Journal of Communication and Computer 14 (2017) 122-128
doi:10.17265/1548-7709/2017.03.003

Surround Sensation Index Based on Differential S-IACF for Listener Envelopment with Multiple Sound Sources

Masato Nakayama[1], Kota Nakahashi[2], Yukoh Wakabayashi[2] and Takanobu Nishiura[1]

1. College of Information Science and Engineering. Ritsumeikan University, Shiga, Japan

2. Graduate School of Information Science and Engineering. Ritsumeikan University, Shiga, Japan

Abstract: LEV (Listener envelopment), which corresponds to surround sensation with independent direct and reflected sounds, is the spatial impression in three-dimensional sound field reproduction systems. However, IACC (interaural cross-correlation), which is conventionally used as an objective index to measure LEV, has difficulty in rating the surround sensation as it assumes a time-invariant sound source. In this paper, we propose a new objective index for surround sensation based on differential short-time IACF (interaural cross-correlation function). We evaluated the effectiveness of the proposed method through evaluation experiments.

Key words: LEV, surround sensation, IACC, differential short-time IACF, objective index.

1. Introduction

Acoustic spatial impression is very important in the field of building acoustics and for surround systems with multiple loudspeakers such as Dolby Surround [1, 2], Dolby Atmos [3], and the 22.2 multichannel (22.2 ch) sound system by NHK Science & Technology Research Laboratories [4]. Acoustic spatial impression consists of at least two aspects, ASW (auditory source width) and LEV (listener envelopment) [5]. ASW is defined as the effect of a sound source being perceived wider than the physical size of the source. LEV is defined as the listener's sensation regarding a surrounding space that is filled with sound. Also, a recent study for 3-D (three-dimensional) sound field reproduction focuses on reproducing LEV, which is regarded as an important factor for 3-D sound field reproduction. The magnitude of LEV is influenced by the level of the reverberant sound and temporal properties of the sound source [6]. Thus, LEV includes surround sensation with independent direct and reflected sounds from multiple sound sources in a wider ASW. Surround sensation corresponds to the temporal change in independent sound sources around a listener. An objective index for LEV is required for evaluating the 3-D sound field reproduction system briefly. IACC (Interaural cross-correlation) [7], which is calculated by an IACF (interaural cross-correlation function), has been used as the objective index for spatial impressions, including LEV. However, it is difficult to evaluate the surround sensation with independent direct and reflected sounds by IACC, because IACC assumes the sound sources are time-invariant. To evaluate LEV with multiple time-variant sound sources, frame analysis is required. In this paper, we focus on statistical analysis of differential short-time IACF (S-IACF).S-IACF denotes IACF analyzed frame by frame, and the differential S-IACF denotes a first-order differential equation of S-IACF. This is because variance of S-IACF might have a temporal change in independent direct and reflected sounds from multiple sound sources. Therefore, we propose a new surround sensation index that is based on differential S-IACF that can evaluate time-variant sound sources. We evaluated the effectiveness of our proposed methods against existing conventional and proposed methods.

Corresponding author: Masato Nakayama (1977-), Ph.D, assistant professor, research field: Acoustic signal processing.

2. Modeling of Listener Envelopment

Fig. 1 shows LEV models in various environments. Fig. 1a shows the model in which the magnitude of LEV is lowest. The magnitude of LEV is lowest in an anechoic environment. In this environment, the observed signals at the left and right ears are given as follows:

$$P_{b,1}(\omega) = V(\omega)H_b(\omega), b \in \{L, R\}, \quad (1)$$

where, ω is the frequency index, $P_{L,1}(\omega)$ and $P_{R,1}(\omega)$are observed signals at theleft and right ears in the anechoic environment, respectively, $V(\omega)$ is the sound source, and $H_L(\omega)$ and $H_R(\omega)$ are head related transfer functions for the left and right ears, respectively. Fig. 1b shows the conventional model of LEV where the magnitude of LEV increases in a reverberant environment. In this environment, the observed signals at the left and right ears are given as follows:

$$P_{b,2}(\omega) = V(\omega)H_b(\omega)R(\omega), b \in \{L, R\}, \quad (2)$$

where, $P_{L,2}(\omega)$ and $P_{R,2}(\omega)$ are observed signals at the left and right ears in the reverberant environment, respectively, and $R(\omega)$ is the room transfer function. The reflected sound is modeled by $R(\omega)$. It is known that IACC has negative correlation to LEV in this environment [7].

