PRIMARY-AMBIENT SIGNAL DECOMPOSITION AND VECTOR-BASED LOCALIZATION FOR SPATIAL AUDIO CODING AND ENHANCEMENT

Michael M. Goodwin, Jean-Marc Jot

Creative Advanced Technology Center
Scotts Valley, CA
mgoodwin,jmj@atc.creative.com

ABSTRACT

Spatial audio coding and enhancement address the growing commercial need to store and distribute multichannel audio and to render content optimally on arbitrary reproduction systems. In this paper, we discuss a spatial analysis-synthesis scheme which applies principal component analysis to an STFT-domain representation of the original audio to separate it into primary and ambient components, which are then respectively analyzed for cues that describe the spatial percept of the audio scene on a per-tile basis; these cues are used by the synthesis to render the audio appropriately on the available playback system. The proposed framework can be tailored for robust spatial audio coding, or it can be applied directly to enhancement scenarios where there are no rate constraints on the intermediate spatial data and audio representation.

Index Terms— spatial audio coding, multichannel audio, upmix, principal component analysis

1. INTRODUCTION

Low-rate coding of audio signals is now a cornerstone of consumer electronic devices and systems. While audio content is still primarily in stereo format, multichannel audio is becoming increasingly available and popular; there is thus a growing need to distribute and store multichannel content, which in turn necessitates further advances in compression technologies – since discrete coding of individual channels is insufficient to satisfy storage and delivery bandwidth constraints. Furthermore, multichannel loudspeaker configurations are becoming more commonly deployed in home theater and music systems, which creates a commercial need for expanding legacy stereo content to a multichannel format to make the best use of the available rendering resources, *i.e.* however many loudspeakers are present. A recently emerging approach known as *spatial audio coding* (SAC) addresses the need to efficiently compress multichannel audio, and a range of upmix techniques are being developed to enhance the reproduction of stereo signals over multichannel loudspeaker formats. In this paper, we describe an approach to spatial audio analysis-synthesis which enables coding at low data rates and flexible rendering and spatial enhancement at the decoder; the framework is also applicable to enhancement scenarios where there is no intermediate coding channel, *i.e.* where there is no need for low-rate communication between the analysis and synthesis modules.

In most SAC schemes proposed in the literature, the multichannel input audio is analyzed in a pairwise fashion to extract interchannel information such as signal level differences and coherence [1]. These inter-channel relationships are sent as side information with a coded downmix signal; at the decoder, the downmix is distributed over the multichannel loudspeaker system using the inter-channel data to approximately recreate the inter-channel relationships of the input signals [2]. If the output format (speaker layout) does not match the input format (channel configuration), the rendered audio scene will be inconsistent with the input signal. The spatial audio coding system proposed in [3, 4] overcomes this difficulty by using format-independent *universal spatial cues* as the side information; these cues describe the audio scene in terms of spatial-perceptual parameters without reference to the channel configuration, thereby enabling consistent rendering on arbitrary output systems as well as flexible modifications or enhancements.

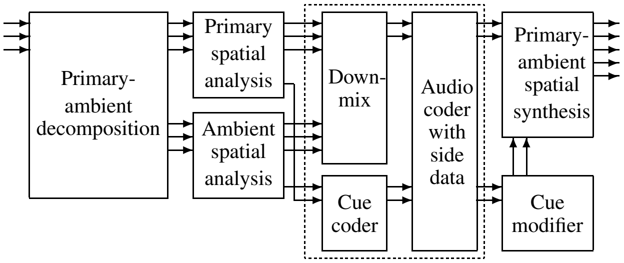

Fig. 1. Block diagram of primary-ambient spatial coding and enhancement system. The modules in the dashed box are needed for coding but not for some enhancement scenarios.

