Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

A frequency-domain framework for spatial audio coding based on universal spatial cues

Michael M. Goodwin[1] and Jean-Marc Jot[1]

[1] *Creative Advanced Technology Center, Scotts Valley, CA, USA*

Correspondence should be addressed to M. Goodwin and J.-M. Jot (mgoodwin,jmj@atc.creative.com)

ABSTRACT

Spatial audio coding (SAC) addresses the emerging need to efficiently represent high-fidelity multichannel audio. The SAC methods previously described involve analyzing the input audio for inter-channel relationships, encoding a downmix signal with these relationships as side information, and using the side data at the decoder for spatial rendering. These approaches are channel-centric in that they are generally designed to reproduce the input channel content over the same output channel configuration. In this paper, we propose a frequency-domain SAC framework based on the perceived spatial audio scene rather than on the channel content. We propose time-frequency spatial direction vectors as cues to describe the input audio scene, present an analysis method for robust estimation of these cues from arbitrary multichannel content, and discuss the use of the cues to achieve accurate spatial decoding and rendering for arbitrary output systems.

1. INTRODUCTION

Recently, spatial audio coding (SAC) has received increasing attention in the literature due to the proliferation of multichannel content and the need for effective bit-rate reduction schemes. The various methods proposed involve a number of common steps: analyzing the set of input audio channels for spatial relationships; downmixing the input audio, perhaps based on the spatial analysis; coding the downmix, typically with a legacy method for the sake of backwards compatibility; incorporating spatial side information in the coded representation; and, using the side information for spatial rendering at the decoder, if it supports such processing [1, 2, 3, 4]. Figure 1 depicts a generic SAC system with these components. In a typical system, the spatial side information is packed with the coded downmix for transmission or storage in the channel; in the figure, the side information is shown explicitly in the bottom branch of the block diagram to highlight the cue coding and decoding blocks since these are of particular interest here. The synthesis block is responsible for the spatial rendering process described above, namely using the spatial cues

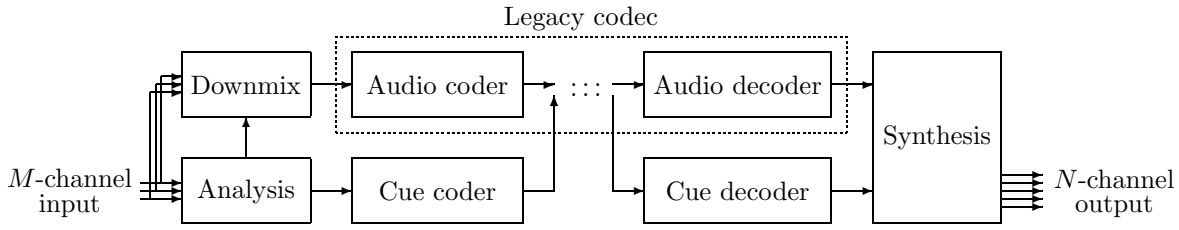


Fig. 1: Block diagram of a generic spatial audio coding and decoding system. Spatial cues are estimated from the M-channel input by the analysis and compressed by the cue coder before being embedded as side information in a legacy coding format. The unpacked and decoded cues are used by the synthesis to recreate the input scene using the N-channel output system.

to distribute the downmixed content to the output channels. To handle cases where the ancillary spatial information is not provided to the decoder, some SAC systems include a so-called *blind decoding* mode which is analogous to an upmix [5, 6]. The synthesis block thus may involve some flexibility based on the availability of the spatial cues and the output format.

Spatial audio coding methods previously described in the literature are channel-centric in that the spatial side information consists of inter-channel signal relationships such as level and time differences, *e.g.* as in binaural cue coding (BCC) [7, 8]. Furthermore, the codecs are designed primarily to reproduce the input audio channel content using the same output channel configuration; some extensions to playback on different loudspeaker formats have been discussed in the literature, but these still rely on prior knowledge of the formats as well as symmetry assumptions [8, 9]. If the actual output configuration does not match that assumed in the analysis, the spatial rendering may be inaccurate, often to an extent that cannot be optimally compensated by decode-side processing and calibration. To avoid such mismatches and enable robust rendering on arbitrary output systems, the SAC framework proposed in this paper uses spatial cues which describe the perceived audio scene rather than the relationships between the input audio channels. Variations of BCC which consider the physical scene parameters have been discussed in the literature; for example, an angle cue is used in [10, 11] as a quantization-robust alternative to level differences; this is essentially still a channel-based parameterization of the spatial information since the cue describes a panning angle between two fixed channel positions.

As stated above, the topic of this paper is spatial audio coding based on cues which describe the actual audio scene rather than specific inter-channel relationships. Essentially, we are proposing a frequency-domain SAC framework based on channel- and format-independent positional cues. The key design goal is to achieve a generic spatial representation that is independent of the number of input or output channels or the speaker layout.

