Virtual Speaker Tracker

Ki-hoon Shin, Sungmok Hwang, Youngjin Park
Korea Advanced Institute of Science and Technology, Daejeon, Korea
ks007b@kaist.ac.kr; tjdahr78@kaist.ac.kr; yjpark@kaist.ac.kr

Abstract

The term *virtualizer* is often used to refer to audio systems that convert multi-channel audio inputs to 2 ch. audio inputs by artificially creating a set of virtual speakers at selected positions around the listener to render a 5.1 ch. (or 7.1 ch.) home theatre system out of just 2 speakers in the front. A virtual speaker is usually generated via convolution with the head-related transfer function associated with the position of each surround speaker and appropriate filtering to eliminate cross-talk. Given a virtualizer incorporated in a TV or PC, however, it is difficult to assess if the system effectively places the virtual speakers at desired positions and thereby conclude which algorithm is "better" in terms of positioning capability. A subjective listening test involving many individuals is time consuming and costly. Therefore, it is necessary to devise a quantitative evaluation technique involving a dummy head microphone system rather than a group of individuals to provide a quick measure of speaker positions to audio engineers who develop virtualizer algorithms. In this paper, a method to track the azimuth of a virtual speaker from the output signals of a B&K HATS is proposed and the evaluation results on 3 selected virtualizer algorithms are shown.

Introduction

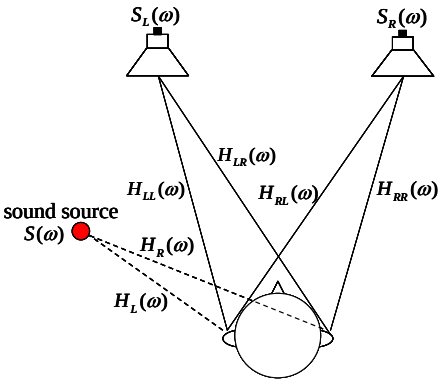
A virtualizer is an algorithm that replaces a multi-channel (5.1 or 7.1 ch.) audio system with a 2 ch. audio system by creating a set of virtual speakers at surround speaker locations through the binaural synthesis based on head-related transfer functions (HRTFs) as depicted in Figure 1 from the work of Hasegawa et al. (2000). The HRTF is a mathematical representation of a sound's transmission path from the source location to the listener's eardrum. In the left diagram of Figure 1, $S(\omega)$ is a source signal, $H_L(\omega)$ and $H_R(\omega)$ are the HRTFs from the source to the left and right ears of the listener, $H_{LL}(\omega)$ and $H_{LR}(\omega)$ are the HRTFs from the left loudspeaker, and $H_{RL}(\omega)$ and $H_{RR}(\omega)$ are the HRTFs from the right loudspeaker respectively. The right diagram of Figure 1 shows how the binaural signals $S_L(\omega)$ and $S_R(\omega)$ are processed in order to render a virtual sound image according to the following set of equations

