JCA2007 4-6 June
The Japan-China Joint Conference of Acoustics 2007

A DYNAMIC VIRTUAL AUDITORY DISPLAY: ITS DESIGN, PERFORMANCE, AND PROBLEMS IN HRTF SWITCHING

Makoto Otani and Tatsuya Hirahara
Faculty of Engineering, Toyama Prefectural University

ABSTRACT

A software-implemented dynamic virtual auditory display (DVAD) has been developed by the authors. The DVAD responds to the listener's head rotation by using a head-tracking device and switching head-related transfer functions (HRTFs), thereby presenting a highly realistic virtual auditory space to the listener. The DVAD operates on Windows XP and does not require high-performance computers. The measured total system latency of our DVAD is 50–60 ms, which is practically small enough for applications and localization experiments. The occurrence of click noises due to HRTF switching is a problem; this is commonly detected in previous DVAD systems. The detectability of click noises depends on both the spatial resolution of the HRTF and the bandwidth of the source signal. An analysis of the excitation patterns of the synthesized binaural signals revealed that click noises are more detectable when a narrower-band source signal is used; this indicates that higher HRTF with higher spatial resolution is necessary for presenting various real-world sound sources.

INTRODUCTION

Sound localization is an important function of human auditory system. By using interaural and spectral cues, we can perceive the direction of the source as well as the distance from the source. These cues are included in the acoustic transfer functions from the sound source to each ear, which are know as head-related transfer functions (HRTFs). In addition to these static cues, the body and head motions provide a strong localization cue, in other words, an auditory-motor integration, as Wallach showed that a head rotation improves the accuracy of sound localization [1].

The binaural technique is a promising approach toward realizing a virtual auditory display (VAD). This technique is achieved by controlling the acoustic signals sent to the listener's ears by using headphones [2]. On the other hand, the main drawback of this technique is that it cannot respond to a listener's head motion, thereby degrading its reality; in other words, lack of auditory-motor integration is its primary disadvantage. In order to present a highly realistic virtual auditory space, dynamic virtual auditory displays (DVADs) have been developed using a head-tracking device. DVADs vary their output signals, *i.e.*, the binaural signals, in response to a listener's head motion, which is tracked using a head-tracking device. However, it is necessary that total system latency (TSL) is small so that a listener does not detect the delay between a head motion and the corresponding auditory inputs. Recently, software-implemented DVADs have been realized without using any specially designed digital signal processing hardwares (*e.g.*, [3]), due to the enhanced computing performance [4, 5, 6, 7]. Therefore, we developed a software-implemented DVAD using a head-tracking device as a research tool for investigating the mechanism of auditory space perception.

The DVAD is implemented by switching the HRTFs in response to the listener's head motion. HRTF switching produces a wave discontinuity in synthesized binaural signals. This discontinuity can be detected in the form of a click noise with a broad bandwidth, leading to degradation in the quality of the presented auditory space.

This paper describes the design of our DVAD and its basic performance, including a discussion on TSL. In addition, an excitation pattern (EPN) analysis is performed to investigate the detectability of click noises due to HRTF switching.

DESIGN

Our DVAD consists of a PC operating on Windows XP, headphones, an audio device, and a head-tracking device (Fig.1). The head-tracking device (NEC/TOKIN MDP-A3U9S), which is connected to the PC via a USB interface, is capable of detecting a listener's head motion. The PC receives the listener's posture angles from the head-tracking device and switches the HRTFs in response to them. The source signal is convolved with the selected HRTFs to synthesize binaural signals for each ear. The PC transmits binaural signals to the headphones using a DA converter (Roland EDIROL UA-101). The posture angles are updated so that the appropriate HRTFs are always selected. A virtual sound image is presented at a certain position in the absolute coordinate system. The posture angles are updated at 125 Hz, *i.e.* at every 8 ms.