The system discussed in this paper expands on the SAC framework of [3, 4] by incorporating a primary-ambient decomposition, distinct spatial analysis for the primary and ambient components, modification of the spatial cues prior to synthesis, and spatial enhancement of the rendered components. Fig. 1 shows a block diagram of the spatial analysis-synthesis system. After conversion to the short-time Fourier transform (STFT) domain or some other time-frequency representation (not shown in Fig. 1), each channel signal is decomposed into primary and ambient components; a multichannel principal component analysis (PCA) algorithm to achieve this separation is discussed in Section 2. The primary and ambient components are then analyzed for spatial information; using a vector theory of sound localization, they are respectively aggregated across the channels into a spatial percept at each time and frequency. Section 3 discusses this spatial analysis as well as consistent synthesis based on the spatial cues derived from the input scene.

2. PRIMARY-AMBIENT DECOMPOSITION

In spatial audio analysis-synthesis, it is effective to treat discrete point-like sources and diffuse sounds differently. For instance, a point source should be rendered at a precise location via an appropriate discrete panning method. Diffuse or ambient sounds call for an alternate rendering technique, perhaps with additional decorrelation introduced to create a desired sense of spaciousness or envelopment;

an example is the use of ambience extracted from stereo signals to generate synthetic surround signals for 2-to-5 upmix [5]. Since a distinction between primary and ambient components is useful for high-fidelity enhancement and reproduction, we are interested in a multichannel primary-ambient signal decomposition. In [5], an approach is proposed which involves creating a time-frequency mask to extract the ambience from a stereo input signals. The mask is based on the cross-correlation between the left-channel and right-channel signals, however, so this approach is not immediately applicable to the problem of extracting ambience from an arbitrary multichannel input. To use any such correlation-based method in this higher-order case would call for a hierarchical pairwise correlation analysis, which would entail a significant computational cost, or some alternate measure of multichannel correlation. Rather than take such an approach, we incorporate the desired properties of the primary and ambient components to derive a multichannel separation algorithm based on principal component analysis (PCA).

2.1. Multichannel decomposition

The multichannel primary-ambient decomposition algorithm proposed in this section is based on a signal model wherein each STFT subband is treated as a vector in time and each channel vector $\vec{X}_m$ is modeled as a sum of a primary component $\vec{P}_m$ and an ambience component $\vec{A}_m$; the primary components of the various channels are scaled versions of a common unit vector $\vec{v}$:

$$\vec{X}_m[k,l] \;=\; [x_m[k,l] \;\; x_m[k,l-1] \;\; \cdots]^T \tag{1}$$

$$\vec{X}_m[k,l] \;=\; \vec{P}_m[k,l] \,+\, \vec{A}_m[k,l] \tag{2}$$

$$\;=\; \rho_m[k,l]\vec{v}\,[k,l] \,+\, \vec{A}_m[k,l] \tag{3}$$

where k is a subband index and l is a time index. In the following, the $[k,l]$ indices will at times be dropped to simplify the notation.

If the channel vectors defined in Eq. (1) are accumulated into a signal matrix

$$\mathbf{X} = \left[\vec{X}_1[k,l] \;\; \vec{X}_2[k,l] \;\; \vec{X}_3[k,l] \;\cdots\; \vec{X}_M[k,l]\right], \tag{4}$$

the primary-ambient signal model can be expressed as

$$\mathbf{X} \;=\; \mathbf{P} \,+\, \mathbf{A} \tag{5}$$

$$\;=\; \vec{v}\,[\rho_1 \; \rho_2 \;\cdots\; \rho_M] \,+\, \left[\vec{A}_1 \;\; \vec{A}_2 \;\cdots\; \vec{A}_M\right]. \tag{6}$$

To fit the signal to this model, we make a number of assumptions that are reasonable for typical audio content: the primary components have higher energy than the ambience; the ambience energy in the various channels is relatively balanced; the primary and ambient components are orthogonal in signal space, *i.e.* uncorrelated.

To derive the primary-ambient signal model of Eq. (6) according to the above assumptions, the key task is to find the unit vector $\vec{v}$ which best describes the set of channel vectors in signal space. Then, each channel can be separated into orthogonal primary and ambient components by projecting onto $\vec{v}$:

$$\vec{P}_m \;=\; \left(\vec{v}^H\vec{X}_m\right)\vec{v} \tag{7}$$

$$\vec{A}_m \;=\; \vec{X}_m - \vec{P}_m. \tag{8}$$

The projection $\vec{P}_m$ is then the primary component of the m-th channel signal, and the difference $\vec{A}_m$ is the ambient component. Note that by definition the primary and ambient components add up to the original, so no signal information is lost in the decomposition.