The conventional model of LEV is modeled by separating ASW from LEV. However, it is conceivable that ASW is very important for LEV. As shown in Fig. 1c, the wider ASW is, the more sound sources there are. Moreover, the more sound sources there are, the higher the LEV is. Thus, we assume that the magnitude of LEV is larger in the environment that includes independent direct and reflected sounds, such as that shown in Fig. 1c. In addition, the surround sensation with independent direct and reflected sounds is important for evaluating the magnitude of LEV. In this environment, the observed signals at the left and right ears are given as follows:

$$P_{b,3}(\omega) = \sum_{k=1}^{K} V_k(\omega)H_{b,k}(\omega)R_k(\omega), b \in \{L, R\}, \quad (3)$$

where, $P_{L,3}(\omega)$ and $P_{R,3}(\omega)$ are observed signals at the left and right ears in this environment, respectively, K is the number of sound sources, $V_k(\omega)$ is the k-th sound source, $H_{L,k}(\omega)$ and $H_{R,k}(\omega)$ are k-th head related transfer functions , and $R_k(\omega)$ is the k-th room transfer function.

3. Subjective Index for LEV

Conventionally, the magnitude of LEV is evaluated by the MOS (mean opinion score). Subjects assign a score, defined as a scale from one (no surround sensation) to five (high surround sensation), to the sound field is. The scores are averaged to obtain the MOS value for the magnitude of LEV in the sound field. However, evaluating using the MOS places a heavy burden on many subjects. Therefore, the objective index to estimate the score of LEV is required to evaluate LEV more quickly.

4. Conventional Objective Index for LEV

IACC has been used as the objective index for spatial impressions including LEV. IACC is defined as follows:

$$\mathrm{IACC} = \max|\mathrm{IACF}(\tau)|, |\tau| \leq 1\mathrm{ms}, \quad (4)$$

$$\mathrm{IACF}(\tau) = \frac{\sum_{t=0}^{L-1} p_R(t)p_L(t+\tau)}{\sqrt{\sum_{t=0}^{L-1} {p_R}^2(t) \sum_{t=0}^{L-1} {p_L}^2(t)}}, \quad (5)$$

where, t is the time index, $p_L(t)$ and $p_R(t)$ are observed signals at theleft and right ears, respectively, τ is the interaural time difference, and L is the length of the input signal. $|\tau| = 1$ ms corresponds to the maximum interaural time difference. $\mathrm{IACF}(\tau)$ denotes the correlation function between left and right ear signals.

IACC values range from 0 to 1. To evaluate LEV quantitatively, we translate IACC into a five-grade score by the second order regression curve. The regression curve is given as follows:

$$S_{\mathrm{IACC}} = \alpha_c \cdot \mathrm{IACC}^2 + \beta_c \cdot \mathrm{IACC} + \gamma_c, \quad (6)$$

where, S_{IACC} indicates the five-grade score that corresponds to the magnitude of LEV, and α_c, β_c, and γ_c are given by the regression analysis between

IACC and the MOS of LEV. S_{IACC} correlates with the MOS of LEV described in Section 3. However, IACC is the objective index that expects a single and time-invariant sound source. Therefore, it is difficult to rate the surround sensation corresponding to time-variant sound sources.

5. Suggestion for a Surround Sensation Index based on Differential Short-Time IACF for LEV

In this paper, we propose a new objective index with the VDSI (variance of differential short-time IACF), which corresponds to the magnitude of LEV including the surround sensation with independent direct and reflected sounds.

To calculate the temporal change in the DOA (direction of arrival) of direct and reflected sounds and the number of sound sources around a listener, the short-time IACF (S-IACF) is formulated as follows:

$$\begin{aligned}&\mathrm{S\text{–}IACF}(k,n) \\ &= \mathrm{IFT}\left[\frac{\mathrm{FT}[p_R(t+nX)]\mathrm{FT}[p_L(t+nX)]^*}{|\mathrm{FT}[p_R(t+nX)]\mathrm{FT}[p_L(t+nX)]|}\right],\end{aligned} \tag{7}$$

where, k ($|k| \leq 1$ ms, which is the interaural time difference) is time index, n is the frame index, $\mathrm{S\text{–}IACF}(k,n)$ is the S-IACF coefficient in the k-th sample and n-th frame, $\mathrm{FT}[\cdot]$ is the Fourier transform, $\mathrm{IFT}[\cdot]$ is the inverse Fourier transform, X is the frame shift length, T ($0 < T \leq 80$ ms, which is early reflection time [8]) is the frame length, $d$is the distance between both ears, and c is the sound velocity. Eq. (7) shows interaural cross-power spectrum phase analysis in short time. We then define VDSI as the variance of the frame differences of S-IACF to rate the temporal change in the DOA of direct and reflected sounds and the number of sound sources around a listener. VDSI is calculated as follows:

$$\begin{aligned}\mathrm{VDSI} &= \frac{1}{(2K+1)N} \\ &\sum_{k=-K}^{K}\sum_{n=0}^{N-1}|\mathrm{S\text{–}IACF}(k,n+1) \\ &\quad - \mathrm{S\text{–}IACF}(k,n)|,\end{aligned} \tag{7}$$