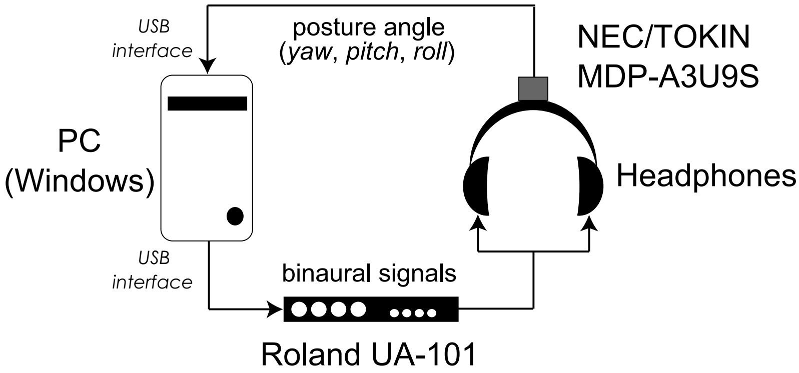


Fig. 1. System configuration.

Binaural signals are synthesized by using the overlap-add method [8]. Microsoft multimedia extended API DirectSound is used as as the audio interface to transmit the sound signals to the DA converter. The entire software is coded using Visual C#, except for the interface of the head-tracking device that is coded in C++. The HRTF database consists of a set of head-related impulse responses (HRIRs), which are time-domain representations of the HRTFs, and header information such as HRIR data length and spatial resolution. The length of the HRIR is set to 512 points. The source signal is divided into 513-point (12-ms) segments, and subsequently convolved with the HRIRs that are being updated. 1024-points segmented binaural signals are sequentially overlap-added, thereby yielding entire binaural signals.

PERFORMANCE

Total system latency (TSL) is the time delay from when a change occurs in the posture angle to when the sound output from the headphones reflects the change. The TSL is a significant factor of DVADs. Extremely large TSLs are perceptually detected, thereby reducing the localization performance and naturality. Brungart *et al.* reported that typical

Table 1. Main features of representative dynamic virtual auditory displays.

	OS	head-tracking device	TSL [ms]
DIVA [4]	Unix, Linux	Ascension Motion Star	110-160
RTVAS [5]	Linux	Intersense IS-900VWT	≤ 7
SLAB [6]	Windows	—	6.1(exc. tracking delay)
Yairi *et al*'s system [7]	Linux	Polhemus FASTRAK	12.01
Our system	Windows	NEC/TOKIN MDP-A3U9S	50-60

subjects did not detect a delay that was less than 80 ms; however, the just-noticeable difference (jnd) decreased to 25 ms when the reference signal having the minimum delay was presented simultaneously [3]. Yairi *et al.* reported that the probability of detecting 80-ms TSL is approximately 50% when subjects shake their heads vigorously [7].

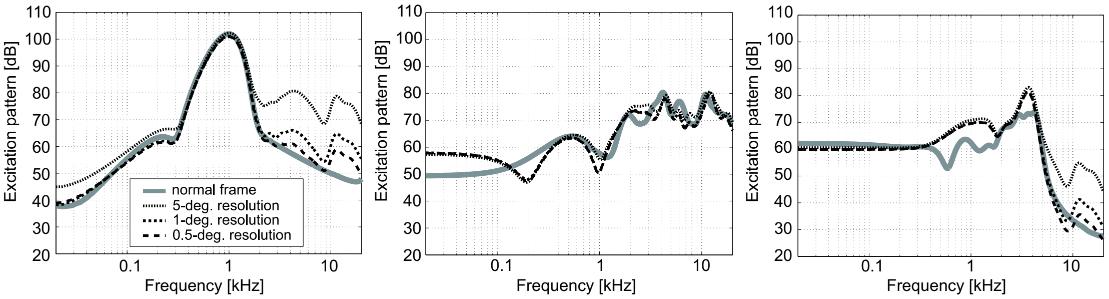
The signal processing delay that occurred in our DVAD was 30–40 ms. The TSL of our system, which comprises updating delay caused by the head-tracking device (8 ms) and a signal processing delay, was approximately 50–60 ms. A TSL of 50–60 ms is smaller than the jnd when a stimulus is presented without a reference. As for the localization performance, Wenzel *et al.* showed that even a 500-ms delay does not prevent accurate localization of 8-s duration stimulus, whereas the same delay increases the front-back errors of 3-s duration stimulus [9]. Therefore, it is expected that the TSL of our system does not have a significant effect on the localization performance.