To this point, we have not yet discussed how to actually determine the primary unit vector $\vec{v}$. The best choice of course depends on the optimization criterion for the signal model. If we adhere to the assumption that the primary component should have maximal energy, then a reasonable optimization criterion is the mean-squared error. Namely, the best $\vec{v}$ is the one that results in the most energy in the primary component, *i.e.* it minimizes the energy in the residual ambience. While the resulting optimization is well known, we outline it here for the sake of completeness. First, it can be easily shown that for any given $\vec{v}$, the minimization of the ambience energy yields

$$\mathrm{tr}(\mathbf{A}^H\mathbf{A}) \;=\; \mathrm{tr}(\mathbf{X}^H\mathbf{X}) \;-\; \vec{v}\,\mathbf{X}\mathbf{X}^H\vec{v}. \tag{9}$$

The optimal $\vec{v}$ in a mean-squared sense is then the vector that maximizes the term $\vec{v}^H\mathbf{X}\mathbf{X}^H\vec{v}$, which is the sum of the magnitude-squared correlations between $\vec{v}$ and the channel signals. The maximum value of $\vec{v}^H\mathbf{X}\mathbf{X}^H\vec{v}$ for a unit-vector $\vec{v}$ is attained when $\vec{v}$ is the eigenvector of $\mathbf{X}\mathbf{X}^H$ with the largest eigenvalue; this is a basic result from linear algebra and is straightforward to prove, *e.g.* using Lagrange multipliers. The eigenvector with the largest eigenvalue is the *principal component* in the PCA representation of a data matrix; thus, maximizing the energy of the primary component leads to a PCA-based primary-ambient decomposition.

A brute-force computation of the PCA primary-ambient decomposition consists of first forming the covariance matrix $\mathbf{R} = \mathbf{X}\mathbf{X}^H$ and computing its eigenvalues $\{\lambda_i\}$ and eigenvectors $\{\vec{v}_i\}$. The largest eigenvalue λ_p and its corresponding eigenvector $\vec{v}_p$ are then determined. The primary component is then computed for each channel as the projection of $\vec{X}_m$ onto $\vec{v}_p$, and the ambience is computed as the projection residual. Of course, this approach to multichannel PCA primary-ambient separation is computationally costly since it involves an explicit eigendecomposition. It can be implemented more efficiently by taking into account that only the eigenvector of $\mathbf{R}$ with the largest eigenvalue is needed for the signal model of Eq. (6). This eigenvector can be derived, or at least approximated, by starting with any $\vec{v}_0$ and iterating the following steps [6]:

$$\vec{v}_0 \;\longleftarrow\; \mathbf{R}\vec{v}_0 \tag{10}$$

$$\vec{v}_0 \;\longleftarrow\; \frac{\vec{v}_0}{\|\vec{v}_0\|} \tag{11}$$

The vector $\vec{v}_0$ will converge to the eigenvector with the largest eigenvalue, with a faster convergence if the eigenvalue spread is large. A practical starting value for $\vec{v}_0$ is the $\vec{X}_m$ with the largest norm, since that will dominate the principal component computation. With this approach, the primary-ambient decomposition can be computed at a more reasonable cost than via a full eigendecomposition.

2.2. Decomposition of stereo signals

The application of PCA to the analysis of multichannel audio for coding or enhancement has not been widely considered in the literature; an approach for scalable representation is proposed in [7], but the PCA is applied in the frequency dimension rather than on the subband time-domain signals as in the method of Section 2.1. For the stereo case, which is important for coding but also essential for enhanced rendering or upmix, PCA decompositions have been applied for spatial analysis-synthesis of time-domain signals [8] and QMF subband signals [9]. In these approaches, the decompositions are derived via gradient-descent iteration; in the following, we simplify the multichannel decomposition for the two-channel case and provide a closed-form solution.

