BLIND UPMIX OF STEREO MUSIC SIGNALS USING MULTI-STEP LINEAR PREDICTION BASED REVERBERATION EXTRACTION

Keisuke Kinoshita, Tomohiro Nakatani

NTT Communication Science Laboratories
Kyoto, Japan

Masato Miyoshi

Kanazawa University
Ishikawa, Japan

ABSTRACT

We propose a blind upmixing method for stereo music signals that utilizes multi-step linear prediction and decomposes the input signals into reverberation and the dereverberated signals. The proposed method is directly motivated by our previously proposed dereverberation algorithm that was shown to dereverberate speech signals well. In this paper, we first analyze the behavior of the multi-step linear prediction and investigate the reverberation reduction/extraction strategy using a stereo music signal model. Based on the analysis, we show that the proposed method can perform a dereverberation and reverberation extraction based on the stereo music signal, and achieve an efficient blind upmix of stereo music by assigning its dereverberated signal to the front channels and extracted reverberation to the rear channels. In the experiment, we apply the proposed upmixing method to real stereo music signals, and confirm its effectiveness with an objective evaluation and a preference test.

Index Terms— blind upmix, stereo music, dereverberation, preference test

1. INTRODUCTION

Despite the increasing popularity of multi-channel audio reproduction systems such as home theater systems and automotive audio, the number of multi-channel audio recordings available to the public is still limited. Although recent movie soundtracks and a few music recordings are available in discrete multi-channel format (e.g. 5.1 surround [1]), most legacy audio recordings are available only in a two-channel (i.e. stereo) formats. Thus, an algorithm that can blindly upmix or convert existing stereo music signals to three or more channel signals is desirable.

One of the easiest ways to achieve blind upmixing is to apply an artificial reverberation to the original stereo signal and assign those signals to the rear channels [2]. However, the resulting impression is essentially that of listening to the original recording in a virtual listening room. This artificial ambience information does not match the conditions in which the original recording was produced [3].

Recent blind upmixers [4][5][6][7] rely on a common principle of extracting reverberation-like components embedded within the recording and assigning them to the rear channels. The methods for accomplishing this rely on the assumption that those sound components that affect our perception of the reverberation (i.e. reverberance) have a relatively lower inter-channel correlation within the stereo audio signal, thus the removal of the correlated sound components will yield the reverberance imagery. However, the extracted reverberation-like components for left and right rear channels are characterized as reversed phase signals, and tend to produce an artificial impression.

In contrast, in this paper, we propose a blind upmixing method that aims to extract *actual reverberation* from stereo music signals, by utilizing our previously proposed dereverberation algorithm based on multi-step linear prediction (MSLP) [8][9]. In previous studies, the dereverberation method is shown to work effectively for speech signals [8].

This paper is organized as follows. We first analyze the behavior of the MSLP based reverberation extraction method using a stereo music signal model, and then confirm experimentally its dereverberation and reverberation extraction effect using real stereo music signals. Finally, we conduct a preference test to obtain a subjective evaluation of the proposed method as an upmixer in comparison with the state-of-the-art method.

2. SIGNAL MODEL

Here we introduce the target signal model dealt with in this paper, which is a stereo music signal with a reverberant center vocal and reverberant accompaniment signals in each channel. The center vocal signal $s(n)$ is recorded on each channel, m ($m = 1, 2$), after being reverberated through a transfer function, $H_m(z)$, and each channel contains an accompaniment signal $\nu_m(n)$ which is the sum of N instrumental or vocal signals. Then the signals on channel m can be represented mathematically as

$$x_m(n) = h_{m,i} * s(n) + \nu_m(n), \tag{1}$$

where $*$ denotes the convolution, and $h_{m,i}$ ($i = 0, 1, \ldots, L$) corresponds to the coefficients of $H_m(z)$. The accompaniment signal $\nu_m(n)$ can be expressed as:

$$\nu_m(n) = \sum_{l=1}^{N} g_{m,j}^{(l)} * \epsilon_m^{(l)}(n), \tag{2}$$

$g_{m,j}^{(l)}$ denotes, similarly to $h_{m,i}$, the coefficient of a transfer function for the l-th musical signal $\epsilon_m^{(l)}(n)$. Equivalently, in vector/matrix form, Eq. (1) can be rewritten as:

$$\mathbf{x}_n^{(m)^T} = \mathbf{s}_n{}^T \mathbf{H}_m + \mathbf{v}_n^{(m)^T}, \tag{3}$$

where T denotes the matrix transpose.

