Blind Depth Estimation Based on Primary-to-Ambient Energy Ratio for 3-D Acoustic Depth Rendering

Se-Woon Jeon*, Dae-Hee Youn* and Young-Cheol Park†
* Electrical and Electronic Eng., Yonsei University, Seoul, Korea
E-mail: jdotsw@dsp.yonsei.ac.kr, dhyoun@yonsei.ac.kr
† Computer and Telecommunication Eng., Yonsei University, Wonju, Korea
E-mail: young00@yonsei.ac.kr

***Abstract*—Since the advent of 3-D video, the acoustic depth rendering for the proximity effect has been an issue of great interest. In this study, we propose an algorithm for estimating acoustic depth cues from stereo audio signal, without a *priori* knowledge about the source-to-listener geometry and room environments. We employ the principal component analysis (PCA) to estimate the acoustic depth based on the primary-to-ambient energy ratio (PAR) which is related with the front-back movement of the sound source. And for the acoustic depth rendering, the distance variation of the sound source is parameterized through tracking the estimated depth cue. The proposed estimation algorithm was evaluated using stereo audio clips extracted from a real 3-D movie, and the results confirmed effectiveness of the proposed acoustic depth estimation algorithm.**

I. Introduction

Recently, 3-D video technologies that can generate the *proximity effect* of visual objects have drawn great attention [1]. By the optical proximity effect, the user can virtually perceive that visual objects are pulled toward him or her from the video screen. To provide more realistic 3-D effect to the user, however, the depth perception of the audio sound corresponding to the video objects being controlled also need to be controllable. It is about the distance alignment problem between the visual object and its sound effect.

The acoustic depth rendering (ADR) is a 3-D audio technique that can create the distant sound effect for the listener. Recently, several ADR algorithms were proposed and their proximity effects were tested via listening tests [2]-[5]. In [5], we proposed a stereophonic loudspeaker system for the acoustic depth control based on two kinds of acoustic parameters; interchannel phase difference (ICPD) and primary-to-ambient energy ratio (PAR). The proposed algorithm creates the proximity effect of the sound source in a distance range between the loudspeakers and the listener. And it can provide users with the perception of the sound image instantly moving front-back by controlling the acoustic parameters. In order to provide the acoustic depth perception, the depth information comprised in the original stereo signal should be estimated a *priori*, because the distance of the sound source should be perceived according to the front-back movement of the visual object in the 3-D scene.

The starting point of the acoustic depth estimation is a knowledge about the distance hearing and acoustic distance cues [6]. Most frequently used distance cues are reverberant cues such as direct-to-reverberant energy ratio (DRR) and early decay time (EDT), because they comprise distance information about the auditory event and the room environment. And more, they are relatively easy to artificially create or to properly adjust for the acoustic distance rendering [7], [8].

Although the reverberant cues are most dominantly related to the distance perception, application of the reverberant cues to the general audio signal is limited. Because the reverberant cue is an absolute cue that is not only determined by the distance of the sound source, but also the specific room environment [7]. Therefore, without a *priori* information or training data about the room environment, the use of reverberant cues is impractical [9], [10]. In most of 3-D multimedia contents, moreover, the sound distance should be estimated from the ready-mixed audio signal. Thus, in the acoustic rendering process, it should be conducted blindly without any side information.

Unlike the distance perception, the depth perception is defined as an overall front-back distance perception of the audio scene or the sound source [6]. Similar to the distance perception, the depth perception is also affected by the room environment. But in a short time duration, it mostly depends on the front-back movement of the sound source, rather than specific properties of the room environment. In real 3-D multimedia contents, it is more plausible to track the front-back movement of the sound source than to find absolute distance without a *priori* knowledge about the room environment of the scene. Therefore, we focused on the depth estimation of the sound source and its usefulness for the acoustic depth rendering.