The proposed spatial audio coding system operates as follows. The input is a set of audio signals and corresponding contextual spatial information. The input could be a multichannel mix obtained with various mixing or spatialization techniques such as conventional amplitude panning or ambisonics; or, it could be unmixed source content. For the former, the contextual information comprises the multichannel format specification, namely standardized speaker locations or channel definitions; for the latter, it consists of arbitrary positions based on sound design or some interactive control. In the analysis, the input signals are transformed into a frequency-domain representation wherein spatial cues are derived for each time-frequency tile based on the signal relationships and the original spatial context. When a given tile corresponds to a single source, the spatial information of that source is preserved by the analysis; when the tile corresponds to a mixture of sources, an appropriate combined spatial cue is derived. These cues are coded as side information with a downmix of the input audio signals. At the decoder, the cues are used to spatially distribute the

downmix so as to accurately recreate the input audio scene. If the cues are not provided, a consistent blind upmix can be derived and rendered.

This paper discusses all of the various components of the proposed SAC framework; the organization is as follows. Section 2 discusses the fundamental design goals and constraints of a "universal" spatial audio coding system. Section 3 proposes a baseline set of universal spatial cues and several extensions. An analysis method to derive the proposed cues is discussed in Section 4. Section 5 considers the downmix component of the SAC system, and how it can be driven by the spatial cue analysis. The spatial synthesis process is treated in Section 6. Applications are discussed in Section 7, and various demonstrations enabled by our SAC prototype are described in Section 8. Conclusions are given in Section 9.

Referring to the title of this paper, note that the term *frequency-domain* is used as a general descriptor of the SAC framework. We focus on the use of the short-time Fourier transform (STFT) for signal decomposition in the spatial analysis, but the approach to be discussed in this paper is applicable to other time-frequency transformations, filter banks, signal models, *etc.* Throughout the paper, we use the term *tile* to indicate a localized time-frequency component, and the term *bin* to describe a frequency channel or subband.

2. DESIGN GOALS AND CONSTRAINTS

In this paper, we are concerned with the general case of analyzing M-channel input, coding it as a downmix with spatial side information, and rendering the decoded audio on an arbitrary N-channel reproduction system. This generality gives rise to a number of design goals and constraints for the system components; these are discussed in the following sections.

2.1. System-level considerations

As discussed in Section 1, the primary design goal of the proposed SAC framework is that the spatial side information provide a physically meaningful description of the perceived audio scene. In the following list of system-level goals, this is restated in terms of channel independence in the first item:

- The spatial information should be independent of the input and output channel configurations.
- The encode-decode should preserve the spatial cues of both point sources and distributed sources, *e.g.* ambience components.
- A stable source should remain stable in the encode-decode process.

The second item relates to prior upmix algorithms based on direct-ambient signal separation [6], and the third item is just a basic requirement for system robustness.

2.2. Spatial cues

For a channel-independent or so-called *universal* spatial audio coding system to be effective, the cues must satisfy a number of constraints:

- Universality: the cues must describe the audio scene, *i.e.* the location and spatial characteristics of sound events, rather than channel relationships.
- Completeness: the cues must capture all of the salient features of the audio scene; the spatial percept of any potential sound event must be representable by the cues.
- Sparsity / compactness: to be most useful, the spatial cues should admit extensive compression so as to minimize the overhead of including the side information in the coded audio stream.
- Consistency: analyzing the output scene should yield the same cues as the input scene (with some limitations).

The universality, completeness, and sparsity constraints are essential. Consistency is basically a validation metric, but it becomes increasingly important in tandem coding scenarios; it is obviously desirable to preserve the spatial cues when the signal undergoes multiple generations of spatial encoding and decoding.

2.3. Downmix

The literature on spatial audio coding systems has covered the use of both mono and stereo downmixes for capturing the audio source content. Recently,

stereo downmix has become prevalent so as to preserve compatibility with standard stereo playback systems [1, 3, 4, 9], but we consider the design issues for both cases here. In general, the design goals for the downmix are as follows:

- The direct playback of the downmix must be of acceptable quality. Prior to encoding, the quality of a stereo downmix should be comparable to an original stereo recording.
- The signal energy and the balance between sources should be preserved in the downmix.
- The spatial information should be preserved in the downmix (to the extent possible).

For the mono case, these reduce to the first two: an acceptable quality for the mono signal and a basic preservation of the signal energy and balance. The key distinction is that spatial cues can be preserved to some extent in a stereo downmix; a mono downmix must rely on spatial side information to render any spatial cues.

2.4. Analysis

The analysis stage is probably the least constrained component of a spatial audio coding system. The short list of considerations includes:

- The analysis approach should be extensible to any number of input channels and to arbitrary channel layouts.
- The analysis approach should be amenable to real-time implementation for a reasonable number of input channels; this can be relaxed for off-line (non-streaming) applications.

Another issue is whether the transformation or model used by the analysis achieves separation of independent sources in the signal representation. Some blind source separation algorithms rely on minimal overlap in the time-frequency representation to extract distinct sources from a multichannel mix [12]. Here, though, complete source separation in the analysis representation is not essential, though it might be of interest for compacting the spatial cue data. Overlapping sources simply yield a composite spatial cue in the overlap region; the scene analysis of the human auditory system is then responsible for interpreting the composite cues and constructing a consistent internal understanding of the scene [13]. Our experiments with simultaneous (and thus substantially overlapping) talkers validate this assumption; further consideration of the effects of source overlap in the time-frequency representation is beyond the scope of this paper.