$$\begin{aligned}
\mathbf{x}_n^{(m)} &= [x_m(n), x_m(n-1), \ldots, x_m(n-p)]^T, \\
\mathbf{s}_n &= [s(n), s(n-1), \ldots, s(n-(L+p))]^T, \\
&\quad \leftarrow \quad p+1 \quad \rightarrow \\
\mathbf{H}_m &= \begin{pmatrix} \mathbf{h}_m & & & \uparrow \\ & \mathbf{h}_m & \mathbf{0} & \\ & & \ddots & L+p+1 \\ & \mathbf{0} & & \\ & & \mathbf{h}_m & \downarrow \end{pmatrix}, \\
\mathbf{h}_m &= [h_{m,0}, h_{m,1}, \ldots, h_{m,L}]^T, \\
\mathbf{v}_n^{(m)} &= [\nu_m(n), \nu_m(n-1), \ldots, \nu_m(n-p)]^T.
\end{aligned}$$

In this study, the center vocal signal and each accompaniment signal are assumed to be uncorrelated.

3. PROPOSED BLIND UPMIXING METHOD

In the proposed method, it is assumed that reverberation can be divided into two parts, namely early and late reflections: early reflections correspond to those caused by the first τ coefficients of $H_m(z)$, $h_{m,0}, \ldots, h_{m,\tau}$, and late reflections correspond to the entire latter reverberation. Note that the dereverberation method that directly motivates the proposed method is designed to suppress the reflections caused by $h_{m,\tau+1}, \ldots, h_{m,L}$ (i.e. late reflections). It

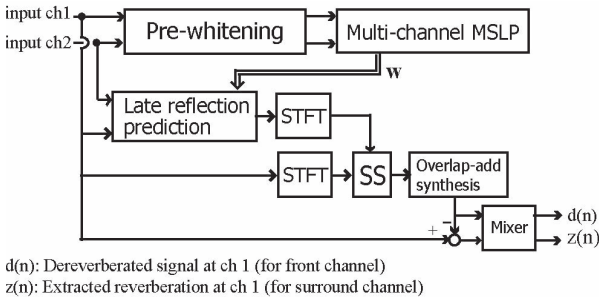

Fig. 1. Schematic diagram of proposed upmix method (processing for channel 1)

Table 1. A summary of the proposed method (for channel 1)

The steps from 1) to 6) correspond to the dereverberation process, and step 7) to reverberation extraction.

1) Pre-whitening is applied to input signals.
2) τ-step predictors for estimating the late reflections are calculated based on the pre-whitened signals.
3) The late reflections are estimated by applying a τ-step predictor to the τ-sample delayed version of the input signals.
4) The estimated late reflections and the input signals at channel 1 are both divided into short time frames with Hamming windows, and their power spectra are calculated with a short term Fourier transform (STFT).
5) The power spectrum of the estimated late reflections is subtracted from that of the input signals. This procedure is referred to as spectral subtraction (SS) in Fig. 1.
6) The resulting spectrum is converted back to a time-domain signal with an inverse STFT and the overlap-add technique. The phase of the observed signal at microphone 1 is used for the signal synthesis.
7) The dereverberated signals are subtracted from the input signals to extract the embedded reverberation.
8) The signals for the front and rear channels are obtained by appropriately mixing the dereverberated signal and the extracted reverberation.

is preferable to preset τ, the control parameter for the performance of late reverberation estimation, at $\tau > \tau_o$, if we can assume a certain time lag, τ_o, after which the autocorrelation of the center vocal signal and accompaniment signal becomes fairly small.

The processing flow of the proposed method is summarized in Fig. 1 and Table 1. As we can see in the table, the proposed method first dereverberates the signal and then extracts the reverberation from the input signal based on the dereverberated signal. This section mainly details the dereverberation process focusing on steps 2) to 6) in Table 1, and outlines its potential performance with stereo music signals.

3.1. Late reflection estimation with multichannel multi-step linear prediction

Now we introduce and analyze the MSLP for estimating late reflections contained in stereo music signals [8].