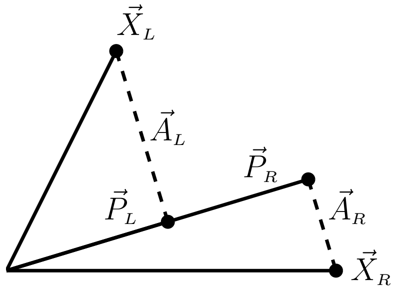

Fig. 2. Orthogonal primary-ambient decomposition of a stereo signal $\{\vec{X}_L, \vec{X}_R\}$ using PCA.

At time l, the first step in the PCA primary-ambient decomposition of stereo signals is to compute the cross-correlation $r_{LR}[k,l]$ and the auto-correlations $r_{LL}[k,l]$ and $r_{RR}[k,l]$ for each k, where

$$r_{ij}[k,l] \;=\; \vec{X}_i^H[k,l]\vec{X}_j[k,l]. \tag{12}$$

Note that the input channels are designated with L and R subscripts (for "left" and "right") instead of numerals. The largest eigenvalue of the matrix $\mathbf{X}\mathbf{X}^H$ is then computed according to

$$\lambda = \frac{1}{2}\left[r_{LL} + r_{RR} \;+\; \left((r_{LL} - r_{RR})^2 \;+\; 4|r_{LR}|^2\right)^{\frac{1}{2}}\right]. \tag{13}$$

This eigenvalue is then used to determine the principal vector

$$\vec{v} \;=\; r_{LR}\vec{X}_L \;+\; (\lambda - r_{LL})\vec{X}_R, \tag{14}$$

which here is not scaled to unit-norm. To enable scaling and an ensuing orthogonal projection, the magnitude of $\vec{v}$ and its correlations with the channel signals are computed:

$$r_{vv} \;=\; \vec{v}^H\vec{v} \tag{15}$$
$$r_{vm} \;=\; \vec{v}^H\vec{X}_m. \tag{16}$$

The primary component is then estimated for each channel as the projection of the channel signal onto the principal vector $\vec{v}$:

$$\vec{P}_m = \left(\frac{r_{vm}}{r_{vv}}\right)\vec{v}, \tag{17}$$

where the division is protected against singularities by thresholding; if $r_{vv}[k,l] < \epsilon$, meaning that no reliable principal component has been found, we set $\vec{P}_m[k,l] = 0$. The ambience is then computed as the projection residual as in Eq. (8): $\vec{A}_m = \vec{X}_m - \vec{P}_m$. A depiction of this PCA-based decomposition is given in Fig. 2.

2.3. Computation and performance considerations

In the stereo case, the correlations needed for the PCA decomposition can be approximated with a recursive formulation:

$$r_{xy}[t] \;=\; \sum_i x[t-i]^* y[t-i] \tag{18}$$
$$\hat{r}_{xy}[t] \;=\; (1-\mu)x[0]^* y[0] \;+\; \mu\,\hat{r}_{xy}[t-1] \tag{19}$$

where these are lag-0 cross-correlations at time t. With this approach to approximating the time-varying inner product between two signal vectors, is not necessary to fix the time duration for the computation, nor is the associated memory required. Instead, the effective time length of the inner product is determined by the forgetting factor μ, which can be adaptively adjusted as needed; and, the computation can be done in place without using buffers to retain the signal history. In the multichannel case, a similarly efficient scheme would be desirable to avoid the computational cost of the iterative principal vector determination; approaches to running PCA decomposition may be of interest in this regard [10].

Beyond the computational concerns, it is also important to note that the performance of the PCA primary-ambient decomposition is limited by how well the input signal satisfies the model assumptions given in Section 2.1. For instance, if the primary component does not have substantially more energy than the ambient component, the ambience will be present in the principal PCA component. This is not usually problematic, but it can lead to suboptimal rendering for purely diffuse sounds as well as incorrect or unstable directional analysis of the true primary content if the ambience is not uniformly spatially distributed. While in most cases the assumptions are largely valid and the PCA method is not compromised, it is of interest to improve the handling of atypical cases beyond the *ad hoc* methods currently employed to detect and treat these cases, which are based on experimentally tuned correlation threshold tests.