2.5. Synthesis

The synthesis block of a universal spatial audio coding system is responsible for using the spatial side information to process and redistribute the downmix signal so as to recreate the input audio scene using the output rendering format. Several constraints are thus inherent to the synthesis design:

- The rendered output scene should be a close perceptual match to the input scene. In some cases, *e.g.* when the input and output formats are identical, exact signal-level equivalence should be achieved for some test signals.
- Spatial analysis of the rendered scene should yield the same spatial cues used to generate it; this corresponds to the consistency constraint discussed in Section 2.2.
- The synthesis algorithm should not introduce any objectionable artifacts.
- The synthesis algorithm should be extensible to any number of output channels and to arbitrary output formats.
- The algorithm must admit real-time implementation on a low-cost platform (for a reasonable number of channels).
- The synthesis must have knowledge of the output rendering format, either via automatic measurement or user input.

Note that the last item is not limiting with respect to the system's universality since the output format knowledge is only used in the synthesis stage and is not incorporated in the analysis of the input audio.

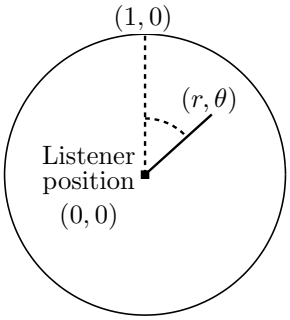


Fig. 2: Depiction of listening scenario upon which the universal spatial cues are based. The coordinates (r, θ) describe the position of a sound event.

3. UNIVERSAL SPATIAL CUES

The goals and constraints established in Section 2 provide a framework for specifying the key aspect of the proposed spatial audio coding system, namely the set of spatial cues. In this section, we define several variants of spatial side information which satisfy the design constraints. These are all based on the simple listening scenario depicted in Figure 2. In this general framework, the listener is situated at the center of a unit circle; the spatial aspects of perceived sound events are described with respect to this circle using the polar coordinates (r, θ), where $0 \leq r \leq 1$ and $-\pi \leq \theta \leq \pi$. The case $r = 1$, *i.e.* on the circle, corresponds to a discrete point source at angle θ. Decreasing r corresponds to source positions inside the circle as in a fly-over sound event. The limit $r = 0$ defines a non-directional percept; note that at $r = 0$ the angle cue θ is not meaningful.

In some spatial audio algorithms, in-the-circle positions ($r < 1$) are treated as a projection of a point on a sphere; the coordinates (r, θ) are representative of azimuth and elevation angles for an on-the-sphere point source [14]. In that case, $r = 0$ corresponds to a point source at the top. Clearly, three-dimensional treatment of sources within the sphere would require a third parameter. This extension is straightforward, but we adhere to the two-dimensional case since it is consistent with standard channel formats and since it allows for clear illustrations. The elevation issue will be discussed further with respect to synthesis in Section 6.

The coordinates (r, θ) define a direction vector. Some systems use such vectors in the time domain to determine a time-varying dominant direction; such approaches tend to have difficulties representing multiple discrete moving sources. To enable robust treatment of multiple moving sources and complex audio scenes, we propose to use the (r, θ) cues on a per-tile basis in a time-frequency domain; we can thus express the cues as $(r[k, l], \theta[k, l])$ where k is a frequency index and l is a time index.

Several design criteria for spatial cues were described in Section 2.2: universality, completeness, sparsity, and consistency. The proposed (r, θ) cues satisfy the universality constraint in that the spatial behavior of sound events is captured without reference to the channel configuration; we anticipate future cue extensions to differentiate between coherent and incoherent sources as in some BCC-based schemes [2, 3, 8, 15]. Completeness is achieved for the two-dimensional listening scenario if the cues can take on any coordinates within or on the unit circle; this issue will come up in defining a robust analysis algorithm in Section 4. With respect to sparsity, a scene with few discrete non-overlapping sources yields correspondingly few dominant angles; in the limiting case where there is one discrete point source in the audio scene, $r = 1 \ \forall \ k$ and θ is likewise a constant. Time-frequency overlap of multiple sources and source widening tends to reduce the apparent cue compactness, but the psychoacoustics of spatial hearing enables significant cue compression based on the resolution limits of the auditory system [7]. Finally, with respect to consistency, if accurate rendering is achievable based on these cues, it is reasonable to expect that similar cues would be estimated from the rendered scene. Ultimately, consistency is a high-level system issue which relates not only to the cue definitions but also to the robustness of the analysis and synthesis algorithms; we discuss synthesis consistency in the context of other SAC system components in Sections 5 and 6.

For the frequency-domain spatial audio coding framework, several variations of the direction vector cues merit consideration:

- (Unimodal) One direction vector per time-frequency tile
- (Bimodal direct-ambient) For each time-frequency tile, the signal is decomposed into di-

rect and ambient components, each of which is assigned a distinct direction vector

- (Continuous) One direction vector for each time-frequency tile with a focus parameter to describe source distribution and/or coherence
- (Multimodal) An extension of the continuous case (and the direct-ambient case) wherein multiple sources with distinct direction vectors and focus parameters are allowed for each time-frequency tile

While separation of direct and ambient components as well as representation of source coherence are relevant to high-fidelity spatial audio coding, we defer detailed treatment to future publications. Listening experiments have confirmed that unimodal cues provide a solid basis for a spatial audio coding system; in the remainder, we thus focus on the case where the spatial cues for each tile consist of one $(r[k,l], \theta[k,l])$ direction vector.

