan, "Applying Active Noise Control Technique for Augmented Reality Headphones," in *Internoise 2014*, Melbourne.

[13] T. Nishino, N. Inoue, K. Takeda, and F. Itakura, "Estimation of HRTFs on the horizontal plane using physical features," *Applied Acoustics,* vol. 68, pp. 897-908, 2007.

[14] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu, "Distance dependent head-related transfer function database of KEMAR," in *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*, 2008, pp. 466-470.