TRANSFORM-DOMAIN DECORRELATION IN DOLBY DIGITAL PLUS

Vinay Melkote, Kuan-Chieh Yen, Matt Fellers, Grant Davidson, and Vivek Kumar

Dolby Laboratories, Inc., San Francisco, CA 94103
{vmelk, kyen, mcf, gad, vzkuma}@dolby.com

ABSTRACT

The Dolby Digital Plus (DDPlus) multichannel audio codec employs channel coupling for parsimonious transmission of high frequency components of the signal, wherein the transform coefficients of discrete channels beyond a coupling-begin frequency are transmitted as a mono-downmix. The decoder reconstructs individual channels by appropriate panning with frequency-banded gains. While channel coupling is a valuable parametric coding tool for low bit-rate encoding of multichannel audio, the resulting high inter-channel coherence can affect audio post-processors that utilize downmixing, such as headphone virtualizers. To mitigate these effects, a new low-complexity spatial audio coding tool is proposed, whose primary objective is to restore phase diversity in the decoder's output. In this new approach, a decorrelation signal is synthesized directly from the decoded real-valued and critically-sampled transform (MDCT) coefficients, avoiding the expense of computing the imaginary counterpart (MDST) or signal transformation to a different domain (such as QMF). The decorrelation signal is then adaptively mixed with the dry signal and inverse-transformed as in a conventional decoder. The degree of mixing depends on spatial parameters that are either sent in the bitstream or estimated in the decoder from the discrete (not coupled) low-frequency spectral coefficients. Listening tests demonstrate the significant performance benefits obtained in either mode of operation of the proposed tool.

Index Terms— decorrelation, audio coding, channel coupling, MDCT, spatial coding

1. INTRODUCTION

The DDPlus codec (as well as its predecessor Dolby Digital) employs channel coupling to reduce coding bit-rate for multichannel content (such as 5.1ch audio), wherein beyond a specific "coupling-begin frequency" the modified discrete cosine transform (MDCT) coefficients of individual channels are transmitted as a mono downmix referred to as the coupling channel [1]. The DDPlus decoder pans the coupling channel back into the high frequency MDCT coefficients of the discrete channels using banded gains referred to as coupling coordinates. Channel coupling thus preserves the spectral magnitude while discarding phase information and is based on the premise that the human ear is insensitive to phase at high frequencies. It is a valuable parametric coding tool when no further processing of the decoded multichannel signal is intended. However, when the decoded signal is subject to any post-processing resulting in a downmix, for instance binaural rendition via headphone virtualization or playback as an Lo/Ro downmix, the coupled channels add-up coherently, leading to a timbre mismatch/unnatural brightness compared to the reference. Additionally the processed signal suffers from reduced externalization and source width. This motivates the requirement for a frequency selective decorrelation tool to restore the phase diversity lost in the coupling frequency range. While the decorrelator could be part of the virtualizer or the DDPlus decoder or an independent component by itself, integration into the DDPlus decoder provides the advantage that the signal is already in the frequency domain (for frequency selectivity), and further, a wealth of information (for instance, the band structure for coupling, or information related to the presence of transients) is readily available to guide the decorrelator. We thus propose an approach for restoring inter-channel correlation by applying decorrelation in the transform domain of the DDPlus codec.

In the proposed method, the sequence of MDCT coefficients at a particular frequency in each decoded channel is filtered via a suitably-defined all-pass filter to generate a decorrelation (or reverberation) signal for that frequency bin. The decorrelation signal is mixed with the direct signal such that the mixed-signal has a desired level of coherence with the direct signal. The modified MDCT coefficients are inverse transformed to generate the time domain signal. The mixing coefficients are determined by spatial parametric information which in one mode of operation, referred to as the forward-adaptive or guided mode, is computed at the encoder and transmitted in the DDPlus bitstream as part of auxiliary metadata (ignored by legacy decoders). Alternatively, in the backward-adaptive or non-guided mode the spatial parameters are estimated at the decoder from the discrete (i.e., non-coupled) lower frequency transform coefficients of the different channels. The backward-adaptive mode of operation enables application of the proposed decorrelation mechanism to legacy DDPlus content, but compared to the guided mode incurs higher decoder complexity and is slightly less efficient in terms of restoring the spaciousness and timbre. Since all-pass filtering is a linear operation, time-domain alias cancellation (TDAC) of the MDCT is almost perfectly preserved and the resultant decorrelation signal is natural sounding. Temporal control of the decorrelation process, in order to avoid excessive smearing of transients, is effected by suitable temporal shaping of the decorrelation signal prior to mixing, and by modification of the mixing coefficients themselves. A system overview is provided by Fig. 1.