4. SPATIAL ANALYSIS

The universality of the spatial cues proposed in the previous section is the key benefit of the SAC system proposed in this paper. Of course, the critical question is whether such cues can actually be estimated robustly from multichannel audio content. We address this issue here for the proposed cues.

The direction-vector spatial cues are based on a vector theory of auditory localization proposed in [16]; the direction-vector concept is incorporated in ambisonics systems and various panning schemes [17, 18, 19, 20]. The idea is simply that the contribution of each channel to the audio scene can be represented by an appropriately scaled direction vector, and the perceived source location is then given by a vector sum of the scaled channel vectors. A depiction of this vector sum is given in Figure 3 for a standard five-channel configuration.

Before mathematically specifying the direction vector analysis, it should be noted that previous direction or *dominance* vector methods were based on the time-domain channel signals as in the matrix encoding-decoding literature [14]. Here, we consider direction vectors on a per-tile basis for an arbitrary

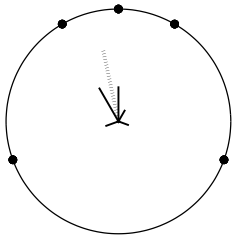


Fig. 3: Depiction of channel vector summation for a standard five-channel layout. The dotted line is the direction vector given by Eq. (1).

time-frequency representation; specifically, we use the STFT, but other representations or signal models are similarly viable. In this context, the input channel signals $x_m[t]$ are transformed into a representation $X_m[k,l]$ where k is a frequency or bin index and l is a time index as established in Section 3; m is the channel index. Note that the $x_m[t]$ are assumed to be speaker-feed signals.

Given the transformed signals, the directional analysis is carried out as follows. First, the channel configuration or source positions, *i.e.* the spatial context of the input audio channels, is described using unit vectors $\vec{p}_m$ pointing to each channel position; if θ is assumed to be 0 at the front center position (the top of the circle in Figure 2) and positive in the clockwise direction, the rectangular coordinates are simply $\vec{p}_m = [\sin\theta_m \;\; \cos\theta_m]^T$ where θ_m is the clockwise angle of the m-th input channel. Then, the direction vector sum is computed as

$$\vec{g}[k,l] \;=\; \sum_m \alpha_m \vec{p}_m. \qquad (1)$$

In the *velocity vector* summation [16, 20], the signal magnitudes are used to weigh the respective channel contributions:

$$\alpha_m \;=\; \frac{|X_m[k,l]|}{\sum_{i=1}^{M} |X_i[k,l]|}. \qquad (2)$$

In the alternate *energy* or *intensity* sum [20],

$$\alpha_m \;=\; \frac{|X_m[k,l]|^2}{\sum_{i=1}^{M} |X_i[k,l]|^2}. \qquad (3)$$

A crossover frequency of approximately $700Hz$ from the velocity model to the energy model is proposed

in [16], but informal listening tests suggest that either formulation is basically effective in the proposed SAC analysis scheme.

In the following, we refer to the direction vector established in Eqs. (1)-(3) simply as the *Gerzon vector* as is common in the spatial audio community. With respect to the (r, θ) spatial cues proposed in Section 3, the magnitude and angle of the Gerzon vector satisfy some of the cue criteria: θ indicates a dominant direction for a sound event, and r describes its radial location; $r = 0$ corresponds to a non-directional event (at the center of the circle), and $r = 1$ corresponds to a discrete directional event (on the circle). However, the Gerzon vector has a significant shortcoming in that its magnitude does not faithfully describe the radial location of discrete pairwise-panned sources. In the pairwise-panned case, the so-called encoding locus is bounded by the inter-channel chord as depicted in Figure 4(a), meaning that the radius is underestimated for pairwise-panned sources, except in the hard-panned case where the direction exactly matches one of the $\vec{p}_m$ vectors. Subsequent decoding based on the Gerzon vector magnitude will thus not render such sources accurately.

To correct the representation of pairwise-panned sources, the Gerzon vector can be rescaled so that it has unit magnitude:

$$\vec{d} \;=\; \frac{\vec{g}}{\|\vec{g}\|}. \tag{4}$$

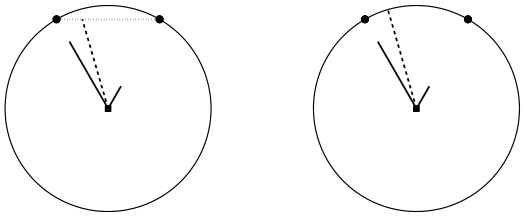
It is straightforward to derive a closed-form expression for this rescaling:

$$\vec{d} \;=\; \Gamma(\alpha_i, \alpha_j, \theta_j - \theta_i)\; \vec{g} \tag{5}$$

$$\Gamma(a_i, a_j, \theta) \;=\; \frac{a_i + a_j}{\left[a_i^2 + a_j^2 + 2a_i a_j \cos\theta\right]^{\frac{1}{2}}} \tag{6}$$

$$\;=\; \|\vec{g}\|^{-1}. \tag{7}$$

In Eq. (5), α_i and α_j are the weights for the channel pair in the vector summation of Eq. (1); θ_i and θ_j are the corresponding channel angles. As illustrated in Figure 4(b), this correction rescales the direction vector to achieve unit magnitude for discrete pairwise-panned sources. Related modifications of the dominance vector for matrix encoding are suggested in [14].