Table 1 shows the main features of the existing DVAD and our DVAD. In these systems, SLAB [6] and our system operate on Windows. The other systems operate on Linux or Unix. From the viewpoint of process priority management, Linux and Unix may be superior to Windows; however, we developed our system on Windows XP for general versatility. Some systems achieve a TSL of less than 10 ms. Although the TSL of our system is 50–60 ms, it is small enough to be used for localization experiments. In addition, our system operates without glitches even on a notebook PC such as a Pentium M 1.5-GHz with 512 MB RAM.

DISCONTINUITY DUE TO HRTF SWITCHING

In the real world, head rotations can produce continuous binaural signals that are fully integrated with motor sensations, which produce strong localization cues. Such continuous binaural signals can be accurately captured by using devices such as TeleHead [10], which is a steerable dummy head that tracks the three-dimensional head movement of a listener; however in DVADs, continuous binaural signals are modeled discretely as discontinuous binaural signals due to an HRTF switching. Such discontinuity produces a detectable click noise, which degrades the quality of the presented sound [5, 11]. Very few researches have been performed for such auditory artifacts that occur due to the discontinuity. Kudo *et al.* suggested spectrum distortion width as a parameter for evaluating the various HRTF switching methods [12]. HRTF differences in amplitude and phase, before and after switching, are considered to have an impact on the generation and a detection of click noises. A higher spatial resolution reduces the HRTF differences; however, it is not clear how high a spatial resolution must be in order to eliminate the click noise perceptually. Furthermore, the bandwidth of the source signal would have significant influence on the click noise detection, although its influence has not been investigated.

In order to evaluate the influences of a source signal and HRTF spatial resolution on



(a) 1-kHz pure tone (b) white noise (c) band-limited noise (–4 kHz)

Fig. 2. Distorted EPNs due to HRTF switching with 5.0-, 1.0-, and 0.5-degree HRTF spatial resolution

click noise detection objectively, EPNs of the synthesized binaural signals are analyzed using an auditory filter [13]. The source signals are 1-kHz pure tone, white noise, band-limited noise having a cut-off frequency of 4 kHz. Numerically calculated HRTFs of a head and torso simulator (Brüel&Kjær 4128C) using the boundary element method [14] were used, thereby enabling the construction of an HRTF database for a 0.5-degree spatial resolution without giving an error due to interpolation. The posture angles obtained from the head-tracking device are not used during HRTF switching. Instead, HRTFs are switched automatically.

RESULTS

EPNs are calculated for each 128-points (2.9-ms) frame, which is obtained from synthesized binaural signals using 64-point (1.5-ms) shifting. In this paper, the frames in which the HRTFs are switched are referred to as “switched frames”, and those in which the HRTFs are not switched are referred to as “normal frames”.