3.1.1. Multi-step linear prediction

Here let us consider a multichannel MSLP system with two input channels. The input signal at channel 1 is predicted with signals from both channels processed with the τ-sample delay units as:

$$x_1(n) = \sum_{m=1}^{2} w_m(n) * x_m(n-\tau) + y_1(n) \qquad (4)$$

Hereafter $w_m(n)$ and $y_1(n)$ are referred to as τ-*step* predictors and the prediction residual. τ-*step* predictors can be obtained by minimizing the mean square energy of the prediction residual.

3.1.2. Analysis of behavior

Here we analyze the behavior of the cost function for τ-*step* predictors using the signal model introduced in section 2, and explain how MSLP may estimate the late reflections. First, let us substitute Eq. (3) into eq (4).

$$\begin{aligned} y_1(n) &= x_1(n) - \sum_{m=1}^{2} \{\mathbf{s}_{n-\tau}{}^T \mathbf{H}_m \mathbf{w}_m + \mathbf{v}_{n-\tau}^{(m)}{}^T \mathbf{w}_m\}, \qquad (5) \\ &= \{\mathbf{s}_n{}^T \mathbf{h}_1 + \nu_1(n)\} - (\mathbf{s}_{n-\tau}{}^T \mathbf{H}\mathbf{w} + \mathbf{v}_{n-\tau}{}^T \mathbf{w}), \end{aligned}$$

where

$$\begin{aligned} &\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2], \\ &\mathbf{v}_n = [(\mathbf{v}_n^{(1)})^T, (\mathbf{v}_n^{(2)})^T]^T. \\ &\mathbf{w}_m = [w_0^{(m)}, w_1^{(m)}, \ldots, w_p^{(m)}]^T, \\ &\mathbf{w} = [\mathbf{w}_1{}^T, \mathbf{w}_2{}^T]^T. \end{aligned}$$

Then, the minimization of the mean square energy of the prediction error $y_1(n)$ leads to the minimization of the following cost function.

$$\begin{aligned} f[\mathbf{w}] &= E\{|y_1(n)|^2\} \\ &= E\{|\mathbf{s}_n{}^T \mathbf{h}_1 - \mathbf{s}_{n-\tau}{}^T \mathbf{H}\mathbf{w}|^2\} + E\{|\nu_1(n) - \mathbf{v}_{n-\tau}{}^T \mathbf{w}|^2\}, \qquad (6) \end{aligned}$$

where $|a|$ and $E\{a\}$ denote the absolute value of a and the time average of a, respectively. With the assumption that $s(n)$ and $v_m(n)$ are uncorrelated, the covariance matrix $E\{\mathbf{s}_n \mathbf{v}_n{}^T\}$ is assumed to be zero. The first term of Eq. (6) coincides with the MSLP cost function for estimating the late reflections of the center vocal signal. The second term, similarly to the first term, can be viewed simply as the cost function of the late reflection estimation for the accompaniment signals. In total, Eq. (6) tells us that this cost function adjusts the accuracy of the late reverberation estimation for each signal component according to the energy of its prediction residual signal, which ideally is the energy of the precisely dereverberated signal. That is, if one signal component in the music has more energy, more accuracy would be assigned to its late reflection estimation, and vice versa. With this cost function, we can obtain a prediction filter that can best predict, in an MMSE sense, the late reflections of all the signal components in the music signal.

3.2. Late reflection removal

In this section, we discuss a method for removing late reflections, and explain the effectiveness of steps 4) to 6) in Table 1.

3.2.1. Late reflection removal in time domain

One way to achieve a dereverberation with $\mathbf{w}$ is to obtain $y_1(n)$ using Eq. (5), which is equivalent to the traditional inverse filtering [9]. Hereafter, this process is referred to as time-domain subtraction.

To perform an exact dereverberation with a prediction filter using time-domain subtraction, the number of source signals contained in the input signal has to be 1 for a 2-channel input (i.e. $\mathbf{v}_n = 0$) [10]. In other words, if the input signal contains more than two source signals (i.e. $\nu_m(n) \neq 0$), it is theoretically impossible for a prediction filter-set to achieve an accurate dereverberation of each source. This means that the estimated late reflections for a certain source signal always contain some degree of unavoidable errors when $\nu_m(n) \neq 0$.

