ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Signal Processing in Acoustics
Session 2aSP: Array Signal Processing for Three-Dimensional Audio Applications II

2aSP4. Influence of height-channel contents on enhancing spatial impressions through virtually elevated loudspeaker

Sungyoung Kim*

*Corresponding author’s address: ECT Engineering Technology, Rochester Institute of Technology, ENT-2151, 78 Lomb Memorial Dr., Rochester, NY 14623, sxkiee@rit.edu

Reproducing early reflections related to ‘height’ information through elevated loudspeakers delivers enhanced presence of auditory images and integrates with a three-dimensional visual content homogeneously. Nonetheless, it is practically difficult for consumers to place loudspeakers required for the height-channel reproduction in a listening room. To overcome this limitation, many academic or commercial institutions propose various methods that render vertical sound images and reproduce them with smaller number of loudspeakers that are typically located in the horizontal plane. The rendered image then could deliver vertically extended impression of a sound field, which is likely related to listeners’ perception of enhanced presence. To better understand this relationship, this paper investigated idiosyncratic difference between one surround sound field with actual height channels and another with virtually elevated channels. The elicitation result revealed that listeners used four salient attributes - ASW, LEV, Powerfulness, and Clarity - to describe the difference. The result implicated that perceived magnitudes of those percepts were accounted for by a physical parameter, correlation coefficient between the elevated signal and the loudspeaker signal that is to feed to the closest loudspeaker in the horizontal plane.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Recent progress of digital audio technology has enabled consumers to enjoy multichannel audio contents in their home. With additional center and two rear channels to conventional stereo, the multichannel audio can provide listeners with enhanced precision of frontal imagery and spaciousness. Moreover, recent studies supported that height channel information could assist listeners in integrating a sound field with a three-dimensional visual content homogeneously [1] as well as providing immersive presence of a new space [2]. Several organizations already proposed extension of channels to 12 (by THX) or 24 (by NHK [3]) for consumer applications.

While these new formats provide listeners with enhanced spaciousness and impression of being fully immersed listening experience, it is practically difficult for a consumer to install loudspeakers in his or her ceiling. As an alternative, researchers investigated the binaural technology which can bring successful periphonic listening when reproduced via headphones [4]. However, for loudspeaker reproduction, the crosstalk component should be cancelled out [5][6].

The author proposed a new crosstalk cancellation method that constrains the left- and right-surround loudspeaker to reproduce binaural left- and right-signal, and lets the center channel loudspeaker to generate cancellation signals [7]. With this method, it is possible to minimize tonal modification caused by crosstalk canceling process. To summarize, the method allowed a listener to perceive elevated localization of sound sources.

This paper introduces the process of new crosstalk cancellation briefly, and presents subsequent subjective evaluation results on idiosyncratic difference between two reproduction conditions: (1) 5-channel plus two additional early reflection channels reproduced using the proposed method, and (2) same as (1) except that the two early reflections were reproduced via physical loudspeakers.

CROSSTALK CANCELLATION FILTERS

Let us denote that the pair of binaurally processed signals as $B_L$ and $B_R$, respectively. The goal of the implementation of two cancellation filters is to render a listener’s ear signals, $E_L$ and $E_R$ as equivalent to $B_L$ and $B_R$. We also denote signals at center, left-surround and right-surround loudspeaker as $C$, $LS$, and $RS$ respectively.

In addition, let us denote the transfer function from either the left- or right-surround speaker to the ipsilateral ear as $H_{ips}$, and to the contralateral ear as $H_{ctra}$, respectively.