Where K ($K = F_S \cdot 0.001$) is the maximum sample of the interaural time difference, F_S is sampling frequency, and N is the total number of frames. VDSI ($0 < VDSI < 2$) tends to be the smaller value because it is calculated by a differential S-IACF sequence. Finally, the score with the second order regression curve of VDSI is calculated as follows:

$$S_{\mathrm{VDSI}} = \alpha_p \cdot \mathrm{VDSI}^2 + \beta_p \cdot \mathrm{VDSI} + \gamma_p, \tag{8}$$

Where S_{VDSI} is the surround sensation index with the five-grade score, and α_p,β_p, and γ_p are calculated by the regression analysis between the five-grade score of MOS with the subjective index of LEV and VDSI. As a result, S_{VDSI} can estimate the five-grade score of the MOS with the subjective index for LEV.

6. Experimental Evaluation

We carried out an evaluation experiment to confirm the effectiveness of the proposed method.

In this experiment, subjects first evaluated the surround sensation of each sound field by its MOS.

We then recorded each sound field using a dummy head and calculated IACC and VDSI with these signals.

Finally, we carried out a regression analysis by the least square method between the MOS, IACC, and VDSI, and estimated the score of the MOS with Eqs. (6) and (9).

6.1 Experimental Conditions

Table 1 shows the conditions of the experiment. Multiple sound fields were generated from loudspeakers

Table 1 Experimental conditions.

Environment (A-weighted sound level)	Soundproof room (19.4 dB)
Subjects	Two females and five males
Sampling frequency	16 kHz
Quantization	16 bit
Number of trials	3
Number of sound sources	1,2,4,6,8
Sound source	White noise, Voice [9], Classical music
Loudspeaker	FOSTEX, FE83En
Dummy head	NEUMANN, KU100

surrounding a subject, as shown in Fig. 2. The radiation angle of each loudspeaker was 60 [deg.]. We carried out evaluation experiments under various sound source conditions and with various sound sources.

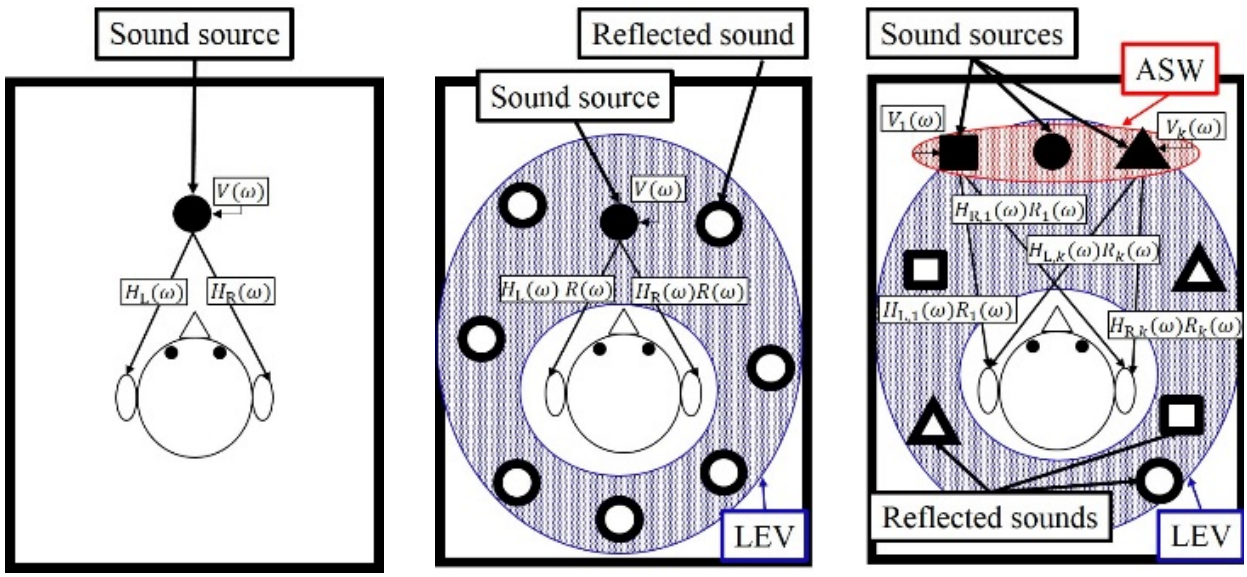
Table 2 shows the relationship between the number of sound sources and indexes of loudspeakers. The evaluation sound sources included a speech voice reading ATR 503 sentences [5], classical music, and two kinds of white noise (with and without sound pressure fluctuation). Moreover, we carried out evaluation experiments under various frame lengths and shifts to investigate the estimation accuracy. Table 3 shows α_c, β_c, and γ_c in Eq. (6). Table 4 shows the conditions of frame length and shift in VDSI and α_p, β_p, and γ_p in Eq. (9) under each condition. We translate IACC and VDSI into a five-grade scale by regression curves, which are given by Eqs. (6) and (9). To evaluate the estimation accuracy of the MOS, we calculated the AEE (average estimation error) as follows:

$$\mathrm{AEE}_a = \frac{1}{M}\sum_{m=1}^{M}\left|\mathrm{MOS}_m - S_{a,m}\right|, \quad a \in \{\mathrm{IACC}, \mathrm{VDSI}\}, \tag{10}$$

Where $\mathrm{AEE}_{\mathrm{IACC}}$ and $\mathrm{AEE}_{\mathrm{VDSI}}$ denote AEE with IACC and VDSI, respectively, M is the number of experimental condition, MOS_m is the MOS in m-th experimental condition, and $S_{\mathrm{IACC},m}$ and $S_{\mathrm{VDSI},m}$ is the calculated score by IACC and VDSI in m-th experimental condition, respectively.

6.2 Modeling of Listener Envelopment

Fig. 3 shows the relationship between the MOS and VDSI under each condition of frame length and shift. A correlation coefficient R^2 tends to increase in the case where the frame length is shorter. For this case, minute changes in the number of sound sources and their DOA can be evaluated. Therefore, it is conceivable that the correlation coefficient is increased.



(a)Single sound source without reflected sound

(b)Single sound source with reflected sounds

(c) Independent sound sources with reflected sounds

Fig. 1 LEV models in various environments.

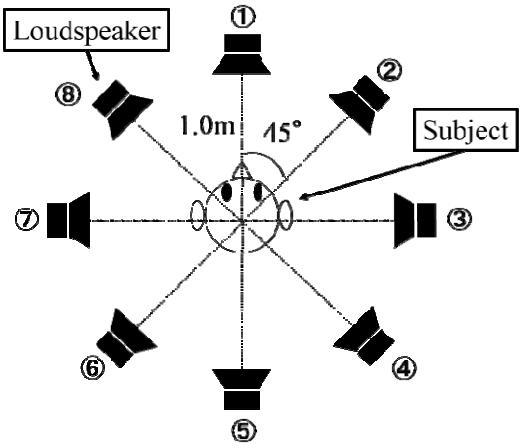


Fig. 2 Arrangement of subject and loudspeaker.

Table 2 Relationship between number of sound sources and indexes of loudspeakers.

Number of sound sources	Indexes of loudspeakers
1	1
2	3,7
4	1,3,5,7
6	1, 3, 4, 6, 7,8
8	1, 2, 3, 4, 5, 6,7, 8

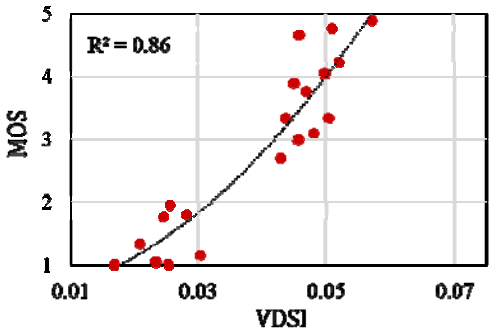
Table 3 α_c, β_c, and γ_c in Eq. (6) for IACC.

α_c	β_c	γ_c
-10.0	9.0	2.1

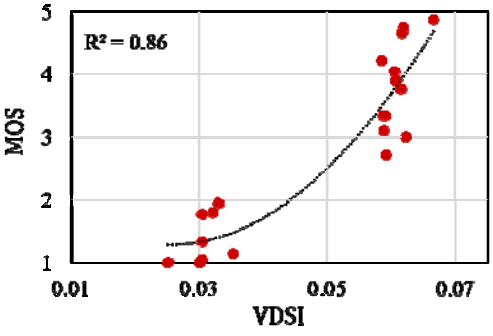
Table 4 α_p, β_p, and γ_p in Eq. (9) for VDSI.

