Audio Engineering Society

Convention Paper

Presented at the 118th Convention
2005 May 28–31 Barcelona, Spain

Modifying Stereo Recordings Using Acoustic Information Obtained with Spot Recordings

Christof Faller[1] and Markus Erne[2]

[1] *Audiovisual Communications Laboratory, EPFL Lausanne, Switzerland*

[2] *Scopein Research, Aarau, Switzerland*

Correspondence should be addressed to C. Faller or M. Erne (`christof.faller@epfl.ch`, `markus.erne@scopein.ch`)

ABSTRACT

We are addressing the following scenario: A concert is recorded with a stereo microphone technique. Additionally, several instruments or groups of instruments are recorded with spot microphones. The proposed technique adaptively in time estimates the impulse response between the spot microphones and the left and right stereo microphones. The spot microphones, filtered with the estimated impulse responses, are scaled and subtracted/added from the stereo microphone signals to attenuate/amplify the corresponding instruments. No amplitude panning and reverberators are needed, while the auditory spatial image attributes of the stereo recording are not altered.

1. INTRODUCTION

A setup for making a stereo recording is illustrated in Figure 1. A stereo signal, $y_1(n)$ and $y_2(n)$, is recorded using a stereo microphone technique. Stereo (or multi-channel) microphone techniques are usually being optimized for "capturing" (in the form of a stereo signal) the auditory spatial image that is perceived by a listener during a musical performance. Attributes of the auditory spatial image are locations [1] and widths [2] of virtual sources and listener envelopment [3].

However, it is likely that the balance of the level between the various instruments in the stereo recording is not optimal. Therefore, usually a number of microphones are placed close to instruments or groups of instruments, denoted spot microphones, yielding signals $x_i(n)$ $(1 \leq i \leq L)$, where L is the number of spot microphones. If an instrument in the stereo recording is not loud enough, the corresponding spot microphone recording is used to amplify it. Usually, amplitude panning [4, 5, 6, 7] is applied to the spot microphone signal to match the direction of

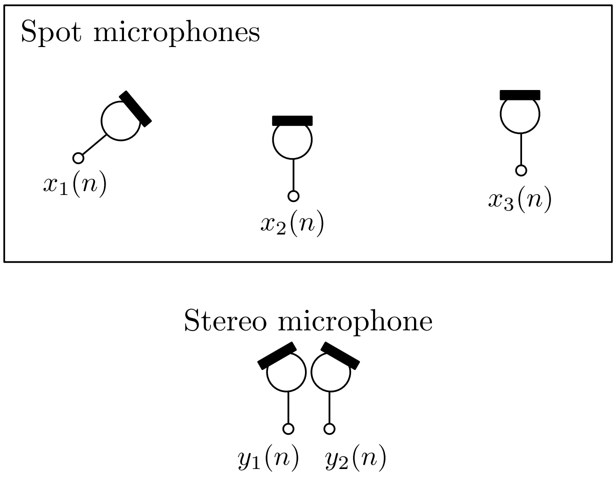

Fig. 1: Setup for carrying out a stereo recording. A stereo signal is captured using a stereo microphone technique. Single instruments or groups of instruments are captured with spot microphones.

the corresponding instruments in the stereo recording. Additionally, reverberators [8, 9, 10] are applied for mimicking the source width and envelopment associated with the corresponding instrument. More sophisticated techniques have been introduced, such as virtual surround panning (VSP) [11], which model the room impulse responses of a room to mimic the impulse responses from the spot microphones to the stereo microphones. Note that a limitation of techniques using reverberators or room models is that they do not consider the effect of the impulse response from the instrument to the spot microphone.

It is a tedious task for the recording engineer to select amplitude panning and reverberator parameters such that an instrument can be amplified in the stereo recording without (negatively) altering the auditory spatial image attributes of the non-modified stereo recording. The previously mentioned VSP may sound a bit more natural than conventional amplitude panning and may be more straight forward to use. However, amplitude panning-based techniques and VSP can only be used to amplify instruments in the stereo recording, but not to attenuate them. This is so, because instruments in the stereo recording can only be attenuated (by subtraction) if accurate estimates of the instrument signal components to be attenuated are available. Since previous techniques do not estimate these signal components accurately they can not be used for attenuation of instruments.

