Angle-Based Virtual Source Location Representation for Spatial Audio Coding

Seungkwon Beack, Jeongil Seo, Hangil Moon, Kyeongok Kang, and Minsoo Hahn

ABSTRACT—Virtual source location information (VSLI) has been newly utilized as a spatial cue for compact representation of multichannel audio. This information is represented as the azimuth of the virtual source vector. The superiority of VSLI is confirmed by comparison of the spectral distances, average bit rates, and subjective assessment with a conventional cue.

Keywords—Spatial audio coding (SAC), binaural cue coding, multi-channel audio.

I. Introduction

Spatial audio coding (SAC) is a technique to compress multi-channel audio signals with a high compression ratio. Recently, binaural cue coding (BCC) has been introduced and has become a representative scheme for SAC [1]. BCC represents multi-channel signals as a mono downmix plus side information representing spatial cues of side information such as the inter-channel level difference (ICLD), inter-channel time difference, and inter-channel coherence. The ICLD plays a pivotal role to remove a lot of redundant information through an approximation of the original signal spectral information. The ICLD, however, can be easily distorted through a quantization process due to the coarse quantization and dynamic range limitation to achieve the desired low bitrate. To solve this problem, some modification techniques have been applied to the SAC [2], [3].

Instead of the ICLD, in this letter, virtual source location information (VSLI) is presented as a spatial cue. Our representation is confined to a multi-channel (5-channnel) case to cope with MPEG-4 SAC [4]. VSLI is analyzed on a semicircle plane and represented as an angle. A spectral distortion measurement is conducted to confirm the usefulness of our VSLI as an approximation of original information with robustness to quantization distortion.

II. VSLI Analysis

VSLI estimation is essentially to approximate the original signal spectral information. The analysis layout in this letter is for multichannel (5-channnel) cases that consist of center (C), left (L), right (R), left surround (Ls), and right surround (Rs) signals. And it is basically assumed that each channel signal is virtually localized on the semicircle plane as shown in Fig. 1(a). Since the main role of VSLI is to estimate each channel power, the analysis layout on the semicircle provides enough information to estimate each channel power with the given four angles.

The frame-based input signals are transformed into the frequency domain, and the VSLI is estimated for a partition of the frame. Our partition and analysis window type are identical to [1]. The magnitude at partition b is estimated as

$$\mathrm{M}_{ch,b} = \sum_{n=B_b}^{B_{b+1}-1} \left|S_{ch,n}\right|, \qquad (1)$$

where $S_{ch,n}$ is the spectral coefficient of channel ch denoted as one of either C, L, R, Ls, or Rs. Boundary B_b is for a partition boundary identical to those of BCC as in [1]. The VSLI is estimated with respect to the sections of the semicircle plane as shown Fig. 1, where five azimuth values, that is, VSLI cues per partition, are extracted. These azimuths are defined as global

Manuscript received Aug. 29, 2005; revised Feb. 09, 2006.

Seungkwon Beack (phone: +82 42 860 1745, email: skbeack@etri.re.kr), Jeongil Seo (email: seoji@etri.re.kr), and Kyeongok Kang (email: kokang@etri.re.kr) are with Digital Broadcasting Research Division, ETRI, Daejeon, Korea.

Hangil Moon (email: fullmoon@acoustics.snu.ac.kr) is with Digital Media Center, Samsung Electronics Co., Ltd., Suwon, Korea.

Minsoo Hahn (email: mshahn@icu.ac.kr) is with Multimedia Group, Information and Communications University, Daejeon, Korea.