Frame length and shift	α_p	β_p	γ_p
$T = 25$ ms, $X = 25$ ms	1,948.2	-115.0	3.0
$T = 25$ ms, $X = 12.5$ ms	2,006.0	-101.6	2.6
$T = 50$ ms, $X = 50$ ms	-220.0	104.1	-1.2
$T = 50$ ms, $X = 25$ ms	1,232.9	9.8	0.4

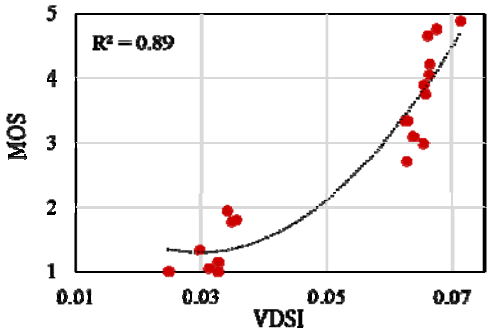
Fig. 4 shows the relationship between the MOS and IACC. The correlation coefficient between the MOS and IACC is lower than that between the MOS and VDSI. Fig. 5 shows the score of MOS, S_{IACC}, and S_{VDSI} undereach condition. VDSI can estimate the MOS of the surround sensation with independent direct and reflected sounds in most cases. However, the score of S_{VDSI} tends to larger than the MOS in the case where the sound source is the white noise, or the number of sound sources is two or less.



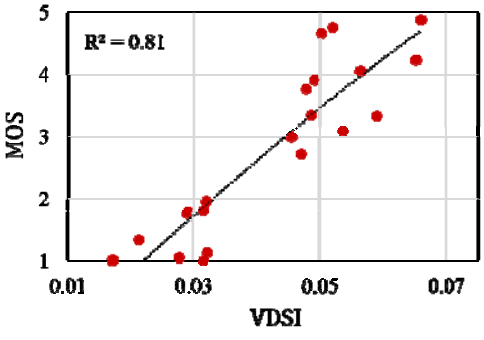
(a) $T = 25$ ms, $X = 25$ ms



(c) $T = 50$ ms, $X = 50$ ms



(b) $T = 25$ ms, $X = 12.5$ ms



(d) $T = 50$ ms, $X = 25$ ms

Fig. 3 Relationship between MOS and VDSI.

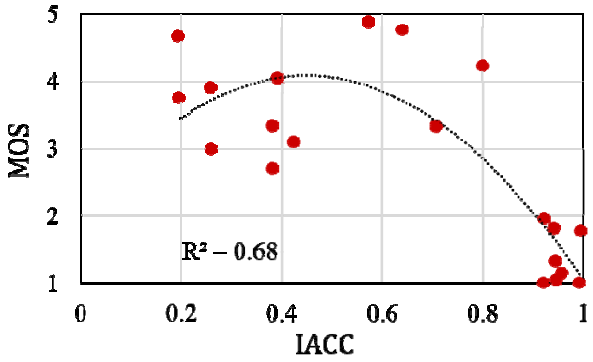


Fig. 4 Relationship between MOS and IACC.