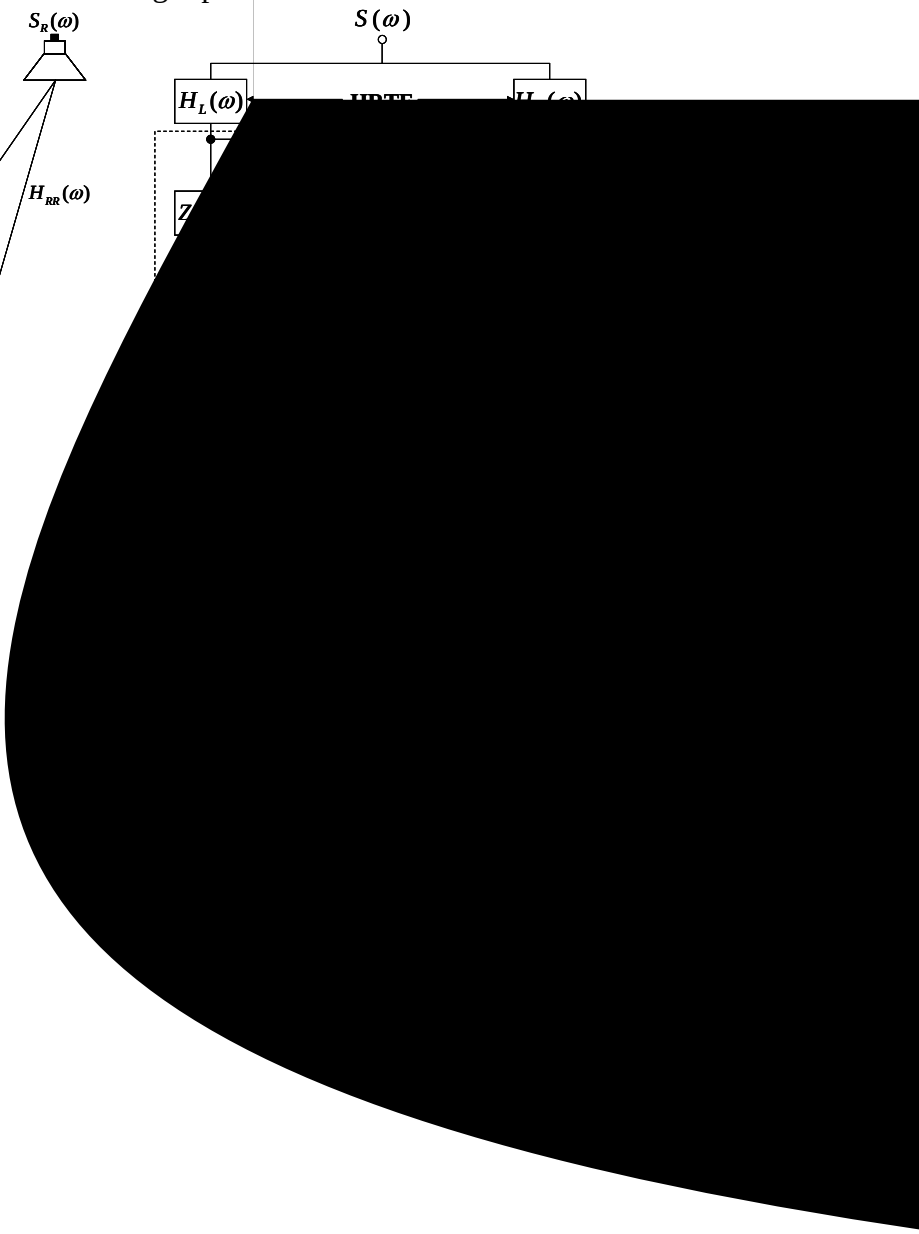
$$\begin{bmatrix} S_L \\ S_R \end{bmatrix} = S \begin{pmatrix} Z_{LL} & Z_{RL} \\ Z_{LR} & Z_{RR} \end{pmatrix} \begin{bmatrix} H_L \\ H_R \end{bmatrix}$$
$$\text{where } \begin{pmatrix} Z_{LL} & Z_{RL} \\ Z_{LR} & Z_{RR} \end{pmatrix} = \begin{pmatrix} H_{LL} & H_{RL} \\ H_{LR} & H_{RR} \end{pmatrix}^{-1}. \quad (1)$$

Z_{LL}, Z_{LR}, Z_{RL}, and Z_{RR} in the above equation constitute what's known as the cross-talk cancellation filter to equalize the sound traveling from one speaker to the ear on the opposite side.