First, the EPNs of the first switched frame and its preceding normal frame, *i.e.*, the frame just before the first switched frame, are shown in Fig. 2. The first switched frame is a frame in which the HRTFs are switched from 0 degrees (front source position) to 5.0, 1.0, and 0.5 degrees left in the azimuth plane. Fig. 2(a) shows EPNs for a 1-kHz pure tone. The normal frame has a single peak at 1 kHz, whereas all the switched frames display broadly distorted EPNs at the off signal frequencies. The peak at 1 kHz is not distorted by the HRTF switching. Therefore, it can be said that HRTF switching distorts an EPN, particularly at the off signal frequencies. In addition, the figures also show that a higher HRTF spatial resolution generates lesser distortion. Fig. 2(b) shows the case of white noise. Unlike the case of the 1-kHz pure tone, in this case, HRTF switching does not significantly distort EPNs. Since white noise yields EPNs at all the characteristic frequencies, HRTF switching does not produce significantly extra EPNs. This phenomenon corresponds to the masking effect in which white noise plays the role of a masker. In addition, HRTF resolution has negligible influence on EPNs. Deviations between EPNs of both frames are due not only to HRTF switching, but also to temporally changing frequency characteristics of white noised. This is because white noise has temporally unstable frequency characteristics in a short time, thereby producing temporally unstable EPNs. Fig. 2(c) shows the case of band-limited noise (cut-off at 4 kHz). As in the case of the 1-kHz pure tone signal, HRTF switching leads to distortion of EPNs at the off signal frequencies, *i.e.*, above 4 kHz. Furthermore, higher HRTF spatial resolution leads to lesser distortion. These results indicate two major aspects pertaining to HRTF switching.

Table 2. Average of SDs for normal frames and switched frame.

Source signal	HRTF spatial resolution [degrees]	Averaged SD [dB]		Averaged SD Difference [dB]
		Switched frames	Normal frames	
1-kHz pure tone	**5.0***	1.79	0.30	1.49
	1.0*	0.81	0.28	0.53
	0.5*	0.63	0.28	0.35
White noise	5.0	1.26	1.24	0.02
	1.0	1.28	1.22	0.06
	0.5	1.27	1.23	0.04
Band-limited noise (–4 kHz)	**5.0***	1.65	1.28	0.37
	1.0	1.33	1.27	0.05
	0.5	1.30	1.28	0.02

* **Click noises were detected in the preliminary subjective experiment**

First, HRTF switching significantly distorts EPN at off signal frequencies. Second, a high HRTF spatial resolution leads to lesser distortions.

Next, spectral distances (SD) of EPNs are calculated between the neighboring frames. SDs are calculated as follows:

$$\mathrm{SD}(n) = \sqrt{\frac{1}{K}\sum_{j=1}^{K}\left(20\log_{10}\frac{E(n,f_{cj})}{E(n-1,f_{cj})}\right)^2},$$

where n is a frame index; K, the number of characteristic frequencies; $E(n, f_{cj})$, the EPN at frame n and characteristic frequency f_{cj}. Subsequently, the average SDs are calculated for both normal and switched frames for 1-s binaural signals. Table 2 shows each averaged SD value for various source signals having a duration of 8 s and HRTF resolutions. The superscripted asterisk in the table indicate that a subject detected click noises in the corresponding stimulus.

For the 1-kHz pure tone and 5.0-degrees resolution, the averaged SD of switched and normal frames are 1.79 dB and 0.3 dB, respectively. The difference between both the frames is 1.49 dB. In comparison to the case of the 5.0-degrees resolution, the averaged SDs of switched frames for 1.0- and 0.5-degree resolutions are smaller, whereas those of the normal frames are almost constant, thereby revealing that higher HRTF resolutions yield lesser distortion in the EPNs due to HRTF switching. For white noise, the differences in the averaged SDs between both the frames are small ($\leq$ 0.1 dB), regardless of the HRTF spatial resolutions. This is because white noise excites all the characteristic frequencies, thereby masking additional EPNs due to HRTF switching. For band-limited noise with a cut-off frequency of 4 kHz, the averaged SDs of normal frames are almost the same as those for white noise, whereas those of switched frames are larger, particularly for a 5.0-degrees resolution. This is because off signal frequencies are excited at switched frames, as shown in Fig. 2(c). This feature is similar to that observed in the case of the 1-kHz pure tone, although the differences in the SDs between both the frames are much smaller since the normal frames have larger averaged SDs.

From these results, it is observed that as the bandwidth of the source signal broadens, the range of the distorted EPNs narrows due to HRTF switching, thereby resulting in smaller values of the averaged SD differences between the normal and switched frames.