The proposed technique aims at addressing both of the mentioned shortcomings (difficulty to select optimal amplitude panning and reverberator parameters and inability to attenuate instruments in the stereo recording). We estimate the impulse responses between the spot microphones and the left and right stereo recording microphones. By convolving each spot microphone signal with the corresponding estimated impulse responses, an estimate of the stereo signal component of each instrument is obtained. By subtracting scaled versions of these estimated stereo signals from the stereo recording, the level of instruments in the stereo recordings can be attenuated. By adding scaled versions of the estimated stereo recordings, instruments in the stereo recording are amplified while maintaining their spatial properties (direction, distance, width, and envelopment) without a need for manually tuning amplitude panning and reverberator parameters.

The proposed technique is not limited to modifying stereo recordings. It can also be applied to multichannel surround recording techniques. In this case, the impulse responses are estimated between each spot microphone and each surround recording channel.

The paper is organized as follows. Section 2 defines the problem of estimating the impulse response between the spot microphones and the stereo microphones and derives the optimal solution in the least squares sense. Section 3 describes how the stereo recordings are modified, given the estimated impulse responses. Implementation details and practical issues are discussed in Section 4. Simulation results, using real stereo and spot recordings are presented in Section 5. The conclusions are presented in Section 6.

2. ESTIMATING THE IMPULSE RESPONSES FROM THE SPOT MICROPHONES TO THE STEREO MICROPHONES

The notation that is used in the following is summarized in Figure 2. The stereo microphone signals and spot microphone signals are denoted $y_i(n)$ and

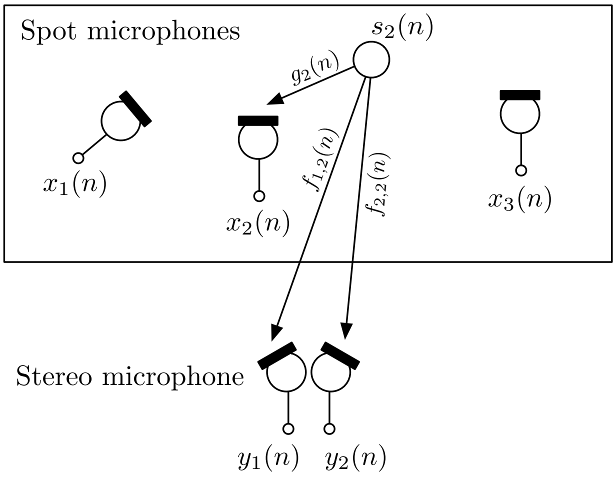

Fig. 2: The notation that is used for signals and room impulse responses.

$x_i(n)$, respectively. The room impulse response from an instrument with index i to its spot microphone (also index i) is denoted $g_i(n)$. The room impulse response from an instrument with index i to the left and right stereo microphones are denoted $f_{1,i}(n)$ and $f_{2,i}(n)$, respectively.

Given the defined variables, the stereo signal which is captured by the stereo microphone technique can be written as

$$\begin{aligned} y_1(n) &= \sum_{l=1}^{L} f_{1,l}(n) \star s_l(n) + v_1(n) \\ y_2(n) &= \sum_{l=1}^{L} f_{2,l}(n) \star s_l(n) + v_2(n) , \end{aligned} \quad (1)$$

where $s_i(n)$ are the sound sources (instruments), L is the number of instruments (for simplicity the same as number of spot microphones), $v_1(n)$ and $v_2(n)$ are noise signals, and $\star$ is the linear convolution operator.