Decorrelation has been widely used in audio coding, for instance, as part of the Parametric Stereo tool in HEAACv2 [2, 3] or in MPEG Surround [4]-[6]. However, unlike proposed here, these codecs apply decorrelation within a quadrature mirror filterbank (QMF) stage that wraps around the core transform domain audio codec - AAC. Decorrelation in the QMF domain does provide improved ability for temporal control of the decorrelator, albeit at increased computational expense. In [7] a transform-domain approach to decorrelation is proposed that derives the modified discrete sine transform (MDST) coefficients - the imaginary counterpart of the MDCT - from the transmitted MDCT coefficients, and the reverb signal is obtained as a linear combination of the two multiplied by the magnitude response of a reverb filter, to mimic the convolution of a reasonably short filter with the time-domain equivalent of the

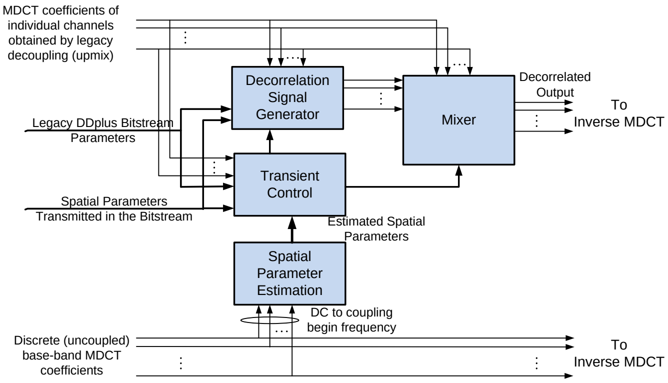
MDCT coefficients of individual channels obtained by legacy decoupling (upmix)
Decorrelation Signal Generator
Mixer
Decorrelated Output
To Inverse MDCT
Legacy DDplus Bitstream Parameters
Transient Control
Spatial Parameters Transmitted in the Bitstream
Estimated Spatial Parameters
Spatial Parameter Estimation
DC to coupling begin frequency
Discrete (uncoupled) base-band MDCT coefficients
To Inverse MDCT

Fig. 1. Overview of the proposed MDCT-domain decorrelator for DDPlus

direct (downmix) signal. The computation of the MDST entails (based on [8]) not just additional complexity, but also an additional delay of one frame since it requires MDCT coefficients from neighboring frames. Other relevant work related to decorrelation in audio coding cover various aspects of the process including choice of reverberation filters, calculation of predicition parameters (i.e., upmix coefficients) for component channels, and temporal shaping of the reverb signal [9]-[13].

2. THEORETICAL BACKGROUND

We will now describe some of the theoretical background that forms the basis of the system design described in Sec. 3.

2.1. Linear filtering in the MDCT domain and TDAC preservation

Let the column vector $\mathbf{x}_k$ denote the k^{th} frame of N MDCT coefficients of an audio channel. A decoder that recieves this frame multiplies $\mathbf{x}_k$ by the inverse MDCT kernel, $\mathbf{P}$, a matrix of dimension $2N \times N$, to create the aliased time-domain vector $\mathbf{u}_k$, of length $2N$. The time-domain signal is constructed by overlap-addition which cancels time-domain aliasing, and the z-transform of the alias-free reconstruction (in its entirety for the audio channel) can be represented as:

$$\mathbf{U}(z) = \sum_k e(z)\mathbf{u}_k z^{-kN} = \sum_k e(z)\mathbf{P}\mathbf{x}_k z^{-kN} \tag{1}$$

where $e(z) = [1\ z^{-1}\ \cdots\ z^{-(2N-1)}]$, and the term z^{-kN} aligns the different frames for overlap-addition. Consider linearly filtering the MDCT coefficients across time, i.e., running a filter in each MDCT bin, with the same filter shared across the entire spectrum. Thus, we generate a second frame of MDCT coefficients, $\mathbf{y}_k$, such that

$$\mathbf{y}_k = \sum_l h(l)\mathbf{x}_{k-l} \tag{2}$$

where $h(k)$ denotes the prototype filter. The z-transform of the audio signal represented by the sequence of frames $\{y_k\}$ is:

$$\begin{aligned} \mathbf{V}(z) &= \sum_k e(z)\mathbf{P}\mathbf{y}_k z^{-kN} \\ &= \sum_l h(l)z^{-lN} \sum_k e(z)\mathbf{P}\mathbf{x}_{k-l} z^{-(k-l)N} \\ &= \mathbf{U}(z) \sum_l h(l) z^{-lN} \end{aligned} \tag{3}$$