3. SPATIAL ANALYSIS-SYNTHESIS

Spatial analysis-synthesis is the process of extracting spatial information from an audio scene and using that information to drive a rendering algorithm. The following sections describe an analysis-synthesis scheme based on a vector theory of sound localization applied on a per-tile basis to the time-frequency signal representation.

3.1. Analysis

As shown in Fig. 1, the primary and ambient components are separately analyzed for spatial information after the primary-ambient decomposition is carried out. In the spatial analysis-synthesis framework, each time-frequency tile is treated as a distinct sound event. The analysis determines a perceived location for each time-frequency event in the audio scene; for each time and frequency, the analysis derives coordinates $(r[k,l], \theta[k,l])$, or equivalently a time-frequency direction vector $\vec{d}[k,l]$, which describes where the time-frequency sound event is located within a circle of unit radius centered at the listener. This derivation is carried out by a weighted sum of format vectors corresponding to the input channel positions, *e.g.* unit vectors in the directions $\{-30°, 30°, 0°, -110°, 110°\}$ for a standard five-channel format. This vector sum is based on the theory proposed in [11]; for a sound event broadcast from directions $\vec{q}_m$ (corresponding to the channel angles), the perceived direction is given by the vector

$$\vec{g}[k,l] \;=\; \sum_m \alpha_m[k,l]\vec{q}_m, \tag{20}$$

where in our framework the weights α_m are given by

$$\alpha_m[k,l] \;=\; \frac{|x_m[k,l]|^2}{\sum_i |x_i[k,l]|^2}. \tag{21}$$

Since the weights are normalized such that $\sum_m \alpha_m = 1$, this vector $\vec{g}[k,l]$ is constrained in magnitude by an inscribed polygon as shown in Fig. 3. A linear algebraic method to expand the *encoding locus* to the full listening circle is proposed in [3] and derived rigorously in [4]; it yields a robust direction vector

$$\vec{d}[k,l] \;=\; r[k,l]\frac{\vec{g}[k,l]}{||\vec{g}[k,l]||} \tag{22}$$

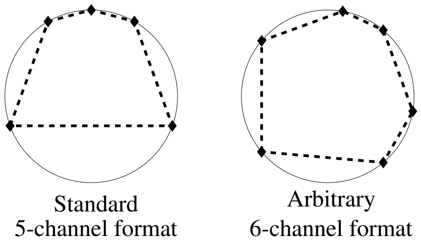

Fig. 3. Depiction of input channel formats (diamonds) and the corresponding encoding loci (dotted) of the Gerzon vector of Eq. (20).

which is in the angular direction of $\vec{g}$ but has a modified radius

$$r = \left\| \left[\vec{q}_i \, \vec{q}_j\right]^{-1} \vec{g} \right\|_1 \tag{23}$$

where $\vec{q}_i$ and $\vec{q}_j$ are the channel format vectors which bracket $\vec{g}$; the radius is thus the sum of the coefficients of the representation of $\vec{g}$ in the basis defined by $\vec{q}_i$ and $\vec{q}_j$. This radius ranges from zero for non-directional events to unity for discrete events that are purely pairwise-panned between adjacent channels [3, 4]. It should be noted that this expanded approach is not necessarily required for the ambience spatial analysis since the ambience is typically not highly directional and is indeed well represented by the vector $\vec{g}$.

3.2. Synthesis

Consistent synthesis of the input audio scene is achieved by deriving multichannel panning coefficients based on the time-frequency direction vectors $\vec{d}[k,l]$ derived in the analysis – such that analysis of the synthesized scene would yield the same direction vectors. One approach for deriving such coefficients is to use radial panning between pairwise weights and non-directional weights:

$$\vec{\beta}[k,l] = r[k,l]\vec{\sigma}[k,l] + (1 - r[k,l])\vec{\epsilon}[k,l] \tag{24}$$

where $\vec{\sigma}$ contains non-zero coefficients only for the two synthesis channels which bracket the direction vector $\vec{d}[k,l]$. The coefficient vector $\vec{\epsilon}$ is in the null space of the synthesis format matrix, *i.e.* the matrix whose columns are the unit format vectors in the directions of the output channels (loudspeaker positions). It is straightforward to show that this panning scheme leads to consistent synthesis [3]. Note that an optimization algorithm to derive non-directional panning weights for arbitrary loudspeaker configurations is given in [4].