What happens if the estimated late reflections contain some degree of error? Fig. 2 summarizes the effect of time-domain subtraction from the viewpoint of impulse response equalization. "Input" stands for an example of an impulse response for a certain source signal. "Estimated late reflections" stands for that of the estimated late

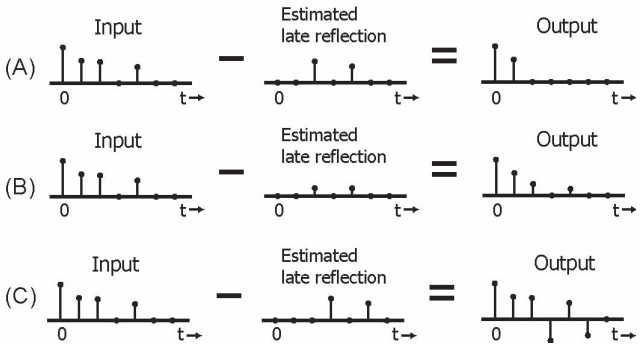

Fig. 2. The effect of time-domain subtraction as regards impulse response equalization: (A) time-domain subtraction without estimation errors in late reflections, (B) time-domain subtraction when an estimated late reflections contain an error in amplitude, and (C) time-domain subtraction when the estimated late reflections contain the error in the phase

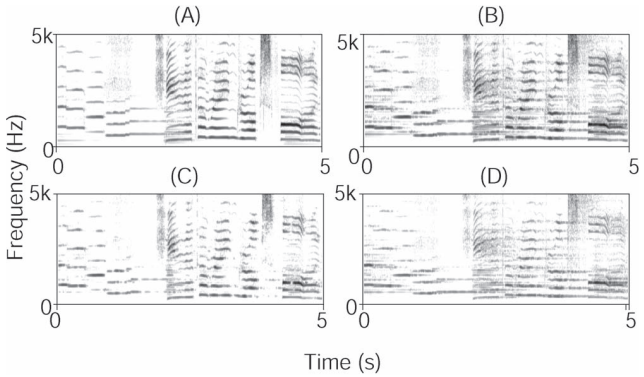

Fig. 3. Spectrograms: (A) music signal without reverberation, (B) music signal with reverberation, (C) dereverberated signal of (B), (D) reverberation extracted from (B)

reflections. "Output" stands for the equalized impulse response (i.e. dereverberated signal), which can be obtained simply by subtracting "Estimated late reflections" from "Input". For the sake of simplicity, τ was set at 2 in this figure. (A) shows the result of time-domain subtraction when the late reflections are accurately estimated. We can see that all the late reflections after the τ-th sample are efficiently suppressed, and accurate dereverberation is achieved. (B) shows the result of time-domain subtraction when the estimated late reflection contains the error only in the amplitude information. In this case, some portion of the late reflections still remains in the "Output" signal. While this problem seems somewhat relevant to all the signal enhancement methods such as noise reduction, phase estimation error causes a more fatal degradation in performance. (C) shows the effect of phase estimation error, where we express the phase error by a shear in time. We can see that the phase estimation error leads not only to the failure of the dereverberation but also to the production of extra reverberant components. Note that the impulse response of "Output" has longer reverberation tails than "Input".

Since these kinds of estimation error cannot be avoided when dealing with stereo music signals, it is better to avoid using time-domain subtraction.

3.2.2. Late reflection removal in power spectral domain

Here we present a method whose performance is less sensitive to the phase estimation error mentioned above. Another way to achieve dereverberation with $\mathbf{w}$ is to substitute subtraction in the power spectral domain for the time domain subtraction in Eq. (5). Hereafter this process is referred to as frequency-domain subtraction.

Frequency-domain subtraction is formulated as:

$$|\hat{S}_1(kM,\omega)| = \begin{cases} \sqrt{|X_1(kM,\omega)|^2 - |R(kM,\omega)|^2}, \\ \quad (\text{if } |X_1(kM,\omega)|^2 - |R(kM,\omega)|^2 \geq 0) \\ 0, \ (\text{otherwise}) \end{cases} \quad (7)$$

where M, k, X_1 and R correspond to the frame length, the frame index, the STFT of the input signal and that of the estimated late reflections, respectively. To synthesize the complex spectrum of the dereverberated signal, we simply employ the phase of the input signal, $\angle X(kM,\omega)$, as

$$\hat{S}_1(kM,\omega) = |\hat{S}_1(kM,\omega)| e^{j\angle X(kM,\omega)}, \quad (8)$$

Note that, even though the late reverberation estimation contains the error in phase (or shear in time) as in Fig. 2 (C), frequency-domain subtraction can ignore these types of errors as long as the shear in time is less than the analysis frame length M. Moreover, even if the shear in time is greater than M, frequency-domain subtraction always reduces the energy of the target signal due to the flooring effect in Eq. (7), thus the process never produces extra reverberant components in contrast to (C) in Fig. 2. Consequently, we can expect frequency-domain subtraction to achieve the joint dereverberation of the center vocal signal and the accompaniment signals simultaneously in combination with the MSLP cost function mentioned in Eq. (6).