In this paper, we propose the blind depth estimation (BDE) and the parametric representation of the estimated depth to use it for the effective 3-D acoustic depth control which can be resulted in the acoustic distance perception. Based on the acoustic depth perception, the proposed algorithm tracks the front-back movement of the sound source. To this end, we first obtain a primary-to-ambient energy ratio (PAR) using the

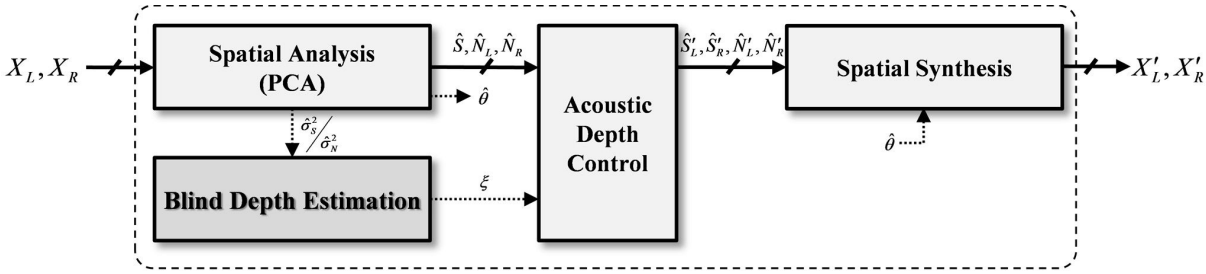


Fig. 1. Block-diagram of the acoustic depth rendering for the stereophonic loudspeaker system.

principal component analysis (PCA) from the stereo signal [12]. Then, the front-back distance of the sound source is estimated by the difference between the minimum and the maximum PARs within a predefined time window. For the parametric representation of the acoustic depth cue, we define a normalized distance given as a function of time-varying PAR, which represents the distance of the sound source in a range between the loudspeakers and the listener in aspects of the proximity effect.

II. Acoustic Depth Rendering

Fig. 1 shows a block-diagram of the acoustic depth rendering algorithm for a stereophonic loudspeaker system [5]. First, to estimate spatial cues and also to separate a primary source from ambient components, the spatial analysis based on PCA is performed [12], [13]. PCA is popularly employed for the multichannel signal processing and its performance is addressed in many previous studies.

In spatial analysis, the azimuthal direction, θ, of the primary source and the PAR, σ_S^2/σ_N^2, are estimated from the original stereo signal. The PAR is delivered to the acoustic depth control process and used for the control of depth perception. The acoustic depth control is based on psychoacoustic properties of the spatial hearing of the human [6], [11]. The interchannel phase difference and/or energy difference can be used as a control methodology. Finally, depth-controlled signals are synthesized in spatial synthesis using the directional information extracted in spatial analysis.

The acoustic distance rendering is still very challenging because none of the previous methods can provide a scalable controllability of the depth perception for the ready-mixed audio signal. The details of the acoustic depth rendering system described in Fig. 1 can be found in [5], and its theoretical foundation about the acoustic depth perception can be also found.

In this paper, we focus on the blind depth estimation and the parametric representation of the estimated cues. The proposed algorithm obtains the depth information from the stereo signal without any a *priori* knowledge about source-to-listener geometry and room acoustics.

III. Blind Depth Estimation

A. Primary-to-Ambient Energy Ratio Estimation

In general, to estimate DRR, that is a dominant distance cue in a room listening environment, the audio signal is deconvolved, and then, the room impulse response is segregated into direct and reverberant parts. But deconvolution is a complex task and it is difficult to estimate accurate long-term impulse response. In case of the multichannel signal, the difference of the statistic properties can be used for the direct-to-reverberant source decomposition [10]. Direct and early parts of the channel signal are highly correlated, while late part is uncorrelated and has less coherence [11]. For the same purpose, in spatial analysis of Fig. 1, the stereo input is decomposed into the primary source and the ambient components using PCA, and respective energy ratio of them is estimated through eigenvalues decomposition [5], [13].