In the spatial synthesis, the above panning scheme is applied independently to the primary and ambient components identified by the analysis. For the ambience synthesis, it is desirable to also include different allpass filters in each channel to increase the sense of spaciousness in the reproduction. Furthermore, for enhanced rendering where consistent synthesis is not a constraint, separate processing can be applied to the primary and ambient components to achieve a variety of effects. For instance, the ambient components extracted from a stereo signal can be distributed to surround channels for upmix [5], or the primary components can be spatially redistributed or otherwise modified prior to synthesis [3, 5].

4. SUMMARY AND FUTURE WORK

We have presented a general spatial analysis-synthesis method based on a PCA primary-ambient decomposition of multichannel audio input; the PCA decomposition algorithm is developed for the multichannel case and a closed-form solution is provided for the stereo case. We described localization analysis based on a vector theory; this is applied separately to the primary and ambient components to derive direction vectors that describe the spatial percept of each time-frequency component of the audio scene. These spatial cues are effective for low-rate spatial audio coding and for enhanced rendering. Future research includes improving the performance of the PCA for atypical signals, reducing the computational cost of the multichannel decomposition, and assessing the fidelity with respect to other analysis-synthesis schemes [12, 13]. A variety of enhancements based on this analysis-synthesis model are also under development, *e.g.* modification of the spatial cues for active upmix, and optimized independent rendering of the primary and ambient components for generalized upmix [14] and improved headphone listening.

5. REFERENCES

[1] F. Baumgarte and C. Faller, “Binaural cue coding – part I: psychoacoustic fundamentals and design principles,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 509–519, November 2003.

[2] J. Herre, C. Faller, et al., “Spatial audio coding: next-generation efficient and compatible coding of multi-channel audio,” *AES 117th Conv.*, October 2004, Preprint 6186.

[3] M. Goodwin and J.-M. Jot, “A frequency-domain framework for spatial audio coding based on universal spatial cues,” *AES 120th Conv.*, May 2006, Preprint 6751.

[4] M. Goodwin and J.-M. Jot, “Analysis and synthesis for universal spatial audio coding,” *AES 121st Conv.*, October 2006, Preprint 6874.

[5] C. Avendano and J. M. Jot, “A frequency-domain approach to multichannel upmix,” *Journal of the Audio Engineering Society*, vol. 52, no. 7/8, pp. 740–749, July/August 2004.

[6] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins Press, Baltimore, MD, 3rd edition, 1996.

[7] M. Briand, D. Virette, and N. Martin, “Parametric representation of multichannel audio based on principal component analysis,” *AES 120th Conv.*, May 2006, Preprint 6813.

[8] R. Irwan and R. Aarts, “Two-to-five channel sound processing,” *Journal of the Audio Engineering Society*, vol. 50, no. 11, pp. 914–927, November 2002.

[9] Y. Li and P. Driessen, “An unsupervised adaptive filtering approach of 2-to-5 channel upmix,” *AES 119th Conv.*, October 2005, Preprint 6611.

[10] D. Erdogmus, Y. N. Rao, et al., “Recursive principal components analysis using eigenvector matrix perturbation,” *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 13, pp. 2034–2041, 2004.

[11] M. A. Gerzon, “General metatheory of auditory localization,” *AES 92nd Conv.*, March 1992, Preprint 3306.

[12] M. Davis, C. Todd, and R. Dolby, “Method and apparatus for encoding and decoding audio information representing three-dimensional sound fields,” *U. S. Patent No.5,909,664*, June 1999.

[13] V. Pulkki and C. Faller, “Directional audio coding: filterbank and STFT-based design,” *AES 120th Conv.*, May 2006, Preprint 6658.

[14] M. Goodwin and J.-M. Jot, “Multichannel surround format conversion and generalized upmix,” *AES 30th Intl. Conf.*, March 2007.