4. OBJECTIVE EVALUATION

In this section, we objectively evaluate the proposed method in terms of dereverberation and reverberation extraction performance. Fig. 3 shows spectrograms of (A) a music signal without reverberation, (B) a music signal with reverberation, (C) the dereverberated signal of (B), and (D) reverberation extracted from (B). The illustrated segment contains the signals of a flute, a piano and a female singing voice. The first half of the segment is dominated by the flute, whereas the latter half is dominated by the singing voice. The music signal (B) is generated by recording each instrumental signal and singing voice separately in the recording studio, then adding a different reverberation to each of them, and finally summing them all up with an appropriately adjusted mixing ratio. The signal (A) is generated in the same manner as (B) but without adding reverberation, thus it can be considered as an ideal dereverberated signal. As we can see, the dereverberated signal (i.e. (C)) exhibits similar characteristics to (A). Furthermore, we see that the extracted reverberation (i.e. (D)) seems to coincide well with the characteristics of "signal (B) minus signal (A)", which is the ideally extracted reverberation.

We calculated the LPC cepstrum distance of (A) and (B), and (A) and (C) of Fig. 3, and found that the dereverberation reduces the distance from 2.1 dB (before) to 1.9 dB (after). If we calculate the LPC cepstrum distance based on the whole music signal (1 min.) including the above segment, the before and after values were 2.6 dB and 2.3 dB, respectively. These results may indicate that the proposed method could achieve the dereverberation of music signals and reverberation extraction with reasonable accuracy.

5. SUBJECTIVE EVALUATION

A preference test was conducted to evaluate the proposed method as an upmixer in comparison with a widely available upmixer, which, in this study, is Dolby Pro Logic II (DPL-II) [7].

5.1. Process parameters and stimuli

5.1ch signals are generated from the stereo music signals using the proposed method and DPL-II with the following parameter settings. The sampling frequency is 48 kHz, and the delay τ and the filter length p for multi-step LP is 5760 (=0.12 ms), 7200 (=0.15 ms), respectively. The frame length used for SS is 1764(=40 ms). The signals for the front and rear channels are generated by mixing the dereverberated signal and extracted reverberation with ratios of 10:1, and 1:10 respectively. A sub-woofer signal is generated by lowpass filtering the original signal with a cutoff frequency of 120 Hz. The

Table 2. List of music signals used for stimuli

Music genre No.	Kind of music.
1	classical (small orchestra)
2	classical (small orchestra)
3	classical (orchestra)
4	classical (orchestra)
5	opera (female singer with orchestra)
6	jazz (quintet including male singer)
7	rock (male singer with a band)
8	pops (female singer)

signal used for the center channel of the proposed method is same as that for DPL-II with an appropriately reduced amplitude.

To generate the upmixed sound based on DPL-II, we used DP564 manufactured by Dolby. The sub-woofer signal is the same as that of the proposed method, which is a simple lowpass filtered input signal.

We used the 8 different music signals listed in the Table 1 to generate sets of stimuli with the above parameters.

5.2. Subjects

Eight professional audio engineers participated in this experiments. They are aged between 29 and 48.

5.3. Procedures

The subjects were asked to sit and listen to the stimuli at the sweet spot in an ITU-R BS77-1-compliant surround listening environment. A set of stimuli consisted of 3 kinds of signals: original stereo music, a 5.1ch signal generated using the proposed method (proposed surround, hereafter), and a 5.1ch signal generated using DPL-II (conventional surround, hereafter). A set of stimuli was presented to a subject without information which surround signal was generated by which process. While evaluating a set of stimuli, the subjects were allowed to play the signal repeatedly, and to listen solely to the sound from each individual loudspeaker. After listening to the stimuli a sufficient number of times, the subjects were asked to report their preferred sound from the original stereo music, the proposed surround and the conventional surround. After the test, they were also asked to describe the criteria they used to decide their preference using such terms as spaciousness and localization.