The spot microphone signals can be written as

$$x_i(n) = g_i(n) \star s_i(n) + w_i(n) , \quad (2)$$

where $w_i(n)$ is noise and $g_i(n)$ is the room impulse response between the instrument and its spot microphone. The noise is composed of ambient noise and

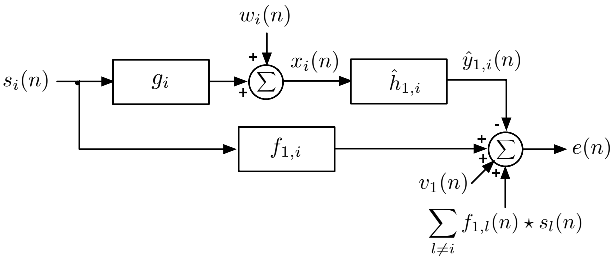

Fig. 3: The relation between the various filters and signals for a spot microphone with index i and the left stereo microphone signal $y_1(n)$.

signal components arising from other instruments than the one to which the spot microphone is associated.

The transfer functions between the recorded stereo signal, $y_1(n)$ and $y_2(n)$, and a spot microphone signal $x_i(n)$ are

$$\begin{aligned} H_{1,i}(j\omega) &= \frac{F_{1,i}(j\omega)}{G_i(j\omega)} \\ H_{2,i}(j\omega) &= \frac{F_{2,i}(j\omega)}{G_i(j\omega)} , \end{aligned} \quad (3)$$

where $G_i(j\omega)$, $F_{1,i}(j\omega)$, and $F_{2,i}(j\omega)$ are the Fourier transforms of $g_i(n)$, $f_{1,i}(n)$, and $f_{2,i}(n)$, respectively.

In the following, analysis is only carried out with respect to the left stereo microphone. The derivations for the right stereo microphone (or other surround microphones) are similar. Figure 3 illustrates the relation between signals and filters for a spot microphone with index i and the left stereo microphone. It is assumed that all signals are zero mean. Given a transfer function estimate $\hat{H}_{1,i}(j\omega)$, the estimate of the spectrum of the signal of the instrument with index i contained in the left stereo signal is

$$\hat{Y}_{1,i}(j\omega) = \hat{H}_{1,i}(j\omega) X_i(j\omega) . \quad (4)$$

The error signal spectrum is defined as

$$E(j\omega) = Y_1(j\omega) - \hat{H}_{1,i}(j\omega) X_i(j\omega) . \quad (5)$$

As error criterion we use the mean square error (MSE),

$$J(\hat{H}_{1,i}(j\omega)) = \mathrm{E}\{E(j\omega)E^{\star}(j\omega)\} , \quad (6)$$

where E{.} denotes expectation and * denotes complex conjugate. (Note that $E(j\omega)$ is the error spectrum and E{.} is the expectation operator). This can be written as

$$\begin{aligned} J(\hat{H}_{1,i}(j\omega)) &= \mathrm{E}\{(Y_1(j\omega) - \hat{H}_{1,i}(j\omega)X_i(j\omega)) \\ &\quad (Y_1(j\omega) - \hat{H}_{1,i}(j\omega)X_i(j\omega))^\star\}\,, \end{aligned} \tag{7}$$

or

$$\begin{aligned} J(\hat{H}_{1,i}(j\omega)) &= \mathrm{E}\{Y_1(j\omega)Y_1^\star(j\omega)\} \\ &\quad -\hat{H}^\star_{1,i}(j\omega)\mathrm{E}\{Y_1(j\omega)X_i^\star(j\omega)\} \\ &\quad -\hat{H}_{1,i}(j\omega)\mathrm{E}\{Y_1^\star(j\omega)X_i(j\omega)\} \\ &\quad +\hat{H}_{1,i}(j\omega)\hat{H}_{1,j}(j\omega)^\star\mathrm{E}\{X_i(j\omega)X_i^\star(j\omega)\}\,. \end{aligned} \tag{8}$$