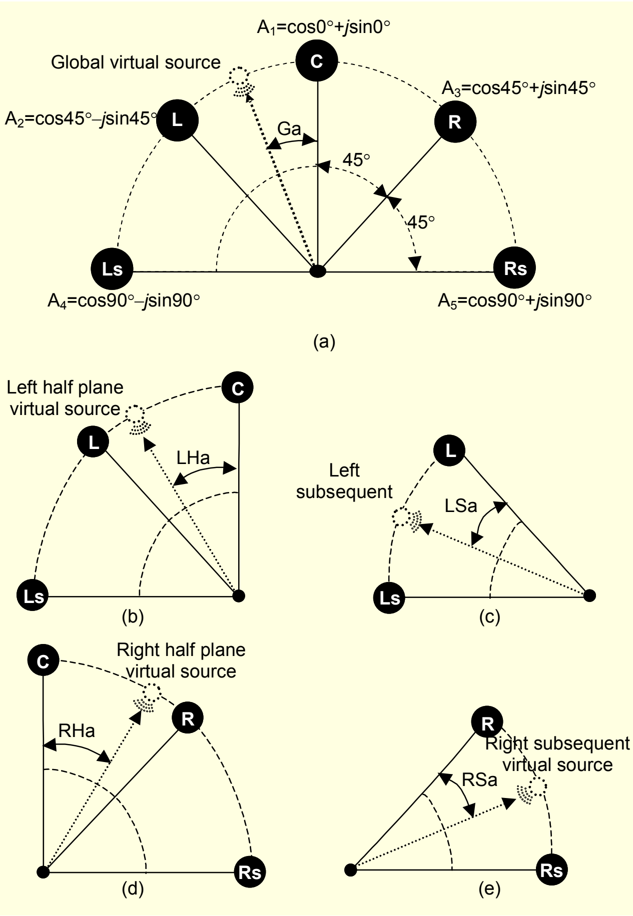


Fig. 1. Representation of virtual source location information: (a) global virtual source vector on the semicircle plane, (b) left-half plane virtual source vector, (c) left-subsequent virtual source vector, (d) right-half plane virtual source vector, and (e) right-subsequent virtual source vector.

angle (Ga), the left-half plane angle (LHa), right-half plane angle (RHa), left subsequent angle (LSa), and right subsequent angle (RSa). Each azimuth is obtained from the virtual source location vector. To estimate the vector, we adopt the vector base amplitude panning (VBAP) in [5]. First, the global vector Gv_b at partition b can be estimated as

$$\begin{aligned}\mathrm{Gv}_b &= \mathrm{A}_1\times\mathrm{M}_{\mathrm{C},b}+\mathrm{A}_2\times\mathrm{M}_{\mathrm{L},b}+\mathrm{A}_3\times\mathrm{M}_{\mathrm{R},b}\\ &\quad+\mathrm{A}_4\times\mathrm{M}_{\mathrm{Ls},b}+\mathrm{A}_5\times\mathrm{M}_{\mathrm{Rs},b}\,.\end{aligned} \tag{2}$$

A_j(j = 1,…,5) is the coordinate of the channel layout on the semicircle plane as shown in Fig. 1(a). The global angle Ga_b at partition b is straightforwardly obtained from (Gv_b). Other vectors, LHv_b, RHv_b, LSv_b, and RSv_b can also be estimated as follows :

$$\mathrm{LHv}_b = \mathrm{A}_1\times\mathrm{M}_{\mathrm{C},b}+\mathrm{A}_2\times\mathrm{M}_{\mathrm{L},b}+\mathrm{A}_4\times\mathrm{M}_{\mathrm{Ls},b}, \tag{3}$$

$$\mathrm{RHv}_b = \mathrm{A}_1\times\mathrm{M}_{\mathrm{C},b}+\mathrm{A}_3\times\mathrm{M}_{\mathrm{R},b}+\mathrm{A}_4\times\mathrm{M}_{\mathrm{Rs},b}, \tag{4}$$

$$\mathrm{LSv}_b = \mathrm{A}_2\times\mathrm{M}_{\mathrm{L},b}+\mathrm{A}_4\times\mathrm{M}_{\mathrm{Ls},b}, \tag{5}$$

and

$$\mathrm{RSv}_b = \mathrm{A}_3\times\mathrm{M}_{\mathrm{R},b}+\mathrm{A}_5\times\mathrm{M}_{\mathrm{Rs},b}\,. \tag{6}$$

Similarly, LHa_b, RHa_b, LSa_b, and RSa_b are the angles of (3), (4), (5), and (6), respectively, as illustrated in Fig. 1. According to the Ga_b position, the type of transmitted angle information per partition varies. Namely, when Ga_b is in the left-half plane, {Ga_b, RHa_b, RSa_b, LSa_b } is estimated and transmitted as the side information. Otherwise, it is changed into {Ga_b, LHa_b, LSa_b, RSa_b }.