In other words $\mathbf{V}(z)$ is the audio channel $\mathbf{U}(z)$ processed with the N-times upsampled filter represented by $\sum_l h(l)z^{-lN}$, and is thus itself free of time-domain aliasing. The transform-domain generation of a decorrelation signal (by applying an all-pass filter to the MDCT coefficients in the coupling frequency range) in the proposed system is inspired by these arguments. We note that theoretically, the process of coupling in itself breaks the TDAC property of the MDCT (even if the coefficients are not quantized), in particular, for signal components on either side and close to the coupling begin frequency. However, any such non-cancellation of aliasing has not been audible in listening experiments.

2.2. Decorrelator guidance parameters and restoration of ICCs

Fig. 2 provides a vector geometry representation of an instance of channel coupling, where two channels denoted as l_{in} and r_{in} are downmixed to a coupling channel x_{mono}. More generally, there could be more than two input channels and the coupling channel may be a mono downmix where the individual channels are not necessarily equally weighted, or the coupling channel may be normalized by a factor such as the time-varying rms value across channels. Further, not all channels need to participate in coupling, and the channels to couple can be chosen separately for each DDPlus block. In Fig. 2, l_{in} could be interpreted, for instance, as a vector of MDCT coefficients in the same coupling band (a frequency region associated with a single coupling coordinate). The correlation coefficient between the two channels and the mono downmix are α_L and α_R, respectively. The encoder transmits (a coded version of) x_{mono} in the bitstream from which a legacy decoder reconstructs l_{in} simply as $g_L x_{mono}$, where g_L is the coupling coordinate for the channel. A new decoder with integrated decorrelator instead generates a reverb signal of MDCT coefficients, y_L, that is substantially uncorrelated with the direct signal, and mixes the direct and reverb signals so that the resultant output l_{out} bears the same correlation α_L with x_{mono} as the original signal l_{in}. (Henceforth, the term *alpha parameter* or *mixing parameter* refers to the coherence of a channel w.r.t the downmix.) The process is similarly extended to the remaining channel.

The end objective of the decorrelator, however, is not the reconstruction of the correct coherence between the mono downmix and each output channel but rather the preservation of inter-channel coherence (ICC). In this example, the ICC between l_{out} and r_{out} should be similar to the ICC between l_{in} and r_{in}. The ICC between output channels can be controlled by appropriately modifying the coherence between the decorrelation signals, y_L and y_R. In general, a choice of decorrelation signals that are orthogonal to each other (achieved by independently designing decorrelation filters for each channel) does not lead to the correct output ICC. In the specific example here, where the coupling channel is a straight downmix of the two input channels, it can be shown that the optimal coherence between y_L and y_R (for ICC preservation) is -1, which is implicit in Fig. 2. Thus, the decorrelation signals for the two channels need be sign flipped versions of each other (up to a scaling). The proposed decorrelator is controlled by transmitting/estimating alpha parameters, while exploiting the above theory of sign-flipped decorrelation signals to approximately preserve the ICCs.

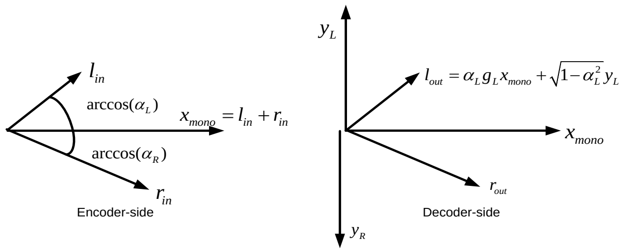

Fig. 2. Coupling two channels and upmixing with decorrelation

3. DESIGN OF THE DECORRELATOR

We now describe individual modules of the proposed decorrelation system in more detail.