In addition, it is also observed that a higher HRTF resolution decreases the averaged SD differences, *i.e.*, distorted EPNs due to HRTF switching. Furthermore, this objective evaluation agrees well with the results obtained from the preliminary subjective experiment; for example, click noises were detected in the cases where the averaged SD differences were relatively large.

A broad-bandwidth source signal such as white noise can mask click noises and reduce their detectability. In contrast, a narrow-bandwidth source signal or a pure tone signal can mask click noises only at around signal frequencies. Consequently, the distorted EPN occurs at off signal frequencies, thereby making the click noises more detectable.

SUMMARY

A software-implemented DVAD developed by the authors is introduced in the paper. Our DVAD operates on Windows XP with a TSL of 50–60 ms, which is practically small enough for actual applications and sound localization experiments.

EPNs are analyzed to investigate click noise detection caused by HRTF switching. The results show that HRTF switching generates a distorted EPN. HRTFs with higher spatial resolution leads to lesser distortion; however, its effect depends on the bandwidth of the source signal. A source signal with a broad bandwidth, such as white noise, can efficiently mask click noise and perceptually eliminate it. However, in order to achieve the more realistic DVADs, which can handle many types of sound source signals, even HRTF with the highest spatial resolution used in this study (0.5-degrees) is insufficient. Hence, it is necessary to significantly increase the spatial resolution of the HRTF or refine the HRTF switching method for smoother switching without producing an unacceptable extra increase in TSL.

REFERENCE

[1] H. Wallach, “On sound localization”, *J. Acoust. Soc. Am.*, **10**, 270–274 (1939).

[2] J. Blauert, *Spatial Hearing* (The MIT Press, London, 1997).

[3] D.S. Brungart, B.D. Simpson, and A.J. Kordik, “The detectability of headtracker latency in virtual audio displays”, in *Proceedings of the 11th International Conference on Auditory Display*, July (2005).

[4] T. Lokki, “Physically-based Auralization -Design Implementation, and Evaluation”, Doctoral dissertation in Helsinki University of Technology, (2002).

[5] J.W. Scarpaci, H.S. Colburn, and J.A. White, “A system for real-time virtual auditory space.”, in *Proceedings of the 11th International Conference on Auditory Display*, July (2005).

[6] E.M. Wenzel, J.D. Miller, and J.S. Abel, “Sound Lab: A real-time, software-based system for the study of spatial hearing.” *Audio Engineering Society 108th Convention*, Pre-print, 5140 (2002).

[7] S. Yairi, Y. Iwaya, and Y. Suzuki, “Development of Virtual Auditory Display Software Responsive to Head Movement,” *Trans. Virtual Reality Society of Japan*, **11**(3) 437–446 (2006).

[8] A.V. Oppenheim and R.W. Schafer, *Discrete-time Signal Processing* (Prentice-Hall International, London, 1989).

[9] E.M. Wenzel, “Effect of increasing system latency on localization of virtual sounds with short and long duration”, in *Proceedings of the 2001 International Conference on Auditory Display*, July (2001).

[10] I. Toshima, H. Uematsu, T. Hirahara, “A steerable dummy head that tracks three-dimensional head movement: TeleHead”, *Acoust. Sci. & Tech.*, **24**(5), 327–329 (2003).

[11] P. F. Hoffmann and H. Møller, “Audibility of Spectral Switching in Head-Related Transfer Functions”, *Audio Engineering Society 119th Convention*, Convention paper, 6537 (2005).

[12] A. Kudo, H. Hokari, and S. Shimada, “A study on switching of the transfer functions focusing on sound quality,” *Acoust. Sci. & Tech.*, **26**(3), 267–278 (2003).

[13] B.C.J. Moore and B.R. Glasberg, “Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns,” *Hearing Research*, **28**, 209–225 (1987).

[14] M. Otani and S. Ise, “Fast calculation system specialized for head-related transfer function based on boundary element method,” *J. Acoust. Soc. Am.*, **119**(5), 2589–2598 (2006).