The derivative of a real-valued function with respect to its complex variable is equal to its partial derivative with respect to the complex conjugate of its variable (while holding the complex variable constant) [12, 13]. Thus, the minimum of (8) is found by computing the partial derivative with respect to $\hat{H}^\star_{1,i}(j\omega)$ and setting it to zero,

$$\left.\frac{\partial J(\hat{H}_{1,i}(j\omega))}{\partial \hat{H}^\star_{1,i}(j\omega)}\right|_{\hat{H}_{1,i}(j\omega)} = 0\,, \tag{9}$$

resulting in

$$\begin{aligned} &-\mathrm{E}\{Y_1(j\omega)X_i^\star(j\omega)\} \\ +\hat{H}_{1,i}(j\omega)&\mathrm{E}\{X_i(j\omega)X_i^\star(j\omega) = 0\,. \end{aligned} \tag{10}$$

The transfer function between the spot microphone and the left stereo microphone, optimal in an MSE sense, is thus

$$\hat{H}_{1,i}(j\omega) = \frac{\mathrm{E}\{Y_1(j\omega)X_i^\star(j\omega)\}}{\mathrm{E}\{X_i(j\omega)X_i^\star(j\omega)\}}\,. \tag{11}$$

Impact of noise on estimation In order to see what impact the noise signals $v_1(n)$, $v_2(n)$, and $w(n)$ have on the transfer function estimate, we are rewriting (11) by substituting Fourier transforms of (1) and (2) into (11),

$$\hat{H}_{1,i} = \frac{\mathrm{E}\{(\sum_{l=1}^{L} F_{1,l}S_l + V_1)(G_iS_i + W_i)^\star\}}{\mathrm{E}\{(G_iS_i + W_i)(G_iS_i + W_i)^\star\}}\,. \tag{12}$$

Note that we ignored the $j\omega$ argument in the spectra for shorter notation. Assuming that the spot microphones contain signal of only one instrument, i.e. $w_i(n)$, $v_1(n)$, and $s_i(n)$ are mutually independent, (12) can be simplified to

$$\hat{H}_{1,i} = \frac{F_{1,i}G_i^\star\mathrm{E}\{S_iS_i^\star\}}{\mathrm{E}\{G_iG_i^\star S_iS_i^\star + W_iW_i^\star\}}\,. \tag{13}$$

When there is no noise in the spot microphone signals ($W_i(j\omega) = 0$), then (13) is equal to the real transfer function (3). Noise in the spot recordings ($W_i(j\omega)$) results in a bias in the estimated transfer function. Thus, the amount of noise in the spot microphones is to be minimized. Noise in the stereo microphone signals ($V_i(j\omega)$) has no effect on the least mean square solution of the transfer function.

When not only the intended instrument signal is contained in the spot microphones, $v_1(n)$ can not be considered as independent of $y_1(n)$. In this case, (12) can be written as

$$\hat{H}_{1,i} = \frac{F_{1,i}G_i^\star\mathrm{E}\{S_iS_i^\star\} + \sum_{l=1}^{L} G_l^\star\mathrm{E}\{V_1S_l^\star\}}{\mathrm{E}\{G_iG_i^\star S_iS_i^\star + W_iW_i^\star\}}\,. \tag{14}$$

This implies that the estimate is more biased and therefore it is desirable to set up the spot microphones such that they mostly contain signal of only one instrument. Otherwise, the transfer function estimate will be inaccurate. The impulse response estimate is the inverse Fourier transform of $\hat{H}_{1,i}(j\omega)$.

3. MODIFYING STEREO RECORDINGS

Given the estimated impulse responses, $\hat{h}_{1,i}(n)$ and $\hat{h}_{2,i}(n)$, the stereo signal is modified by

$$\begin{aligned} \tilde{y}_1(n) &= y_1(n) + \sum_{i=1}^{L}(10^{\frac{a_i}{20}} - 1)\hat{h}_{1,i}(n) \star x_i(n) \\ \tilde{y}_2(n) &= y_2(n) + \sum_{i=1}^{L}(10^{\frac{a_i}{20}} - 1)\hat{h}_{2,i}(n) \star x_i(n)\,, \end{aligned} \tag{15}$$

where a_i are the desired gain values in dB for the instruments corresponding to the spot microphones. Positive gain values increase the level of an instrument and negative gain values attenuate the level of an instrument (if the estimated $\hat{h}_{1,i}(n)$ and $\hat{h}_{2,i}(n)$ are precise enough).