3.1. Decorrelation Signal Generator

A three-pole all-pass filter preceded by a fixed delay of 1 MDCT sample (equivalent in the time-domain to the stride of a DDPlus block, 256 samples) is employed to filter the MDCT coefficients of each bin in the coupling frequency range of a channel. The MDCT coefficients of the coupling channel are first upmixed into individual channels via straight-forward panning as in a legacy decoder, and the upmixed coefficients for each channel are then filtered to generate the channel-specific reverb signals. The primary reason for this strategy (as opposed to directly filtering the coupling channel) is to undo typical time-varying normalization applied during construction of the coupling channel at the encoder: filtering MDCT samples that have been normalized differently across time results in artifacts.

The center channel is not decorrelated to avoid smearing of voice, and the LFE channel is not coupled. Therefore, in the 5.1ch case decorrelation signals are produced only for the L, R, Ls, and Rs channels. Appealing to the arguments in Sec. 2.2, we choose the all-pass filters for the four channels to be the same while the outputs of the filter (in each MDCT bin) for R and Ls are multiplied by -1. If the coupling channel is dominated by contributions from two adjacent channels then it can be seen that this approach preserves the spatial image (and the most important ICCs) quite well.

3.2. Mixer

The use of all-pass filters results in reverb signals for each channel that are quite uncorrelated with the input to the filters, while retaining the same power profile. Thus, the mixer generates a modified set of MDCT coefficients by simply mixing the direct and reverb signals in a power preservation sense, i.e., by weighting with appropriate α and $\sqrt{1-\alpha^2}$ as exemplified in Fig. 2. The alpha parameters themselves may have been modified in other modules as described below.

3.3. Transient Control

The assumption that the all-pass filter generates an output with the same temporal power distribution as the input, at each frequency, holds well for stationary or quasi-stationary signals. Transients, on the other hand, appear delayed in the reverb signal (not in synchrony with the direct signal), while the natural response of the all-pass filter results in time-smearing. Thus, audible double-hits and reduced sharpness are introduced. Spaciousness, however, is less important during transient events. These observations motivate the transient control features implemented in the proposed decorrelator.

First, a set of transient flags based on existing DDPlus bitstream elements and decoder-based energy measurements are calculated per-channel per-DDPlus block. The transient flag is really a confidence measure in the detector, and takes values ≥ 0 and ≤ 1, where 1 indicates a transient with high confidence and 0 indicates non-transient behavior with high probability. The flag values are exponentially decayed back to 0 over a transition period, and reset at an event that triggers an instantaneous value of the flag that is higher than the decayed value at that time. Block-switching is employed in DDPlus to cope with hard transients, where the typical DDPlus block with a stride of 256 samples is split into two smaller blocks of 128 samples stride each. Since the resolution of MDCT coefficients of the smaller blocks is different from the more typical long block, channels that are in block-switching are not coupled. When the transient control module sees a set block-switch flag, the transient flag for the channel is set to to 1 and additionally the input samples to the filters for that channel are zeroed out, so that the all-pass filters linearly combine MDCT coefficients of the same frequency resolution always. To avoid double hits, decorrelation is completely stopped (by changing alphas of the appropriate channel to 1) when the transient flag for the channel is evaluated as unity, while normal mixing operation is affected when the value is 0. In between values of the flag result in proportionately reduced decorrelation.

Second, in order to ameliorate transient-smearing, per-channel per-block per-coupling band temporal shaping gains are calculated based on the decay characteristics of the all-pass filter and the power envelop of its input (the direct signal), and applied to the decorrelation signal to “duck” its temporal envelop below that of the direct signal during a transient event.

3.4. Forward-adaptive mode: quantization of alphas

The encoder calculates the alpha parameters for a channel at the same frequency resolution (banding structure) as the coupling coordinates. Alphas are calculated as the correlation-coefficient between the transform coefficients of a band of the original channel and the coupling channel, and can be shared between adjacent DDPlus blocks, in which case they are averaged across the time-share before quantization.