The above cross-talk cancellation algorithm is built based on the assumption that the listener is not moving his/her head while listening, and the HRTF embedded in a typical virtualizer is the non-individualized HRTF measured from a dummy head microphone system. Therefore, the performance of a virtualizer depends on a number

of complicated aspects such as HRTF compatibility, speaker dynamics, and speaker arrangements. The HRTFs used in the virtualizer may not match those of each individual listener and the implemented cross-talk cancellation may fail in case the listener evades the proper sweet spot. Moreover, the original input tracks for surround speakers can be severely distorted both in phase and magnitude when filtered through the HRTF and cross-talk cancellation filter blocks. Therefore, an extensive series of subjective listening test is necessary in order to verify if the given virtualizer system properly deploys the virtual speakers in the right places.





angle between the left and right speakers was set at 30˚ and the distance from the TV panel to the listener (dummy head) was fixed at 1 m. A pink noise with a bandwidth covering the entire audible frequency range (20 Hz ~ 20 kHz) was used as the general input stimulus. For each virtualizer under test, all speakers including the virtual speakers were driven one at a time in order to evaluate the location of each speaker independently.

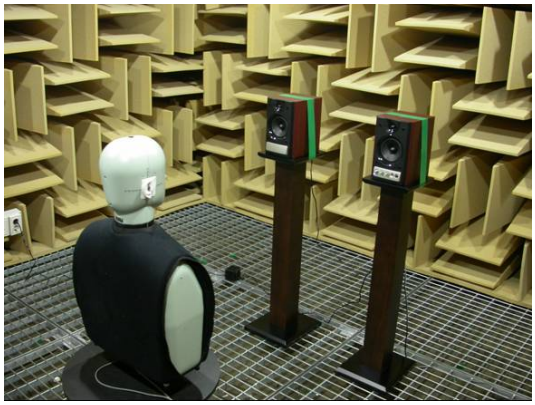


Figure 3. Experimental setting for virtualizer evaluation.

Human auditory localization depends primarily on the interaural time difference (ITD) and interaural level difference (ILD) that result from diffraction of incoming sound waves around the head and pinna (Rayleigh, 1905; Middlebrooks and Green, 1990). In addition to these interaural differences, the direction-dependent spectral notches and peaks in the HRTF caused by the filtering action of the pinna are thought to provide cues to sound source localization (Musicant and Butler, 1985). Of the primary sound localization cues, it is now believed by many researchers that ITD is the most dominant cue in determining the source location on the horizontal plane, and that ITD is encoded mostly by low frequency auditory neurons (Middlebrooks and Green, 1990). Wightman and Kistler (1991) also confirmed that the apparent direction of a sound source almost always followed the ITD cue as long as low frequencies were included in the testing stimuli through a carefully designed subjective listening test. In light of the above observations, our method to track each speaker in a given virtualizer system includes first finding the low frequency ITD from the output signals of the dummy head and then recovering the speaker angle from the ITD using a ray-tracing formula (RTF) which will be mentioned in detail in subsequent sections.

Low Frequency ITD Estimation

The ITD for each speaker was obtained from the low frequency (~ 1500 Hz) phase response of the interaural transfer function (ITF) between the output signals from the 2 ears of the dummy head using the following relationship

$$\mathrm{ITD} = -\frac{1}{2\pi}\frac{d\angle \mathrm{ITF}(f)}{df} \tag{2}$$

where $\angle \mathrm{ITF}$ and f are the phase response of ITF and frequency respectively. By introducing a linear phase function that best fits the measured phase response in the mean square sense, the group delay can be obtained from its slope as illustrated in Figure 4. This group delay can be substituted in the above equation to compute the ITD associated with each speaker.

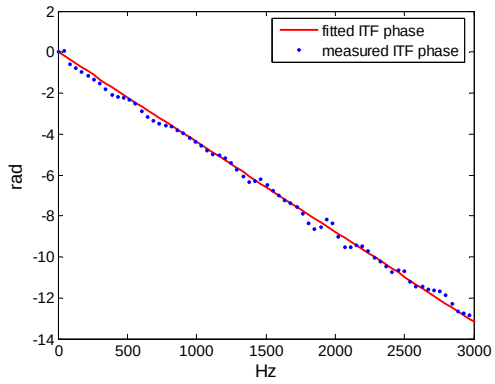


Figure 4. Measured ITF phase and its linear fit.