4. IMPLEMENTATION DETAILS AND PRACTICAL ISSUES

4.1. Implementation details

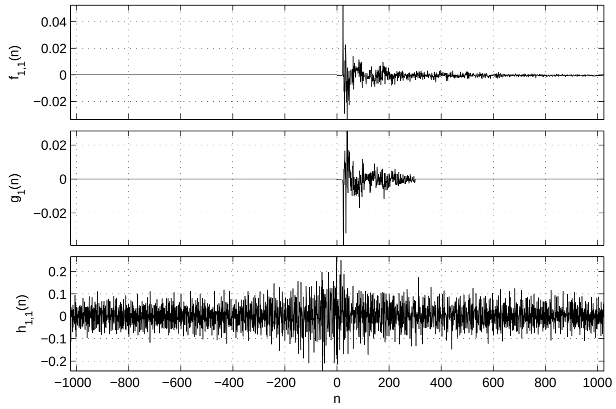

Fig. 4: Given two measured room impulse responses for $f_{1,1}(n)$ and g_1(n), the impulse response between the spot and stereo microphone, $h_{1,1}(n)$, is shown.

The signal spectra $X_i(j\omega)$ and $Y_i(j\omega)$, needed for the estimation of the transfer functions (11), are estimated using a discrete short-time Fourier transform (STFT). We are using a Hann analysis window with 50% overlap. The length of the STFT window is determined experimentally for each specific recording. For simplicity, in the following, we are denoting the discrete estimated signal spectra the same as previously ($j\omega$ argument as opposed to discrete frequency index). The time index k is used for numbering STFT spectra in time.

Due to thermal fluctuation room impulse responses are changing all the time [14]. Thus, the estimations need to be carried out in a time adaptive manner. The transfer function $H_{1,i}(j\omega)$ (11) is in practice estimated by

$$\hat{H}_{1,i}(j\omega, k) = B_i(\omega)\frac{\mathrm{E}\{Y_1(j\omega, k)X_i^\star(j\omega, k)\}}{\mathrm{E}\{X_i(j\omega, k)X_i^\star(j\omega, k)\} + r}\,, \tag{16}$$

where $B_i(\omega)$ is a real valued gain function, r is a positive real regularization constant, and the time index k indicates that the estimation is carried out in a time adaptive manner. The gain function $B_i(\omega)$ is selected such that the resulting impulse response will be bandpass filtered to the spectral range in which the instrument of the spot recording has non-negligible energy. The cross-spectrum and power spectrum means are estimated iteratively by

$$\begin{aligned}\mathrm{E}\{Y_1(j\omega, k+1)X_i^\star(j\omega, k+1)\} &= \alpha X_i^\star(j\omega, k)Y_i(j\omega, k)\\ &\quad +(1-\alpha)\mathrm{E}\{Y_1(j\omega, k)X_i^\star(j\omega, k)\}\\ \mathrm{E}\{X_i(j\omega, k+1)X_i^\star(j\omega, k+1)\} &= \alpha X_i^\star(j\omega, k)X_i(j\omega, k)\\ &\quad +(1-\alpha)\mathrm{E}\{X_i(j\omega, k)X_i^\star(j\omega, k)\}\,.\end{aligned} \tag{17}$$

The factor α determines the degree of smoothing of the estimation over time. The time constant of the exponential decay in seconds is

$$T = \frac{M}{\alpha f_s}\,, \tag{18}$$

where f_s is the sampling frequency and M is the STFT window hop size.

5. SIMULATIONS

5.1. What impulse responses to expect

For illustrative purposes we took two measured impulse responses for $f_{1,1}(n)$ and $g_1(n)$ and computed the impulse response $h_{1,1}(n)$ between the spot microphone and stereo microphone (3). We reduced the length of the measured impulse response for $g_1(n)$ since the spot microphone is typically set up such that it captures as little reverberation as possible. Figure 4 shows the three impulse responses. The inversion of $g_1(n)$ results in that $h_{1,1}(n)$ is not causal anymore.