III. VSLI Synthesis

The main purpose of VSLI synthesis is to convert the transmitted angle information into power gain factors. The constant power panning (CPP) law [6] is adopted to obtain the gain factors per partition of the channels. Different synthesis processes are applied according to the Ga_b position. The VSLI is used to estimate the inverse panning angles, $\theta_1,\theta_2,\theta_3,\theta_4$, in Table 1. From the inverse panning angles, the power gain factors can be calculated as shown in Table 2, and can be obtained straightforwardly from [6].

Table 1. Power panning angle calculation.

$\mathrm{Ga}_b \geq 0$	$\mathrm{Ga}_b < 0$
$\theta_1 = \left(\dfrac{\mathrm{Ga}_b-\mathrm{LHa}_b}{\mathrm{RSa}_b-\mathrm{LHa}_b+\delta}\right)\times\dfrac{\pi}{2}$	$\theta_1 = \left(\dfrac{\mathrm{Ga}_b-\mathrm{RHa}_b}{\mathrm{LSa}_b-\mathrm{RHa}_b+\delta}\right)\times\dfrac{\pi}{2}$
$\theta_2 = \left(\dfrac{\mathrm{LHa}_b-\mathrm{LSa}_b}{0-\mathrm{LHa}_b+\delta}\right)\times\dfrac{\pi}{2}$	$\theta_2 = \left(\dfrac{\mathrm{RHa}_b-\mathrm{RSa}_b}{0-\mathrm{RSa}_b+\delta}\right)\times\dfrac{\pi}{2}$
$\theta_3 = \left(\dfrac{\mathrm{LSa}_b+\pi/2}{-\pi/4+\pi/2}\right)\times\dfrac{\pi}{2}$	$\theta_3 = \left(\dfrac{\mathrm{RSa}_b-\pi/2}{\pi/4-\pi/2}\right)\times\dfrac{\pi}{2}$
$\theta_4 = \left(\dfrac{\mathrm{RSa}_b-\pi/2}{\pi/4-\pi/2}\right)\times\dfrac{\pi}{2}$	$\theta_4 = \left(\dfrac{\mathrm{LSa}_b+\pi/2}{-\pi/4+\pi/2}\right)\times\dfrac{\pi}{2}$

Table 2. Channel power gain factor calculation.

$\mathrm{Ga}_b \geq 0$	$\mathrm{Ga}_b < 0$
$F_{\mathrm{C,b}} = \cos(\theta_1)\sin(\theta_2)$	$F_{\mathrm{C,b}} = \cos(\theta_1)\sin(\theta_2)$
$F_{\mathrm{L,b}} = \cos(\theta_1)\cos(\theta_2)\sin(\theta_3)$	$F_{\mathrm{L,b}} = \sin(\theta_1)\sin(\theta_4)$
$F_{\mathrm{Ls,b}} = \cos(\theta_1)\cos(\theta_2)\cos(\theta_3)$	$F_{\mathrm{Ls,b}} = \sin(\theta_1)\cos(\theta_4)$
$F_{\mathrm{R,b}} = \sin(\theta_1)\sin(\theta_4)$	$F_{\mathrm{R,b}} = \cos(\theta_1)\cos(\theta_2)\sin(\theta_3)$
$F_{\mathrm{Rs,b}} = \sin(\theta_1)\cos(\theta_4)$	$F_{\mathrm{Rs,b}} = \cos(\theta_1)\cos(\theta_2)\cos(\theta_3)$