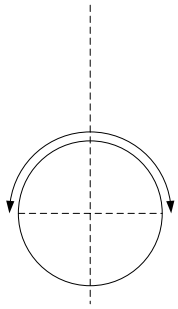
Angle Estimation from ITD

Once the ITD is computed for each speaker, the next step involves recovering the speaker angle from the ITD by the RTF of Woodworth and Schlosberg (1962) which describes a sound's propagation path around the head to the ears reasonably well for a source oriented at any angle on the horizontal plane. Figure 5 shows the illustration of the ray-tracing formulas both for a near source and a distant source. For a near source located at an angle θ with respect to the median plane bisecting the interaural axis, the sound propagates along the face of the head, which can be approximated as a sphere, to the ears and the difference in distance traveled by the waves reaching the 2 ears d will simply be

$$d = 2r\theta \tag{3}$$

where r is the radius of the head obtained by assuming that our sphere head model has an equal circumference with the head of the B&K HATS. On the other hand for a distant source, the waves from the source can be viewed as plane waves traveling straight to the ipsilateral (close) ear and around the head to the contralateral (far) ear after reaching the tangential point as illustrated in Figure 5. In this case, the difference in distance traveled by the waves will be

$$d = r(\theta + \sin\theta)\,. \tag{4}$$



in Figure 6. Although the distant source version of the RTF is known to provide a close fit to the ITD obtained from broadband stimuli as shown in the figure, the near source version seems to provide a better fit to the low frequency ITD and this is subject to a future study. As a result, the near source version of the RTF was used to compute the speaker angle. By multiplying the ITD from Equation (2) with the speed of sound (343 m/s) to obtain d, the speaker angle θ was computed from Equation (3).

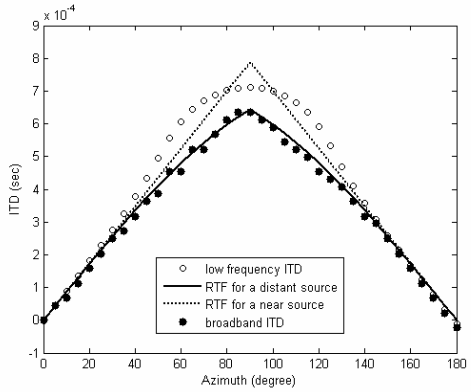


Figure 6. Low frequency ITDs vs. ITD estimates from the RTF.

Front-back Distinction Using HRTF Database

The speaker angle θ from Equation (3) can be dual because every source on the horizontal plane shares an equal ITD with its mirror image counterpart. However, we can resort to the HRTF database of the B&K HATS to resolve this problem since the pinna reacts differently for frontal sources and rear sources. The difference in pinna response from 2 kHz to 14 kHz in the ipsilateral HRTFs is quite salient for 2 sources at mirror image positions with respect to each other as illustrated in Figure 7.

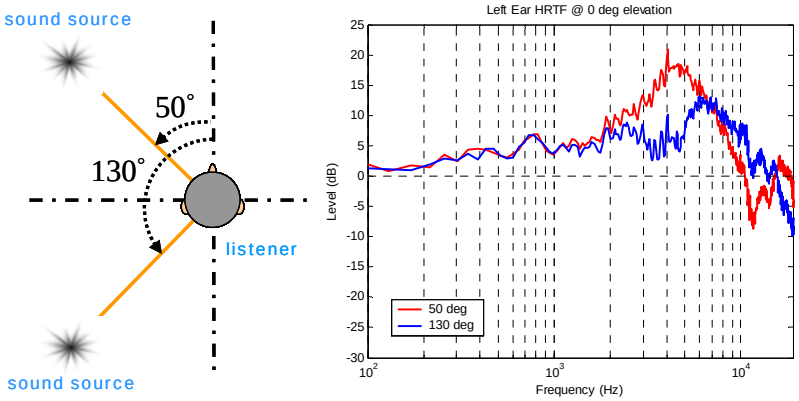


Figure 7. Left ear HRTFs for a source (at 50°) and its mirror image (at 130°).