Note that the impulse response $h_{1,1}(n)$ does not look like a typical room impulse response (unless the spot microphone is such that $g_1(n) \approx \delta(n)$). Thus, early and late parts of $f_{1,1}(n)$ can not be directly related to early and late parts of $h_{1,1}(n)$. This also indicates the limitation of previous techniques, since their rationale is always the modeling of a room impulse response to be applied to the spot microphones.

5.2. Processing a classical stereo recording

We experimented with a real classical recording which was recorded as is described in the following. A total of 16 microphones have been used for the recording of the “Turangalîla-Symphony” from Olivier Messiaen, a French composer, who wrote the symphony between 1946 and 1948. Almost 100 musicians, most of them students from the “Hochschule für Musik & Theater” in Zurich and Geneva played this impressive piece of music.

Two Jecklin-disks, equipped with B&K4006 microphones, an AMS-Soundfield microphone using Ambisonic B-format, and 11 complementary spot-microphones have been used for the recording.

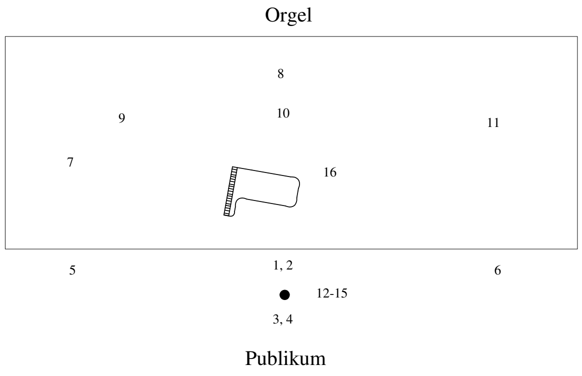

Fig. 5: Microphone setup of the classical recording that we processed.

Capturing the instrument "Ondes Martenot" accurately was a challenge and the microphone setup is highlighted in Figure 5.

Mic. No.	Specification
1 and 2	Jecklin Disk OSS, B&K4006
3 and 4	Jecklin Disk OSS, Sanken
5 and 6	Schoeps MK2S
7	Harp, Neumann U87
8	Base drum, Neumann KM84
9-11	Wood instruments, Neumann U87
12-15	AMS-Soundfield
16	"Ondes Martenot", Neumann KM84

Table 1: Purpose and type of the microphones in Figure 5.

The 16-tracks have been recorded, using a Mackie SDR24/96-Harddisk Recorder, in parallel with a Pyramix-Recording System from Merging Technologies, operating in DSD-Recording mode.

Figure 6 shows an example of estimated impulse responses between the Base drum spot and the left and right Jecklin Disk microphones 1 and 2, $\hat{h}_{1,1}$ and $\hat{h}_{2,1}$.

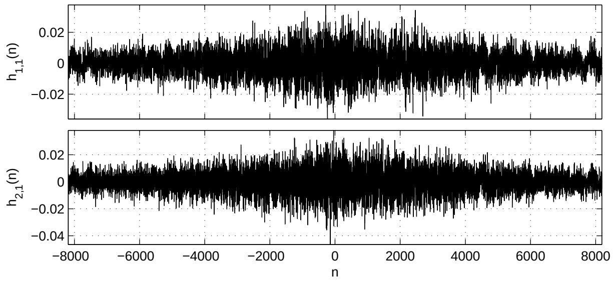

Fig. 6: The estimated impulse responses for a classical stereo recording with a spot microphone for the Base drum spot microphone.

To get correct spatialization when increasing the gain of instruments ($a_i > 0$ dB in (15)) is rather unproblematic and does not need very high precision of the impulse response estimates. It is important to choose the gain filter $G_i(\omega)$ and regularization r appropriately for each spot microphone in order to get a reasonable estimate. We were also able to attenuate instruments in the stereo recording. If the choice of the time constant T is short enough (i.e. only very time-local information is used for impulse response estimates) instruments can always be attenuated. But if T is chosen too short, artifacts occur since then the estimated impulse responses are only effective very locally and do not correspond to the real impulse responses.