(a) Gerzon vector (b) Modified vector

Fig. 4: Direction vectors for pairwise-panned sources. The Gerzon vector specified in Eqs. (1)-(3) is limited in magnitude by the dotted chord shown in (a). Diagram (b) shows the modification of Eq. (4) rescaling the vector to unit magnitude ($r = 1$) for pairwise-panned sources.

The rescaling modification of Eq. (4) corrects the direction vector magnitude for the case of pairwise panning in a two-channel encoding. In a multichannel scenario, the same modification of the Gerzon vector could be applied if only two adjacent channels were active. In the common case where all channels are active, however, the scaling is not directly applicable. One approach to consider is to first find the inter-channel arc which includes the direction vector angle and to then rescale the two corresponding channel weights, but this overestimates the magnitude and also changes the direction of the sum vector. A rescaling method which does not incur these problems is discussed in the following.

In the multichannel scenario, a robust Gerzon vector rescaling can be achieved by decomposing the vector into a directional component and a non-directional component. Consider again the unit channel vectors $\vec{p}_m$. The unmodified Gerzon vector $\vec{g}$ is simply a weighted sum of these vectors with $\sum_m \alpha_m = 1$ as specified in Eqs. (1)-(3). The vector sum can be equivalently expressed in matrix form as

$$\vec{g} \;=\; P\vec{\alpha} \tag{8}$$

where the m-th column of the matrix P is the channel vector $\vec{p}_m$. Note that P is of rank two for a planar channel format (if not all of the channel vectors are coincident or colinear) or of rank three for three-dimensional formats.

Since the format matrix P is rank-deficient (when the number of channels is sufficiently large as in typ-

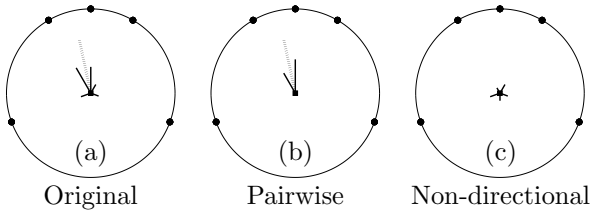


Fig. 5: Direction vector decomposition into a pairwise-panned and non-directional component. Diagram (a) shows the scaled channel vectors and Gerzon direction vector from Figure 3; (b) and (c) show the pairwise-panned and non-directional components, respectively, according to the decomposition specified in Eqs. (9) and (10).

ical multichannel scenarios), the direction vector $\vec{g}$ can be decomposed as

$$\vec{g} \;=\; P\vec{\alpha} \;=\; P\vec{\rho} \;+\; P\vec{\epsilon} \qquad (9)$$

where $\vec{\alpha} = \vec{\rho} + \vec{\epsilon}$ and where the vector $\vec{\epsilon}$ is in the null space of P, *i.e.* $P\vec{\epsilon} = 0$ with $\|\vec{\epsilon}\| > 0$. Of the infinite number of possibilities here, there is a uniquely specifiable decomposition of particular value for our application: if the coefficient vector $\vec{\rho}$ is chosen to only have nonzero elements for the channels which are adjacent to the vector $\vec{g}$, the resulting decomposition gives a pairwise-panned component with the same direction as $\vec{g}$ and a non-directional component whose Gerzon vector sum is zero. Denoting the channel vectors adjacent to $\vec{g}$ as $\vec{p}_i$ and $\vec{p}_j$, we can write:

$$\begin{bmatrix} \rho_i \\ \rho_j \end{bmatrix} = \begin{bmatrix} \vec{p}_i \; \vec{p}_j \end{bmatrix}^{-1} \vec{g} \qquad (10)$$

where ρ_i and ρ_j are the nonzero coefficients in $\vec{\rho}$, which correspond to the i-th and j-th channels. Here, we are finding the unique expansion of $\vec{g}$ in the basis defined by the adjacent channel vectors; the remainder $\vec{\epsilon} = \vec{\alpha} - \vec{\rho}$ is in the null space of P by construction. An example of the decomposition is shown in Figure 5.

Given the decomposition into pairwise and non-directional components, the norm of the pairwise coefficient vector $\vec{\rho}$ can be used to provide a robust rescaling of the Gerzon vector:

$$\vec{d} \;=\; \|\vec{\rho}\|_1 \left(\frac{\vec{g}}{\|\vec{g}\|} \right) \qquad (11)$$

In this formulation, the magnitude of $\vec{\rho}$ indicates the radial sound position. The boundary conditions meet the desired behavior: when $\|\vec{\rho}\|_1 = 0$, the sound event is non-directional and the direction vector $\vec{d}$ has zero magnitude; when $\|\vec{\rho}\|_1 = 1$, the direction vector $\vec{d}$ has unit magnitude. Note that we are assuming that the weights in $\vec{\rho}$ are energy weights, such that $\|\vec{\rho}\|_1 = 1$ for a discrete pairwise-panned source; this assumption is consistent with standard panning methods.