To cancel the crosstalk, we need to design inverse filters, and let's denote these filters $f_1$ and $f_2$. As shown in Fig. 1, the binaural left ($B_L$) and right ($B_R$) signals are directly reproduced via the left- and right-surround speakers. These signals are then processed with one of inverse filters, $f_1$. From the center speaker, we reproduce a unknown crosstalk-cancelation signal, $C$, processed by another inverse filter, $f_2$. Then the left ear signals, $E_L$, will be:

$$E_L = f_1 * B_L * H_{ips} + f_1 * B_R * H_{ctra} + f_2 * C * H_c$$

Since our purpose is to make $E_L$ same as $B_L$, we can reduce the previous equation as follows:

$$B_L = f_1 * B_L * H_{ips} + f_1 * B_R * H_{ctra} + f_2 * C * H_c$$

(1)
FIGURE 1: Loudspeaker layout recommended by ITU-R BS 775.1. Three loudspeakers indicated with solid lines (center, left-, and right-surround) cancel the crosstalk component and generate a virtually elevated auditory image. The transfer function of a signal radiating from the one surround speaker to the ipsilateral ear of a listener is labelled as $H_{ips}$; to the contralateral ear as $H_{ctra}$. The transfer function from the center speaker to a listener’s ear is labelled as $H_c$. Two filters for crosstalk cancellation are denoted as $f_1$ and $f_2$ (the details of two filters are explained in the equation 2).

Solving this equation would generate solutions of $f_1$ and $f_2$ such that:

$$f_1 = \frac{1}{H_{ips} - H_{ctra}}$$

$$f_2 = -\frac{H_{ctra}}{H_c}$$

With filters $f_1$ and $f_2$, the three loudspeaker signals would be:

$$LS = f_1 * B_L$$

$$RS = f_1 * B_R$$

$$CT = f_2 * (LS + RS)$$

A subsequent task then was to implement two filters, $f_1$ and $f_2$. Each filter has a denominator which can be solved by implementing its inverse filter. We calculated these filters based on a frequency domain least-squares approximation.

SUBJECTIVE EVALUATION

Methods

In this section, we report the design of subjective evaluation that compared auditory images of the proposed method with auditory images of real loudspeakers.

In order to create binaurally processed stimuli, we convolved a monaural sound source with a Head Related Impulse Response (HRIR) measured by Gardner and Martin [8] with the target azimuth 90° and elevation 50°. The position along the frontal plane (also known as the coronal plane) was chosen based on the previous study results that loudspeakers located at those
positions “generate effective localization and height envelopment[9].” The authors also empirically found that sound sources elevated along the frontal plane effectively enhance spatial impression, compared to the ones along the median plane. Subsequently, the binaurally processed sound was filtered with $f_1$ and $f_2$ to create three loudspeaker signals ($CT$, $LS$, and $RS$) as expressed in the equation 3.

The listening experiment was conducted at the Yamaha Corporation’s Auralization Experiment Room, located in Iwata, Japan, the dimensions of which are 7.2 m wide, 7.4 m long, and 3.8 m high. For the reproduction of virtually elevated sound, three speakers were attached along the horizontal plane (at a height of 1.1 m) with the azimuths of $0^\circ$, $110^\circ$, and $250^\circ$. For comparison, the real loudspeakers was placed at the azimuth of $90^\circ$ (counterclockwise) at elevation of $50^\circ$ (one of loudspeakers located along the frontal plane) as shown in the bottom panel of Fig. 2.

![Figure 2](image-url.com)

**Figure 2:** (Left panel) Top view of the experiment room. Three loudspeakers were mounted to a semi-spherical metal frame with the azimuths of center, left-surround, and right-surround speakers in the ITU-R 5.1 layout. The metal frame consists of multiple layers of circles and bars, represented as dotted lines in this figure. (Right panel) Rear view of the Auralization Experiment Room. On the left side bar, four speakers were attached with the azimuth of $90^\circ$ (counterclockwise) and the elevations of $0^\circ$, $30^\circ$, $50^\circ$, and $70^\circ$, respectively.