A multi-stage quantization approach motivated by the emperically observed correlation between alpha parameters across channels and frequency is employed. The alphas of the first coupling band of all channels to be decorrelated are jointly quantized using a vector quantizer (VQ). The alphas for subsequent coupling bands, within each channel, are differentially encoded. The differentials of at most 4 consecutive bands are collected together and vector quantized. For the first group of 4 bands the difference is calculated w.r.t the reconstructed alpha of the first band of the channel (that is quantized using the inter-channel VQ), and for subsequent groups it is w.r.t the quantized alpha of the last band already processed. The number of channels downmixed into the coupling channel influences the value of the alphas, and hence a retransmission of alphas is mandatory whenever channels move in and out of coupling. When encoding 5.1ch audio content at 192 kbps the additional bit-rate required for the decorrelation guidance information was on an average a mere 1.5 kbps.

3.5. Backward-adaptive mode: estimation of alphas

The backward-adaptive mode provides a means for applying decorrelation to legacy bitstreams that do not contain side-chain data required by the forward-adaptive mode. The decoder uses the decoded MDCT coefficients in the discrete lower frequency range, i.e. below the coupling-begin frequency, to calculate the alpha parameters between an individual channel and a composite downmix of the channels locally created at the decoder, which are then extrapolated to obtain the alphas at frequencies above the coupling begin frequency. The extrapolation is a scaled copy of the alpha from the lower frequencies, with the scaling constant successively decreasing for higher frequencies. The extrapolated alpha values are then dithered via empirically derived rules based on statistics gathered from a large set of diverse audio content, so as to impose the same second-order statistics to the alphas across time and frequency that existed for the analyzed content. The variance of the dither depends on the extrapolated alpha value for the band (closer this is to 1, lower the added dither: content that is coherent between channels at lower frequencies results in the higher frequencies being coherent as well) and the band index (the alphas tend to get more chaotic moving up the frequency scale).

4. EXPERIMENTAL RESULTS

A blind subjective test was conducted to evaluate the performance of the proposed approach compared to legacy DDPlus when the decoded audio is rendered via headphone virtualization. The source content used was 5.1ch, 48kHz sampling rate PCM audio. Dolby Headphone v1 (DHv1) was the choice of virtualizer. Six expert listeners graded the three systems:

- Legacy DDPlus followed by DHv1
- Proposed DDPlus with backward-adaptive decorrelation (BA Decorr. in Fig. 3) followed by DHv1
- Proposed DDPlus with forward-adaptive decorrelation (FA Decorr. in Fig. 3) followed by DHv1

against the reference original (also rendered via DHv1) in terms of the following attributes relevant to characterization of headphone virtualization: overall quality, dialog clarity, spectral naturalness (or timbre), externalization (out-of-head experience), source width (perceived width of the audio scene), and sound scene consistency (preservation of motion-trajectory/placement of objects in the scene). The order of codecs was randomized and grading was on a 5-point scale defined in Table. 1. The encoding bit-rate for all three systems under test was 192 kbps and the coupling-begin frequency, 3.42kHz. A large corpus of content coded with DDPlus was screened to select critical items, i.e., in which artifacts due to channel coupling were most perceptible. The test eventually employed seven audio sequences: (a) Applause, (b) dialog section from the movie Inception with helicopter in the background, (c) soundtrack from an Indiana Jones movie, (d) rain, (e) sea-wash, (f) crowd noise and applause from a Superbowl Halftime clip, and (g) crowd chanting (speech in surround channels) from the movie Gladiator.

The mean per-attribute scores from the subjective test are provided in Fig. 3 along with their 95% confidence intervals. The average is computed across test-subjects and audio items. The positive impact of integrating the decorrelator into the DDPlus decoder is evident. The improvement on four of the five spatial attributes spectral naturalness, externalization, source-width, sound scene consistency is significant, with both systems with decorrelation (forward-adaptive and backward-adaptive) judged on an average as perceptibly different from the uncoded virtualized reference, but hard to distinguish. In contrast, the legacy DDPlus system scores about 1 point lower on the 5-point scale on all four attributes. The Overall Quality is improved due to decorrelation by about 0.8-0.9 points in a statistically significant sense, from between Poor and Fair to between Fair and Good, while dialog quality is not impacted.