In PCA, eigenvalues are first calculated from elements of a covariance matrix R of the stereo signal given by

$$R = \begin{bmatrix} r_{LL} & r_{LR} \\ r_{RL} & r_{RR} \end{bmatrix}. \tag{1}$$

Then, eigenvalues are expressed as

$$\lambda_i = 0.5\Big\{r_{LL}+r_{RR}\pm\sqrt{(r_{LL}-r_{RR})^2+4r_{LR}r_{RL}}\Big\}, \quad i=0,1. \tag{2}$$

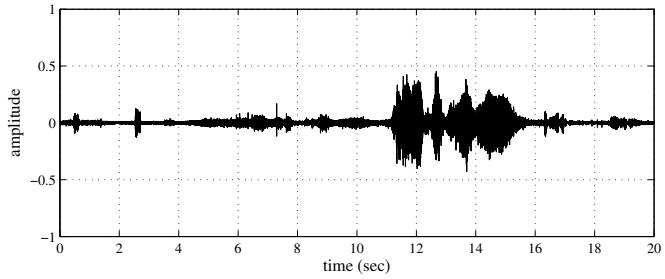
The eigenvalues represent the energies of the signal components included in the mixed channel signals [5], [12]. From (2), the energy estimates of the primary source S and the ambient component N can be easily obtained as

$$\hat{\sigma}_S^2 = \lambda_0 - \lambda_1 \quad \text{and} \quad \hat{\sigma}_N^2 = \lambda_1. \tag{3}$$

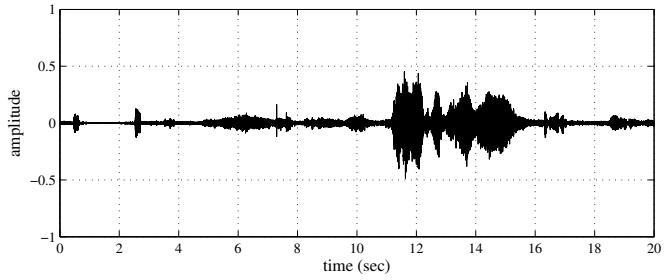
Thus, the energy ratio of the primary source and the ambient component, PAR, can be estimated as

$$\text{PAR} = \frac{\hat{\sigma}_S^2}{\hat{\sigma}_N^2} = \frac{\lambda_0 - \lambda_1}{\lambda_1}. \tag{4}$$

Since PAR represents the energy ratio between the correlated and the uncorrelated signal components in the stereo signal, it has high similarity to DRR. Thus, PAR also has close relationship with the distance to the primary source, and furthermore, time-variation of the sound source distance can be obtained by analyzing the time-variation of PAR. In general, the proximity effect of target object in 3-D multimedia contents sustains over hundreds of milliseconds. Thus, the variation of PAR should be watched over a time window longer than the time duration which the room characteristics are constants.



(a) left signal, x_L



(b) right signal, x_R

Fig. 2. Stereo input signals; the audio clip example of the movie "Avatar (2009)".

B. Blind Depth Estimation

A number of factors can contribute to the depth perception. The term *depth* can be defined as overall front-back distance perception of a scene or individual sources [6]. In most cases, the movie scene changes unpredictably and therefore, stationary of the sound scene cannot always satisfied. Therefore, we use several assumptions for the ready-mixed audio signal in order to obtain plausible blind depth estimation method:

- *Room dimension and spatial properties are fixed in the limited time duration.*
- *Depth perception is determined by front-back distance perception in the limited time duration.*
- *Reverberant cue corresponding to the distance of sound source, e.g. direct-to-reverberant energy ratio, is only affected by front-back movement of the sound source.*

Under these assumptions, we first define a depth cue factor obtained by the maximum difference between PARs within a time window as given as

$$\eta(k,b) = \max_{m}\{\text{PAR}(k,b)\} - \min_{m}\{\text{PAR}(k,b)\}, \quad \text{with } m \subset [k-N+1,\, k], \tag{5}$$

where k and b, respectively, are the time frame and frequency subband indices, N is the number of previous time frames within a time window for tracking the front-back movement of the sound source. The depth factor is defined using PARs which are parameters related to the absolute distance. Though, since the front-back movement of the sound source creates the depth perception, the depth estimate can be conducted by tracking PARs. If there is a front-back movement within the watched window, the depth factor will have a greater value than certain threshold. On the contrary, if the sound source stays at a location, the depth factor will be close to one.