6. CONCLUSIONS

We proposed a technique for modifying stereo (and multi-channel) microphone recordings using spot microphone recordings. The impulse responses from each spot microphone to the stereo microphones are estimated and used for estimating the signal components of instruments in the stereo recording. By adding or subtracting these estimated signal components, the level of instruments in the stereo recording can be increased or decreased, respectively, without altering the auditory spatial image associated with the stereo recording.

We carried out a number of proof of concept experiments, which show that in principle the proposed method works on real recorded audio material. The selection of the estimation parameters, such as the regularization, gain filter, and estimation time constant, is critical and more research is required for finding optimal or nearly optimal solutions.

ACKNOWLEDGMENTS

Thanks to Prof. Jürg Jecklin who performed the recording with our AES-Sound-technician Students from Zepra and thanks to the “Hochschule für Musik & Theater Zürich” which organized the concert.

7. REFERENCES

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, The MIT Press, Cambridge, Massachusetts, USA, revised edition, 1997.

[2] T. Okano, L. L. Beranek, and T. Hidaka, “Relations among interaural cross-correlation coefficient (IACC$_{\mathrm{E}}$), lateral fraction (LF$_{\mathrm{E}}$), and apparent source width (asw) in concernt halls,” *J. Acoust. Soc. Am.*, vol. 104, no. 1, pp. 255–265, July 1998.

[3] M. Morimoto and Z. Maekawa, “Auditory spaciousness and envelopment,” in *Proc. 13th Int. Congr. on Acoustics*, Belgrade, 1989, vol. 2, pp. 215–218.

[4] B. B. Bauer, “Phasor analysis of some stereophonic phenomena,” *J. Acoust. Soc. Am.*, vol. 33, pp. 1536–1539, Nov. 1961.

[5] B. Bernfeld, “Attempts for better understanding of the directional stereophonic listening mechanism,” in *Preprint 44th Conv. Aud. Eng. Soc.*, Feb. 1973.

[6] J. C. Bennett, K. Barker, and F. O. Edeko, “A new approach to the assessment of stereophonic sound system performance,” *J. Audio Eng. Soc.*, vol. 33, no. 5, pp. 314–321, May 1985.

[7] V. Pulkki, “Localization of amplitude-panned sources I: Stereophonic panning,” *J. Audio Eng. Soc.*, vol. 49, no. 9, pp. 739–752, 2001.

[8] M. R. Schroeder, “Improved quasi-stereophony and ”colorless” artificial reverberation,” *J. Acoust. Soc. Am.*, vol. 43, no. 9, pp. 1061–1064, Aug. 1961.

[9] M. R. Schroeder, “Natural sounding artificial reverberation,” *J. Aud. Eng. Soc.*, vol. 10, no. 3, pp. 219–223, 1962.

[10] W. G. Gardner, “Reverberation algorithms,” in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds., chapter 2. Kluwer Academic Publishing, Norwell, MA, USA, 1998.

[11] U. Horbach, “Virtual Surround Panning (VSP) for the D950 digital mixing system,” *Studer Profession Audio, Switzerland*, http://www.studer.ch.

[12] T. A. Arias, “Notes on the ab initio theory of molecules and solids: Density functional theory (DFT),” *Physics 480/680, Cornell University, Department of Physics*, Jan. 2004, http://people.ccmr.cornell.edu/ muchomas/P480/Notes/dft/dft.html.

[13] J. S. Bendat and A. G. Piersol, John Wiley & Sons, New York, 2nd edition, 1986.

[14] G. W. Elko, E. Diethorn, and T. Gänsler, “Room impulse response variation due to thermal fluctuation and its impact on acoustic echo cancellation,” in *Proc. Intl. Workshop on Acoust. Echo and Noise Control (IWAENC), Kyoto, Japan*, Sept. 2003.