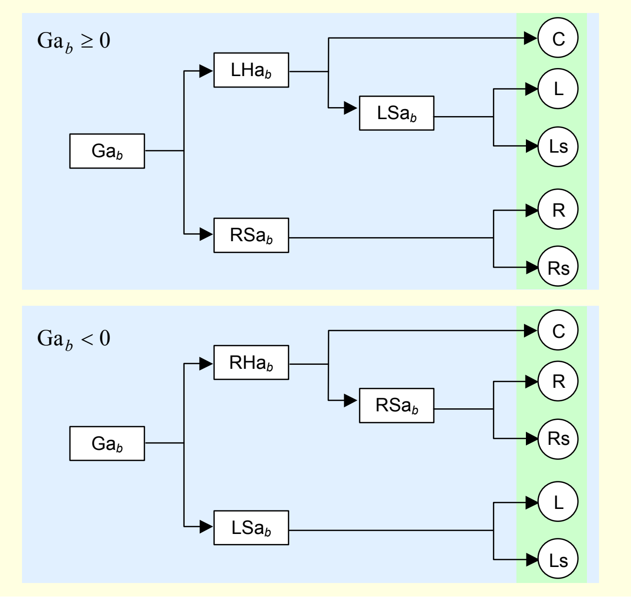


Fig. 2. Schematic diagram of the synthesis procedure of the VSLI scheme.

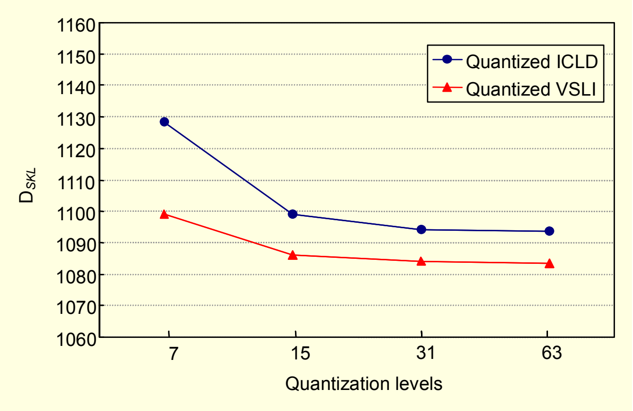


Fig. 3. Comparison of DSKL between quantized VSLI and quantized ICLD.

The δ is a small positive constant value to keep those values in Table 1 non-singular.

The synthesis procedure is graphically illustrated in Fig. 2. If angle Ga$_b$ is greater or equal to zero degrees, the total power of the transmitted downmix signal is projected into the position of LHa$_b$ and RSa$_b$ by CPP. Subsequently, the power gain at LHa$_b$ is projected into the partition of C and position of LSa$_b$ by CPP. The power gain at LSa$_b$ is projected into the partitions of the L and Ls channels. On the other side, the power gain at the position of RSa$_b$ is decomposed into the partitions of R and Rs. Consequently, all the channel power per partition can be estimated by the transmitted angles. The other case when Ga$_b$ is in the right-half plane undergoes the same procedure as RHa$_b$, instead of that of LHa$_b$.

Finally, the spectra of output channels can be reproduced as

$$U_{ch,k} = F_{ch,b} S'_k, \quad B_b \leq k \leq B_{b+1} - 1, \tag{7}$$

where S'_k and $U_{ch,k}$ are spectral coefficients of the down-mixed mono signal and output signal of channel *ch*, respectively.

IV. Experimental Result

Objective tests were carried out to prove that the newly presented VSLI is more reliable to approximate spectra than the conventional ICLD. As a measurement, the symmetric Kullback-Leibler distance (SKL) between the original and output signals is estimated [7]. SKL's quantity is highly correlated with audible distortion and is defined as

$$\mathrm{D}_{SKL} = \int (P(\omega) - Q(\omega)) \log \frac{P(\omega)}{Q(\omega)} d\omega, \tag{8}$$

where $P(\omega)$ and $Q(\omega)$ are the power spectra of the reference and decoded signal, respectively.

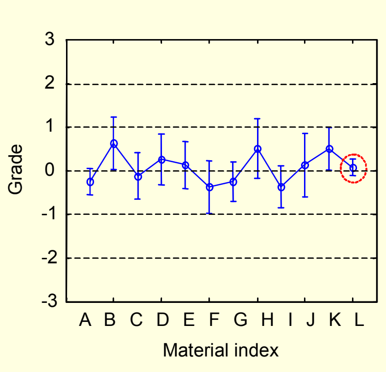
All eleven five-channel items offered by the MPEG audio group [4] were used for our test. All test items have a sampling rate of 44.1 kHz. Several accumulative D$_{SKL}$ are calculated between the original and reconstructed signals, which are decoded by using both the quantized VSLI and the quantized ICLD with a ± 32 dB dynamic range. The VSLI is quantized by a midtread uniform quantizer as in [8], and the analysis hop size is 512 points. Figure 3 shows an accumulative *SKL* according to several quantization levels. The spectral distortion of the VSLI is considerably lower than that of the ICLD for all quantization levels.