Therefore, the speaker angle estimate θ can be redeemed from its front-back ambiguity simply by comparing the transfer function estimate $Y(f)$ between the target speaker and the ipsilateral ear to the set of ipsilateral HRTFs $H(f,\theta)$ and $H(f,180^\circ-\theta)$ from the HRTF database and minimizing the error between them in the mean square sense as follows (Shin and Park, 2005). The speaker is in the front hemisphere if $errf$ is smaller than $errb$ and the speaker is in the rear hemisphere if otherwise.

$$errf = \sum_{f=2\ kHz}^{14\ kHz} \left(|H(f,\theta)| - |Y(f)| \right)^2$$
$$errb = \sum_{f=2\ kHz}^{14\ kHz} \left(|H(f,180^\circ - \theta)| - |Y(f)| \right)^2 \quad (5)$$

Result and Conclusion

In this study, 3 different virtualizer algorithms of different makes were tested with the proposed evaluation method and the results are shown in Figure 8. The left and right real speakers L and R were always positioned 30˚ wide for all 3 virtualizer systems tested so the estimation results of the proposed evaluation method seem quite reasonable. Virtualizers B and C are shown to render the virtual speakers SL and SR at wider angles compared to virtualizer A and we can thus conclude that virtualizers B and C are better than virtualizer A in terms of source positioning capability.

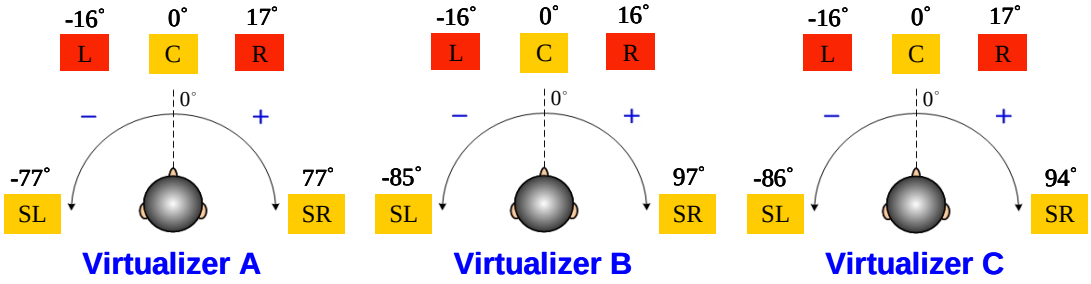


Figure 8. Tracking results for 3 virtualizer algorithms A, B, and C.

Acknowledgements

This work was supported by the Ministry of Science and Technology through the National Research Laboratory (NRL) program.

References

Hasegawa, H., Kasuga, M., Matsumoto, S., and Koike, A., 2000, “Simply Realization of Sound Localization Using HRTF Approximated by IIR Filter,” IEICE Trans. Fundamentals, Vol. E83-A, No. 6, pp. 973-978.

Middlebrooks, J.C. and Green, D.M., 1990, “Directional Dependence of Interaural Envelope Delays,” J. Acoust. Soc. Am., Vol. 87, pp. 2149-2162.

Musicant, A.D. and Butler, R.A., 1985, “Influence of Monoaural Spectral Cues on Binaural Localization,” J. Acoust. Soc. Am., Vol. 77, pp. 202-208.

Rayleigh, L., 1907, “On Our Perception of Sound Direction,” *Philosoph. Mag.*, Vol. 13.

Shin, K.H. and Park, Y., 2005, “Development of a Quantitative Evaluation Technique to Assess Virtual Audio Systems on Their Source Positioning Capability,” Proc. 34th Int. Congress and Exposition on Noise Control Engineering, Rio de Janeiro, Brazil.

Wightman, F.L. and Kistler, D.J., 1991, “The Dominant Role of Low-frequency Interaural Time Differences in Sound Localization,” J. Acoust. Soc. Am., Vol. 91, No. 3, pp. 1648-1661.

Woodworth, R.S. and Schlosberg, G., 1962, *Experimental Psychology*, Holt, Rinehard and Winston, NY, pp. 349-361.