In aspects of creating the proximity effect, the acoustic depth should be represented as a relative parameter that can be mapped onto a location in the distance range between the loudspeakers and listener. Thus, we define a normalized distance cue factor for the parametric representation of the distance perception, which is given by

$$\xi(k,b) = \big(\text{PAR}(k,b) - \min_{m}\{\text{PAR}(k,b)\}\big) \Big/ \eta(k,b), \quad \text{with } m \subset [k-N+1,\, k]. \tag{6}$$

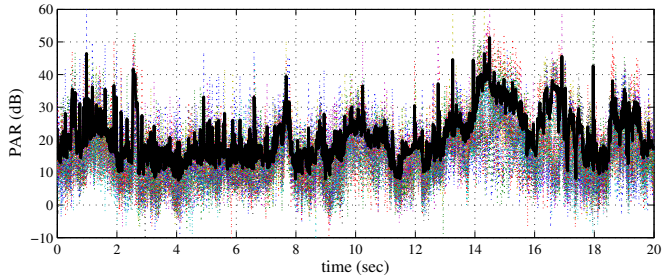
Equation (6) provides a normalized value from zero to one, which represents the perceived distance of the sound source. When the perceived distance changes far-to-near from the listener, the normalized parameter ξ increases from zero to one, and vice versa. Now, the normalized distance can conveniently represent the relative distance between the loudspeakers and listener.

IV. Evaluation and Discussion

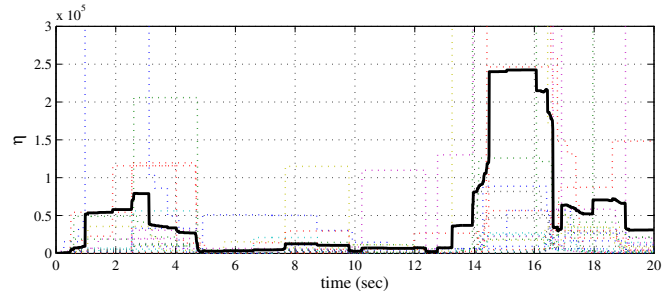
The performance of the proposed algorithm was evaluated using stereo audio clips extracted from the real 3-D movie without any side information regarding the acoustic scene. Fig. 2 shows an example of the stereo input signals extracted from the movie scene in which *one archaeopteryx suddenly appears through a dense thicket toward a hero*. The primary sound requiring the acoustic depth control is a cry of the animal and the ambient sounds are composed of a background noise and other sound effects.

First, using PCA, the PAR was calculated in each critical subband and the obtained PARs are presented in Fig. 3-(a), where PARs of all bands are displayed. It was assumed that only one dominant source exists in each band [5]. However, PARs indicate that it is not clear to identify the movement of sound source, although a general trend is visible. A thick line indicates mean values of all subbands. Now, the maximum PAR period is visible, which is between fourteen and fifteen seconds. Although the estimated PAR somewhat reflects the distance of the primary source, it is still not suitable to identify depth cue due to the estimation noises of PARs along the time frames.

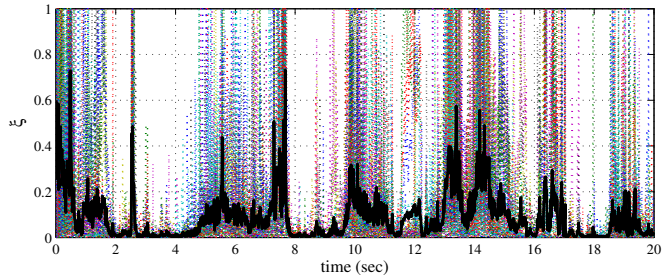
In the example of Fig. 2, the front-back movement of the primary source instantaneously happens in about two seconds. Therefore, to estimate the acoustic depth cue, the number of previous time frames N was set as 200 when a frame size was 1024 with 50% frame overlap. Fig. 3-(b) shows acoustic depth cues obtained by calculating the difference between the minimum and the maximum PARs of (5). Now, it represents the dynamic movement range of the primary source more definitely than the PAR results of Fig. 3-(a). Finally, in Fig. 3-(c), the normalized distance results of (6) in the critical subbands are shown. It is based on the variation of the primary-to-ambient energy ratio and it represents the perceived distance variation in aspects of the proximity effect.