The listening experiment consisted of two parts: identifying salient attributes and quantifying magnitude of identified attributes for analysis of relationship across the reproduction conditions. For this experiments, we used a classic piano sound captured with multiple microphone including height channels. The stimulus was a 7-channel sound clip, which contains conventional 5-channel plus additional two height channels. The condition where height channels were reproduced via real height-speakers was denoted as R7 and via virtual speakers as V7 respectively. We created four 30-second long clips that represented various sound pressure levels and spectra.

**Results**

As the first step, four expert panelists elicited salient attributes which differentiated sound fields with real height channel from virtual height channels. The used method was a consensus vocabulary technique [10, p.44]. After in-depth discussion, four panelists selected four descriptors related to salient perceptual attributes of the given stimuli. The selected descriptors and their definitions were:
- Apparent Source Width (ASW): the apparent horizontal and vertical spatial extent of the sound sources.
- Listener Envelopment (LEV): the degree of spatial impression of sound sources that envelops and surrounds around a listener.
- Powerfulness: the degree of integration of spatial and tonal impression of sound sources that enhances a listener’s impact associated with the sound field.
- Clarity: the apparent integration of the sound sources into a single unified image.

While identification and definition of perceptual attributes were conducted with selected expert panelists, more participants (total of seventeen subjects) participated in the following quantification experiment. All subjects are researchers conducting audio and/or acoustics aging from 27 to 57. Although no formal audiological test was made on participants, all participants have normal hearing to conduct everyday life and acoustic-related research activities. They were asked to choose one between R7 and V7 condition, which gave higher perceived magnitude for each percept.

Fig. 3 shows the sum of pair-wise choice for 17 subjects of Stimulus 1 and Stimulus 2, which implicated that R7 condition not always delivered increased magnitude of the elicited attributes. V7 condition delivered higher magnitudes of attributes to listeners for Stimulus 2; more evaluated that powerfulness (denoted as “Pow” in the figure), ASW, and LEV of V7 had stronger perceived magnitude than R7.

![Figure 3: Sum of pair-wise choices between real (R7) and virtual (V7) height-channel condition on four elicited attributes. While more subjects perceived stronger Powerfulness (denoted as Pow), ASW, and LEV of R7 condition for the Stimulus 1, the result of Stimulus 2 (right panel) were opposite.](image)

These results might suggest several points regarding V7 condition, especially related to physical properties of sound sources. First, so-called height signals (radiated from either physically or virtually elevated speaker) were captured via microphones above the microphones for front channels (left-, right- and center-channel). Therefore, the height signals happened to be correlated to the front-channel signals. And as explained previous section, some portions of virtual height channels are reproduced via the center loudspeaker. This mixture of the original center signal and virtual height signals might cause listeners to perceive change in loudness and localization, especially when the height signals are highly correlated with the center signal. Thus, this loudness variation might affect powerfulness of the reproduced sound fields. When correlation was small, subjects perceived as being enveloped (bigger LEV), while high correlation caused subjects to feel narrow and loud sound field. This observation implicated that an ideal virtual 7-channel sound could be manipulated through a systematic analysis on
physical characters of original 7-channel content, which is the current research topic of the author. Based on this finding, the author plans to investigate and unfold perceptual attributes that deliver listeners an impression of immersed presence and a corresponding signal rendering method.

**CONCLUSION**

In order to render a binaural signal in a conventional 5-channel reproduction system, we have devised a signal processing method which cancels the crosstalk signal using the center, left-, and right-surround speakers. We subsequently compared the sound field with virtually elevated height channels and the now with actual height channels when height-channel contents were reflections from the ceiling. The listening test results showed that four perceptual attributes - ASW, LEV, Powerfulness, and Clarity - differentiated the 3D auditory images with virtually elevated channels from their counterparts. Moreover, perceived magnitude of those attributes were dependent to inter-channel relation of horizontal surround contents. This implies that a wider, more enveloping and powerful, and clearer sound field utilizing virtual height channels could be rendered via unfolding optimal degree of difference between height components and horizontal components.

**REFERENCES**