Average bit rates for various quantization levels are also given in Table 3. The bit rate is calculated from the amount of the encoded quantizer index differences by the Huffman coding. A more detailed procedure can be found in [1]. Huffman codebooks for the difference index of ICLD and VSLI are trained under the same training condition, using exactly the same large amount of training data set with the exception of the eleven test sequences. From Table 3, it can be easily verified that the VSLI can be encoded more efficiently than the ICLD.

An ITU-R-recommended blind triple-stimulus test to grade the difference with respect to a reference (that is, based on a VSLI scheme) was performed for the subjective assessment. For instance, the +3 point indicates the case when the reference is clearly better than the compared one, the +2 point for when the reference is quite a bit better, and the +1 point when it has only slightly better quality. A negative grade indicates the reverse case of the reference comparison. Twelve subjects participated in this assessment. The results of the assessment are shown in Fig. 4 with a 95% confidence level corresponding to the mean score. For some test materials, VSLI cases are superior to ICLD ones, and while the average grading of the VSLI is slightly better, there is no significant difference between the two schemes.

Table 3. Average bit rate comparison between ICLD and VSLI.

Q (kbps)	7	15	31	63
ICLD	12.47	18.62	24.86	31.50
VSLI	11.85	17.67	23.83	29.45



Index	Material
A	applause
B	ARL applause
C	chostakovitch
D	gountain music
E	glock
F	Indie2
G	jackson1
H	pops
I	poluenc
J	rock concert
K	stomp
L	Total average grade

Fig. 4. Subjective assessment result.

V. Conclusion

VSLI has been newly proposed as a spatial cue to compress multichannel audio signals. By testing the SKL measure, the superiority of our VSLI is confirmed in the sense of both the accuracy of the original signal estimation and the robustness to quantization process. It is shown that the bit rate for the VSLI is lower than that for the ICLD. From the subjective assessment, it is also observed that the quality of the VSLI-based decoded signals is slightly better than that of the ICLD-based ones.

Considering all the above, we can strongly recommend the use of our VSLI-based representation method for SAC.

References

[1] C. Faller, F. Baumgarte, “Binaural Cue Coding-Part II: Schemes and Application,” *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, Nov. 2003.

[2] S. Beack, J. Seo, K. Kang, and M. Hahn, “An Efficient Representation Method for ICLD with Robustness to Spectral Distortion,” *ETRI J.*, vol.27, no.3, June 2005, pp.330-333

[3] J. Seo, H. Moon, S. Beack, K. Kang, and J. Hong, “Multi-channel Audio Service in a Terrestrial-DMB System Using VSLI-Based Spatial Audio Coding,” *ETRI J.*, vol.27, no.5, Oct. 2005, pp.635-638.

[4] ISO/IEC JTC1/SC29/WG11 (MPEG), *Procedures for the Evaluation of Spatial Audio Coding Systems*, Document N6691, Redmond, July 2004.

[5] V. Pulkki, “Virtual Sound Source Positioning Using Vector Base Amplitude Panning,” *J. Audio Eng. Soc.*, vol. 45, June 1997, pp. 456-466.

[6] J. R, West, *Five-Channel Panning Laws: an Analytic and Experimental Comparison*, Master’s Thesis, Music Engineering, University of Miami, 1998.

[7] E. Klabbers and R. Veldhuis, “Reducing Audible Spectral Discontinuities,” *IEEE Trans. on Speech and Audio Proc.*, vol. 9, no. 1, Jan. 2001, pp. 39 – 51.

[8] C. Faller and F. Baumgarte, “Binaural Cue Coding Applied to Stereo and Multi-channel Audio Compression,” *Preprint 112th Conv. Aud. Eng. Soc.*, May 2002.

[9] E. Zwicker and H. Fastl, *Psychoacoustics*, Springer-Verlag, 1999.