Before the test, the subjects were provided with the following question:Which sound would you like the most as a consumer, if these sounds were played through your surround-sound system. When the presented signal was stereo, it was simply played through the front 2 loudspeakers of the surround-sound system.

5.4. Results

Figure 4 shows the results of the preference test. The preference score was calculated by adding 1 point to the best liked sound in a trial, and adding together all the points across the subjects for each music genre. The results show that the proposed surround was preferred for 6 of 8 genres. To analyze the results in more detail, we took a further look at the subjects' comments about genre numbers 5 and 6, where the conventional surround or the original stereo was preferred to the proposed surround. For opera (genre number 5), most of the evaluations seemed to attach more importance to the localization of the center vocal signal, thus the conventional surround that carefully maintains or even emphasizes the localization of the center signal was preferred. For jazz (genre number 6), some subjects commented that jazz music should not be upmixed in the first place. This is a general problem with surround reproduction, so it may be better to discard the results related to jazz when evaluating blind upmixing methods. For the rest of the music genres where the proposed surround was preferred, the subjects seemed to prefer the surround sound that is closest to their personal imagination of a natural reproduction of a stereo music signal in surround.

We conducted the Tukey's HSD (Honestly Significant Difference) test and ANOVA (ANalysis Of VAriance) with respect to the

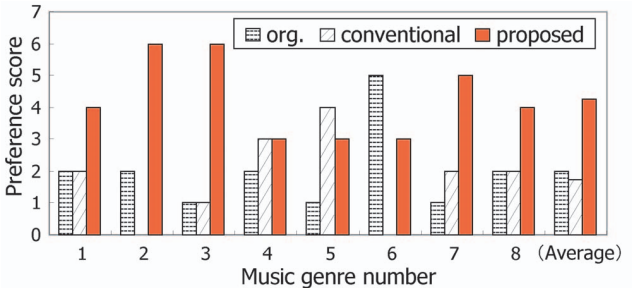

Fig. 4. Results of the preference test

average differences between the proposed surround and the conventional surround/original stereo music, and found that the differences were statistically significant ($p < 0.01$). On the other hand, the difference between the conventional surround and the original stereo music was not significant. Although currently the number of subjects is not large enough to conclude that the proposed surround is better than the conventional system, the obtained results encourage us to pursue a further investigation of the proposed method as a new upmixing strategy in contrast to the conventional methods.

6. SUMMARY

In this paper, we proposed a new blind upmixing method based on a multi-step linear prediction based reverberation extraction scheme that is motivated by our previously proposed dereverberation method. A mathematical analysis and objective evaluations showed that the proposed method could achieve the dereverberation of music signals and reverberation extraction with reasonable accuracy. To evaluate the proposed method subjectively as an upmixer, we then conducted a preference test with 8 professional audio engineers. In the test, the proposed method was compared with a widely available upmixer and an unprocessed original stereo music signal, and encouraging results were obtained revealing that the proposed surround was preferred to the conventional surround and the original stereo in 6 genres of 8. In the future, we will conduct preference tests with a larger number of subjects to obtain more statistically reliable results.

7. REFERENCES

[1] ITU-R BS.775.1, "Multichannel Stereophonic Sound System With and Without Accompanying Picture," 1994.

[2] D. R. Begault, "3-d sound for virtual reality and multimedia," *Academic Press, Cambridge*, pp. 226–229, 1994.

[3] C. Avendano and J.-M. Jot, "Ambience extraction and synthesis from stereo signals for multichannel audio upmix," in *Proc. Int'l Conf. Acoust., Speech, Signal Processing*, 2002, pp. 1957–1960.

[4] R. Irwan and R. M. Aarts, "Two-to-five channel sound processing," *J. Audio Eng. Soc.*, vol. 50(11), pp. 914–926, 2002.

[5] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1051–1064, 2006.

[6] J. Usher and J. Benesty, "Enhancement of spatial sound quality: A new reverberation-extraction audio upmixer," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2141–2150, 2007.

[7] R. Dressler, "Pro logic surround decoder principles of operation," `http:www.dolby.com`.

[8] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.

[9] G. B. Giannakis, Y. Hua, P. Stoica, and L. Tong, *Signal processing advances in wireless and mobile communications*, Prentice Hall, Upper Saddle River, NJ, 2001.

[10] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Speech Audio Processing*, vol. 36(2), pp. 145–152, 1988.