The angle and magnitude of the rescaled vector in Eq. (11) are computed for each time-frequency tile in the signal representation; these are used as the $(r[k,l], \theta[k,l])$ spatial cues in the proposed SAC system. The following section discusses the downmix component, while Section 6 describes the use of the cues to recreate the input audio scene from the downmix signal. There, we demonstrate that the rescaled Gerzon vector provides a consistent synthesis; that is, the synthesized scene yields the same spatial cues.

5. DOWNMIX

Various downmix schemes for spatial audio coding have been proposed in the literature; early systems were based on a mono downmix, and later extensions incorporated stereo downmix for compatible playback on legacy stereo reproduction systems [2]. Some recent methods allow for a custom downmix to be provided in conjunction with the multichannel input; the spatial side information then serves as a map from the custom downmix to the multichannel signal [3]. In this section, we consider three downmix options for the proposed SAC system: mono, stereo, and *guided* stereo.

The proposed spatial audio coder can operate effectively with a mono downmix signal generated as a direct sum of the input channels. To counteract the possibility of frequency-dependent signal cancellation (or amplification) in the downmix, dynamic equalization can be applied as in [8, 9, 21]. Such equalization serves to preserve the signal energy and balance in the downmix.

Though robust SAC performance is achievable with a monophonic downmix, the applications are somewhat limited in that the downmix is not optimal for

playback on stereo systems. To enable compatibility of spatially encoded material with stereo playback systems not equipped to decode and process the spatial cues, a stereo downmix is called for [1, 2]. In some cases, this downmix is generated by left- and right-side sums of the input channels, perhaps with equalization similar to that described above. In our system, however, such a direct stereo downmix is somewhat problematic since it relies on assumptions about the input channel format, but this can be managed by analyzing the input configuration for left-side and right-side contributions.

While an acceptable direct downmix can be derived, it does not specifically satisfy the design goal of preserving spatial cues in the stereo downmix; directional cues may be compromised due to the input channel format or the mixing operation. An alternate approach which preserves the cues, at least to the extent possible in a two-channel signal, is to use the spatial cues extracted from the multichannel analysis to synthesize the downmix. The frontal cues are maintained in this guided downmix, and other directional cues are folded into the frontal scene. Future extensions may include a phase-amplitude downmix wherein non-frontal components are phase-shifted as in matrix encoders; robust embedding of positional information in the downmix would enable improved blind decode and a reduction in the amount of side information needed for the SAC codec.

6. SYNTHESIS

The synthesis engine of a spatial audio coding system applies the spatial side information to the downmix signal to generate a set of reproduction signals. This spatial decoding process amounts to synthesis of a multichannel signal from the downmix; in this regard, it is often referred to as a *guided upmix*. In this section, we describe the spatial decode of a mono downmix based on the proposed universal spatial cues; extensions to the stereo downmix case are straightforward.

Given the downmix signal $T[k,l]$ and the cues $r[k,l]$ and $\theta[k,l]$, the goal of the spatial synthesis is to derive output signals $Y_n[k,l]$ for N speakers positioned at angles θ_n so as to recreate the input audio scene represented by the downmix and the cues. These output signals are generated on a per-tile basis using the following procedure. First, the output channels adjacent to $\theta[k,l]$ are identified. The corresponding channel vectors $\vec{q}_i$ and $\vec{q}_j$ are then used in a vector-based panning method to derive pairwise panning coefficients [19, 20]; this panning is similar to the process described in Eq. (10). Here, though, the resulting panning vector σ is scaled such that $\|\vec{\sigma}\|_1 = 1$. These pairwise panning coefficients capture the angle cue $\theta[k,l]$; they represent an on-the-circle point, and using these coefficients directly to generate a pair of synthesis signals would render a point source at $\theta[k,l]$ and $r = 1$.

To correctly render the radial position of the source as represented by the magnitude cue $r[k,l]$, a second panning is carried out between the pairwise vector $\vec{\sigma}$ and a non-directional set of panning coefficients, *i.e.* a set of weights which render a non-directional sound event over the given output configuration. Denoting the non-directional set by $\vec{\delta}$, the weights resulting from a linear pan are given by

$$\vec{\beta} \;=\; r\vec{\sigma} \;+\; (1-r)\vec{\delta}. \qquad (12)$$

Once the panning vector $\vec{\beta}$ is computed, the synthesis signals can be generated by amplitude-scaling and distributing the mono downmix accordingly:

$$Y_n[k,l] \;=\; \sqrt{\beta_n}\; T[k,l]. \qquad (13)$$

The consistency of the synthesized scene can be verified by considering a directional analysis based on the output format matrix, denoted in the following by Q. The Gerzon vector is given by

$$\vec{g}_s \;=\; Q\vec{\beta} \;=\; rQ\vec{\sigma} \;+\; (1-r)Q\vec{\delta}. \qquad (14)$$