Grade	Overall Quality	Other Attributes
5	Excellent	Imperceptible
4	Good	Perceptible difference, but hard to distinguish
3	Fair	Slightly different
2	Poor	Different
1	Bad	Very Different

Table 1. Correspondence between 5-point scale in the test and subjective quality

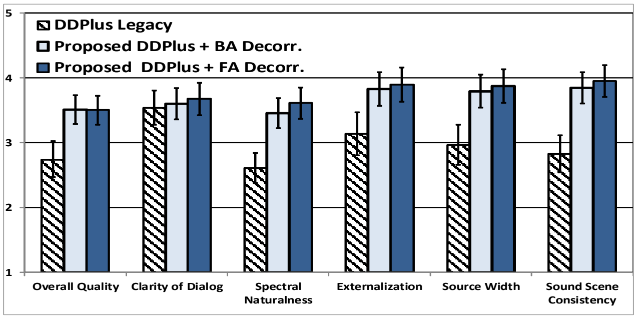

Fig. 3. Subjective comparison of legacy DDPlus and the proposed decorrelator in backward adaptive and forward adaptive modes, when followed by the DHv1 virtualizer

In Fig. 3 the forward-adaptive method appears to fare only marginally better than the backward-adaptive decorrelator. However, we note that the benefits of transmitting side-information can be significant based on the content. For the Inception dialog test item, the forward-adaptive approach scored better than the backward-adaptive method by about 0.75 pts on spectral naturalness and sound scene consistency. The remaining attributes were improved compared to the backward-adaptive approaching by about 0.2-0.5 pts on the 5 point scale. The forward-adaptive approach provides benefit in particular for signals where the higher frequencies predominantly belong to a different source than the lower frequencies. In this item the higher frequencies primarily correspond to the helicopter and ambient noise, while the lower frequencies mostly contain dialog. Thus the error in the estimation of correlation parameters in the backward-adaptive mode is higher.

5. CONCLUSION

A transform-domain decorrelator to alleviate timbral artifacts and spatial image collapse caused by channel coupling in DDPlus, observed when the codec is followed by a downmixing process, is proposed. In a forward-adaptive mode, the decorrelator utilizes side-information sent by the encoder to guide the mixing of direct and decorrelation signals, while in an alternate backward-adaptive mode it estimates the mixing weights from the discrete (non-coupled) low frequency MDCT coefficients. In the latter mode the decorrelator can act on legacy bitstreams as well. A subjective test to evaluate the efficacy of the decorrelator with headphone virtualization evidences the significant performance improvement over legacy DDPlus.

6. REFERENCES

[1] Advanced Television Systems Committee, Document A/52B, *"Digital Audio Compression Standard (AC-3, E-AC-3), Revision B"*, June 2005.

[2] ISO/IEC 14496-3:2001/Amd 2:2004, *"Parametric coding for high-quality audio"*, 2004.

[3] J. Breebart et al, "Parametric Coding of Stereo Audio," *EURASIP Journal on Applied Signal Processing*, September 2005.

[4] ISO/IEC 23003-1:2007, *"Information technology – MPEG audio technologies – Part 1: MPEG Surround"*, 2007.

[5] S. Quackenbush and J. Herre, "MPEG Surround," *IEEE Multimedia*, October 2005.

[6] J. Herre et al., "The Reference Model Architecture for MPEG Spatial Audio Coding," in *118th AES Convention, preprint 6477*, May 2005.

[7] K. Suresh and T.V. Sreenivas, "MDCT Domain Analysis and Synthesis of Reverberation for Parametric Stereo Audio," in *123rd AES Convention, preprint 7281*, October 2007.

[8] C. I. Cheng, "Method for estimating magnitude and phase in the MDCT domain," in *116th AES Convention, preprint 6091*, May 2004.

[9] G. Hotho, L. Villemoes, and J. Breebart, "A Backward-Compatible Multichannel Audio Codec," *IEEE Tran Audio Speech and Lang Proc*, January 2008.

[10] C. Faller and F. Baumgarte, "Binaural Cue CodingPart II: Schemes and Applications," *IEEE Tran on Speech and Audio Proc*, November 2003.

[11] H. Purnhagen, "Low Complexity Parametric Stereo Coding in MPEG-4," *Proc. 7th Intl. Conf. Digital Audio Effects*, October 2004.

[12] M. R. Schroeder, "Synthesis of Low-peak-factor Signals and Binary Sequences with Low Autocorrelation," *IEEE Trans. Information Theory*, January 1970.

[13] J-M. Jot and A. Chaigne, "Digital Delay Networks for Desiging Artificial Reverberators," in *90th AES Convention, preprint 3030*, February 1991.