(a) primary-to-ambient energy ratio, PAR(k, b)



(b) acoustic depth cue, $\eta(k, b)$



(c) normalized distance parameter, $\xi(k, b)$

Fig. 3. Primary-to-ambient energy ratio (PAR), acoustic depth cue, and normalized distance parameter in the critical subbands of time frames (k, b) for the audio clip in Fig. 2; dotted lines represent estimates of each subbands and a thick line represents mean value in all subbands.

The tracking equations of (5) and (6) efficiently represent the variation of the perceived depth and distance. However, in the dynamic case, that is a scene in which the sound image moves fast and the room environment frequently changes, the performance of the depth estimation can be degraded due to the low estimate speed. For this reason, the number of the previous frames N should be carefully chosen corresponding to the properties of the acoustic scene. In general, less number of N is more suitable for the dynamic scene. The proper method to determine the number of N is still not defined in this study, but we experimentally found that the choice of the time duration from about one to five seconds via N could satisfy most of the dynamic scene cases.

V. Conclusions

Distance hearing is a hot issue with respect to the sound localization for 3-D audio. Until now, many researches about the acoustic distance perception of the human have been studies. Direct-to-reverberant energy ratio is one of the dominant cue for the acoustic distance estimation and also the acoustic depth rendering. However, it is difficult to estimate the accurate distance of the sound source without a *priori* information of the source-to-listener geometry and the specific room environment, because the direct-to-reverberant energy ratio is an absolute cue that is only available in the same acoustic environment. Recently, the depth perception is also focused in aspects of the proximity effect of 3-D multimedia contents. However, similar to the distance perception, since the conventional audio formats do not provide the distance or depth information, it is difficult to properly control the depth perception in the sound reproduction. In this paper, we focused on how to blindly estimate the acoustic depth information for the use of the acoustic depth rendering and its adjustment. PCA was used to extract the primary-to-ambient energy ratio which represented the perceived distance as like the direct-to-reverberant energy ratio. And the acoustic depth cue was estimated by the tracking equation of the front-back distance based on the variation of the primary-to-ambient energy ratio. In addition, the perceived distance was parameterized by the normalized distance which described the sound source was perceived at proximate or furthermore location between the loudspeakers and the listener. The proposed blind depth estimation algorithm can provide the more immersive effect of the 3-D multimedia contents and more, contribute the distance alignment between the visual object and its sound effect without a *priori* information.

References

[1] L. Zhang and W. J. Tam, "Stereoscopic image generation based on depth images for 3D TV," *IEEE Trans. on Broadcasting*, vol. 51, no. 2, Jun. 2005, pp. 191-199.

[2] H.-G. Moon, "Auditory depth control using reverberation cue in virtual audio environment," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E91-A, no. 4, Apr. 2008, pp. 1212-1217.

[3] A. Härmä, S. van de Par, and W. de Bruijn, "On the use of directional loudspeakers to create a sound source close to the listener," *AES 124th Conv.*, Amsterdam, The Netherlands, May 2008.

[4] S. Koyama, Y. Hiwasaki, K. Furuya, and Y. Haneda, "Inverse wave propagation for reproducing virtual sources in front of loudspeaker array," *19th European Signal Processing Conference (EUSIPCO 2011)*, Barcelona, Aug.-Sep. 2011.

[5] S.-W. Jeon, Y.-C. Park, and D.-H. Youn, "Acoustic depth rendering for 3D multimedia applications," *IEEE 30th Int. Conf. on Consumer Electronics (ICCE 2012)*, Jan. 2012.

[6] F. Rumsey, *Spatial Audio*. Focal Press, 2001.

[7] S. H. Nielsen, "Auditory distance perception in different rooms," *J. Audio Eng. Soc.*, vol. 41, no. 10, Oct. 1993, pp. 755-770.

[8] J. Michelsen and P. Rubak, "Parameters of distance perception in stereo loudspeaker scenario," *AES 102nd Conv.*, Munich, Germany, Mar. 1997.

[9] S. Vesa, "Binaural sound source distance learning in rooms," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 8, Nov. 2009, pp. 1498-1507.

[10] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, Sep. 2010, pp. 1793-1805.

[11] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1997.

[12] I. T. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 2002.

[13] M. M. Goodwin and J.-M. Jot, "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2007.