This corresponds to the analysis decomposition in Eq. (9); by construction, $rQ\vec{\sigma}$ is the pairwise component and $(1-r)Q\vec{\delta}$ is the non-directional component. Since $Q\vec{\delta} = 0$, we have

$$\vec{g}_s \;=\; rQ\vec{\sigma}. \qquad (15)$$

We see here that $r\vec{\sigma}$ corresponds to the $\vec{\rho}$ pairwise vector in the analysis decomposition. Rescaling the Gerzon vector according to Eq. (11), we have

$$\vec{d}_s \;=\; \|r\vec{\sigma}\|_1 \left(\frac{\vec{g}_s}{\|\vec{g}_s\|}\right) \;=\; r\left(\frac{\vec{g}_s}{\|\vec{g}_s\|}\right). \qquad (16)$$

This direction vector has magnitude r, verifying that the synthesis method preserves the radial position cue; the angle cue is preserved by the pairwise-panning construction of $\vec{\sigma}$.

The flexible rendering approach described above yields a synthesized scene which is perceptually and mathematically consistent with the input audio scene; the universal spatial cues estimated from the synthesized scene indeed match those estimated from the input audio.

If source elevation angles are incorporated in the set of spatial cues, the rendering can be extended to include panning with respect to both the non-directional center point and a top-of-the-sphere point; such a top speaker, if not present in the reproduction systems, can be realized as a virtual speaker.

7. APPLICATIONS

In this section, we discuss several applications of the spatial coding system and the universal spatial cues.

Flexible multichannel rendering

In channel-centric spatial audio coding approaches, the configuration of output speakers is assumed at the encoder; spatial cues are derived for rendering the input content with the assumed output format. As a result, the spatial rendering may be inaccurate if the actual output format differs from the assumption. The issue of format mismatch is addressed in some commercial receiver systems which determine speaker locations in a calibration stage and then apply compensatory processing to improve the reproduction; a variety of methods have been described for such speaker location estimation and system calibration [22, 23]. The multichannel audio decoded from a channel-centric SAC representation could of course be processed in this way to compensate for output format mismatch. The proposed SAC system, however, can integrate the calibration information directly in the decoding stage and thereby eliminate the need for the compensation processing. Indeed, the problem of the output format is addressed directly by the proposed framework: given a source component (tile) and its spatial cue information, the spatial decoding can be carried out to yield a robust spatial image for the given output configuration, be it a multichannel speaker system, headphones with virtualization, *etc.* It should be noted that the spatial cues and flexible synthesis in the proposed SAC system are related to recent methods described for room response rendering [24].

Upmix

Given the growing adoption of multichannel listening systems in home entertainment setups, algorithms for enhanced rendering of stereo content over such systems is of great commercial interest. A number of methods for this *upmix* operation have been described in the literature [5, 6, 25]. The spatial decoding process in SAC systems is often referred to as a guided upmix since the side information is used to control the synthesis of the output channels; conversely, a non-guided upmix is tantamount to a *blind decode* of a stereo signal. It is straightforward to apply the universal spatial cues described in this paper for 2-to-N upmixing. Indeed, for the case $M = 2$ and $N > 2$, the M-to-N SAC system of Figure 1 is simply an upmix with an intermediate transmission channel. In such upmix schemes, the frontal imaging is preserved and indeed stabilized for rendering over standard multichannel speaker layouts. If front-back information is phase-encoded in the stereo, side and rear content can also be identified and robustly rendered using a matrix-decode methodology [14]. Furthermore, ambience extraction and redistribution can be incorporated for enhanced envelopment [6].

Directional source extraction and enhancement

The localization information provided by the universal spatial cues can be used to extract and enhance sources in multichannel mixes. A frequency-domain panning-based analysis was used in [26] to enhance, suppress, and re-pan discrete sources in a stereo signal. The time-frequency direction vector cues discussed in this paper provide a potential extension to enable source manipulation and modification in multichannel scenarios.

Transcoding of inter-channel spatial cues

To enable flexible output rendering, the channel-centric side information used in standard SAC system could be converted to universal spatial cues before synthesis. It will be a subject of future work to specify this conversion and to compare the spatial

fidelity of the universal cues extracted from an original source to that of cues derived by transcoding channel-centric data.

8. DEMONSTRATIONS

The real-time prototype implementation of the proposed SAC system enables several demonstrations which illustrate the effectiveness of the proposed universal spatial cues. These include:

- Spatial encode-decode for multichannel music content: this illustrates the basic robustness of the system.
- On-the-fly azimuth rotation of the output scene: to avoid the impracticality of physically moving the speakers, the entire *scene* can be moved to demonstrate flexible rendering on arbitrary speaker layouts.
- Speaker selection/de-selection: the rendered scene is preserved as speakers are removed from or added to the synthesis configuration.
- Input variability: given an input with a discrete music source in each channel, the rendered scene positions the sources accurately as the input channel angles are changed on the fly; this demonstrates robustness to the input configuration and to moving input sources.
- Input variability and time-frequency overlap: given an input with a discrete talker in each channel, the scene is robustly rendered as the input channel angles are adjusted on the fly; this verifies that time-frequency source overlap is not fundamentally problematic.