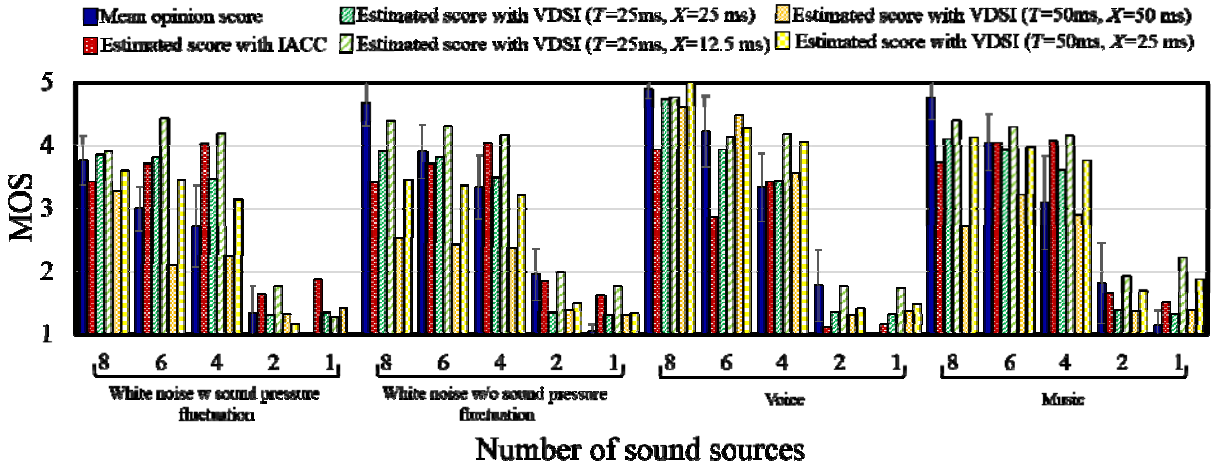


Fig. 5 Score of MOS, S_{IACC}, and S_{VDSI} in each experimental condition.

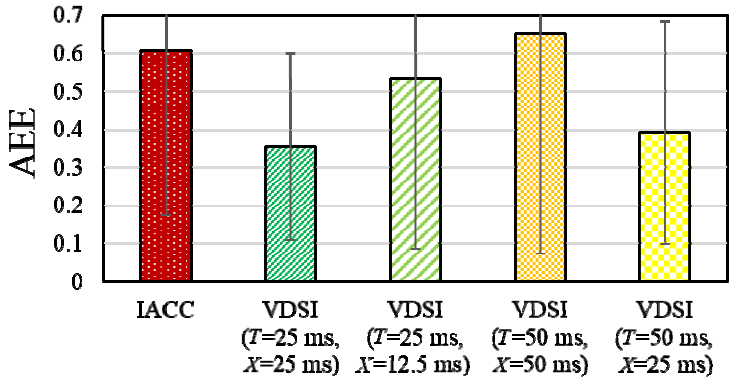


Fig. 6 AEE of each index.

Fig. 6 shows the AEE under each condition. The AEE tends to decrease where VDSI is utilized. Moreover, the AEE is the lowest in the case where $T = 25$ ms and $X = 25$ ms. However, the AEE of VDSI is higher than that of IACC in the case where $T = 50$ ms and $X = 50$ ms.Therefore, it is conceivable that the estimation accuracy of the MOS by VDSI decreases in the case where the frame length and shift is longer.

7. Conclusions

In this paper, we proposed the objective index of the surround sensation with independent direct and reflected sounds on the basis of VDSI. We confirmed

that VDSI can estimate the MOS of the surround sensation with independent sound sources in LEV more accurately than IACC. In the future, we intend to evaluate the proposed method with the diffuse sound field.

Acknowledgement

This work was partly supported by JSPS KAKENHI Grant Numbers JP24220004, JP26280065, and JP15K16030.

References

[1] Dolby Laboratories, Inc., Dolby Digital 5.1: https://www.dolby.com/us/en/technologies/dolby-digital.html.

[2] Dolby Laboratories, Inc., Dolby Surround 7.1: https://www.dolby.com/us/en/technologies/dolby-surround-7-1.html.

[3] Dolby Laboratories, Inc., Dolby Atmos: http://www.dolby.com/us/en/technologies/home/dolby-atmos.html.

[4] NHK Science & Technology Research Laboratories, 2011. “22.2 Multichannel Audio Format Standardization Activity.” *Broadcast Technology*, 45.

[5] Bradley, J. S., and Soulodre, G. A. 1995. “The Influence of Late Arriving Energy on Spatial Impression.”*The Journal of the Acoustical Society of America* 97 (4): 2263-71.

[6] Van DorpSchuitman, J. 2011. “Auditory Modelling for Assessing Room Acoustics.” Ph.D. thesis, TU Delft, Delft University of Technology.

[7] Hidaka, T., eranek, L. L., B and Okano, T. 1995. “Interaural Cross-Correlation, Lateral Fraction, and Low- and High-Frequency Sound Levels as Measures of Acoustical Quality in Concert Halls.” *The Journal of the Acoustical Society of America* 98 (2): 988-1007.

[8] Barron, M., and Lee, L. J. 1988. “Energy Relations in Concert Auditoriums. I.” *The Journal of the Acoustical Society of America* 84 (2): 618-28.

[9] Sagisaka, Y., Takeda, K., Abe, M., Katagiri, S., Umeda, T., and Kuwabara, H. 1990. “A Large-Scale Japanese Speech Database.” *First International Conference on Spoken Language Processing*, 1089-92.