A number of these demonstrations were given at the convention, and some static sound examples are available online [27].

9. CONCLUSION

In this paper, we presented a system for spatial audio coding based on channel- and format-agnostic spatial cues, which we term *universal spatial cues*; these cues directly describe the properties of the input audio scene rather than the relationships between the input audio channels, thereby enabling robust representation of arbitrary input content and flexible rendering on arbitrary output systems. We discussed the fundamental design goals and constraints of the various components of universal spatial coding systems, and described each of the components of the proposed system in light of the established goals.

Various real-time demonstrations of the proposed spatial coder were given at the convention to illustrate its robustness and independence to the input and output formats; static sound examples are available online [27]. These serve as a proof-of-concept that universal spatial cues can be used in spatial audio coding systems. Various issues that merit further investigation have been mentioned throughout the text; such future work also includes optimization of cue compression and formal listening assessments.

10. REFERENCES

[1] J. Herre, C. Faller, et al. MP3 Surround: efficient and compatible coding of multi-channel audio. *116th Convention of the Audio Engineering Society*, May 2004. Preprint 6049.

[2] J. Herre, C. Faller, et al. Spatial audio coding: next-generation efficient and compatible coding of multi-channel audio. *117th Convention of the Audio Engineering Society*, October 2004. Preprint 6186.

[3] J. Breebaart, J. Herre, et al. MPEG spatial audio coding / MPEG surround: overview and current status. *119th Convention of the Audio Engineering Society*, October 2005. Preprint 6599.

[4] A. Seefeldt, M. Vinton, and C. Robinson. New techniques in spatial audio coding. *119th Convention of the Audio Engineering Society*, October 2005. Preprint 6587.

[5] R. Irwan and R. Aarts. Two-to-five channel sound processing. *Journal of the Audio Engineering Society*, 50(11):914–927, November 2002.

[6] C. Avendano and J. M. Jot. A frequency-domain approach to multichannel upmix. *Journal of the Audio Engineering Society*, 52(7/8):740–749, July/August 2004.

[7] F. Baumgarte and C. Faller. Binaural cue coding – part I: psychoacoustic fundamentals and design principles. *IEEE Transactions on Speech and Audio Processing*, 11(6):509–519, November 2003.

[8] C. Faller and F. Baumgarte. Binaural cue coding – part II: schemes and applications. *IEEE Transactions on Speech and Audio Processing*, 11(6):520–531, November 2003.

[9] C. Faller. Coding of spatial audio compatible with different playback formats. *117th Convention of the Audio Engineering Society*, October 2004. Preprint 6187.

[10] S. Chon et al. Virtual source location information for binaural cue coding. *119th Convention of the Audio Engineering Society*, October 2005. Preprint 6538.

[11] J. Seo, I. Jang, and K. Kang. Spatial audio coding system based on virtual source location information. *119th Convention of the Audio Engineering Society*, October 2005. Preprint 6576.

[12] S. Rickard and O. Yilmaz. On the approximate W-disjoint orthogonality of speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2002.

[13] A. Bregman. *Auditory Scene Analysis*. The MIT Press, Cambridge, MA, 1990.

[14] S. Julstrom. A high-performance surround sound process for home video. *Journal of the Audio Engineering Society*, 35(7/8):536–549, July/August 1987.

[15] C. Faller. Parametric multichannel audio coding: synthesis of coherence cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):299–310, January 2006.

[16] M. A. Gerzon. General metatheory of auditory localization. *92nd Convention of the Audio Engineering Society*, March 1992. Preprint 3306.

[17] M. A. Gerzon. Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871, November 1985.

[18] M. A. Gerzon. Panpot laws for multispeaker stereo. *92nd Convention of the Audio Engineering Society*, March 1992. Preprint 3309.

[19] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, June 1997.

[20] J.-M. Jot, V. Larcher, and J.-M. Pernaux. A comparative study of 3-D audio encoding and rendering techniques. *AES 16th International Conference on Spatial Sound Reproduction*, April 1999.

[21] F. Baumgarte, C. Faller, and P. Kroon. Audio coder enhancement using scalable binaural cue coding with equalized mixing. *116th Convention of the Audio Engineering Society*, May 2004. Preprint 6060.

[22] S. Wilson, J. Walters, and J. Abel. Speaker locations from inter-speaker range measurements: closed-form estimator and performance relative to the Cramer-Rao lower bound. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2004.

[23] R. Bruno, A. Laborie, and S. Montoya. Reproducing multichannel sound on any speaker layout. *118th Convention of the Audio Engineering Society*, May 2005. Preprint 6375.

[24] J. Merimaa and V. Pulkki. Spatial impulse response rendering I: analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127, December 2005.

[25] Y. Li and P. Driessen. An unsupervised adaptive filtering approach of 2-to-5 channel upmix. *119th Convention of the Audio Engineering Society*, October 2005. Preprint 6611.

[26] C. Avendano. Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression, and re-panning applications. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2003.

[27] M. Goodwin. Spatial audio coding demo, May 2006. URL: www.atc.creative.com/users/mgoodwin/demo.html.