Spatial Enhancement of Audio Recordings

Jean-Marc Jot and Carlos Avendano

Creative Advanced Technology Center
Scotts Valley, CA 95067, USA.

ABSTRACT

This paper reviews signal processing methods for improving the presentation of two-channel and multi-channel recordings by spatial processing in consideration of the playback system. The approaches discussed include: artificial reverberation; virtualization of recordings for playback over headphones or two loudspeakers; "upmix" or "stereo widening" of two-channel recordings for playback over multi-channel or two-channel loudspeaker systems. The goal is to enhance the listener's experience by attempting to match the audio signal to the playback system, while reproducing the intended audio image as faithfully as possible. Recently developed frequency-domain algorithms for ambience enhancement and directional enhancement are reviewed, and their possible benefits are discussed.

1. INTRODUCTION

Today and for the foreseeable future, multi-channel recorded audio content coexists with traditional two-channel formats. Consumers may indifferently want to play these recordings over multi-channel or two-channel loudspeaker systems, or over headphones. Assumptions made at the production stage are often not met at the playback end: the playback system may provide too few or too many channels, or the playback channels may not be employed as preferred by the producer.

Although any multi-channel DVD or DVD-A player supports basic downmix conventions to ensure that all the audio information in a multi-channel recording will actually be heard over a two-channel system, this and other format mismatch scenarios are opportunities for significantly improving the listener's experience.

In this paper, we are more particularly concerned with the following scenarios:

- ***Upmix*** of two-channel content played back over a multi-channel surround sound system (Fig. 1). The expectation is that the center channel can be exploited to make the frontal stereo image more robust against lateral displacements of the listener away from the reference listening position ("sweet spot"), and that the surround channels can be exploited to provide a more immersive listening experience.
- ***Stereo widening*** of two-channel recordings, when played back over two frontal loudspeakers forming an angle significantly narrower than 60º. This scenario may be encountered in computer audio systems or portable players and boomboxes.
- ***Virtual surround*** playback of five-channel content over two loudspeakers. This requires a downmix process where it is desirable to create the illusion that the surround channels are reproduced over loudspeakers located in the rear half-plane on the left and right sides of the listener's head.
- ***Binaural virtualization*** of two-channel or multi-channel content for playback over headphones, seeking to create the illusion that the recording is actually being heard over loudspeakers located at the due positions in the horizontal plane.

Ideally, the signal processing methods should involve no assumption on the type of audio material (music or movie soundtrack) or on the spatial encoding techniques used in the recording. Possible spatial encoding techniques include live microphone recording methods and multi-track mixing involving artificial reverberation or other studio effects. Two-channel material may be provided in a low-bit-rate audio format (e. g. MP3) or in a matrix-encoded surround sound format. Multi-channel material may also be in low-bit-rate audio format or may be the multi-channel output of a matrix surround decoder. In this paper, we restrict our discussion of multi-channel formats to the standard five-channel layout (Fig. 1), without significant loss of generality.

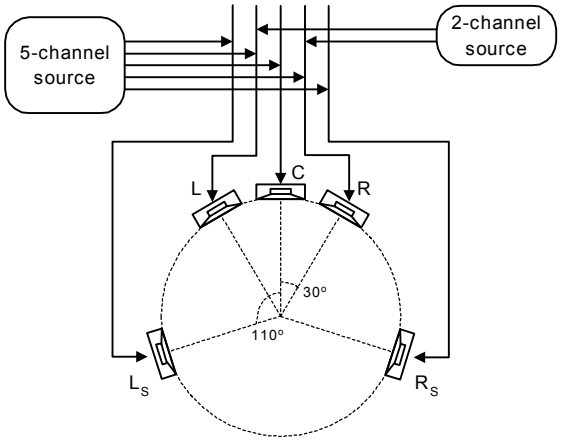

Fig. 1: 2-channel and 5-channel layout geometries.

The "enhancements" discussed here need not be obviously noticeable, and indeed may often be subtle. Ideally, an uninformed listener should not discern any alteration of the audio signal and the intention of the producer should be reproduced as faithfully as possible (including the timbre, localization and relative balance of the different sound events). However, if the enhancement is successful, switching from *on* to *off* should result in a noticeable degradation in the spatial qualities of the sound image.

In the next section of this paper, we provide an overview of audio signal processing techniques that can be used to achieve such enhancements. We then define general assumptions and design criteria for upmix processes, based on the decomposition of the recorded signal into *ambient* components and *primary* components. In sections 4 and 5, we discuss techniques for enhanced reproduction of the ambient components and for directional enhancement of the primary components. The existing time-domain approaches are reviewed, and a frequency-domain analysis/synthesis approach is presented.

2. SIGNAL PROCESSING APPROACHES

The "toolbox" of signal processing techniques that can be employed and combined together to achieve the enhancements described above includes well-known 3D audio spatialization and matrix decoding methods. In this section, we review the principles of these approaches, some of which will be examined in more detail in sections 4 and 5. We conclude this section by reviewing the rationale for a frequency-domain technique recently proposed by the authors.

2.1. Head-related virtualization

The purpose of head-related stereophony is to reconstruct the intended audio signals at the ears of the listener. The general principles have been extensively covered in the literature (e. g. [1]-[4]). In this section, we review the relevance and challenges of this approach in the scenarios under consideration.

For headphone playback, each source channel is passed through a pair of filters that model the head-related transfer functions (HRTF) for the desired perceived direction of a "virtual loudspeaker". The resulting binaural signal pairs are summed to feed the headphones. With this binaural synthesis technique, the illusion of listening over actual loudspeakers is challenged by two well-known phenomena [1]: in-head-localization of sound events and front-back ambiguities (front channels heard in the back or vice-versa). Possible approaches for reducing the occurrence of these phenomena include [1]: (a) head tracking; (b) simulating the reflections of a "virtual room"; (c) adapting the HRTF models to the listener.

A typical virtual surround sound process for two loudspeakers includes a binaural synthesis stage followed by a "transaural" cross-talk cancellation stage that compensates for the loudspeaker-to-ear acoustic transfer functions. If the process simulates front virtual loudspeakers spaced more widely than the physical loudspeakers, a stereo widening effect can be achieved. While simple "head shadow" filter models are sufficient to obtain a convincing simulation of virtual loudspeakers localized at the front and sides [2] [5], the production of rear images over a frontal pair of loudspeakers requires more accurate HRTF models. The main drawbacks of this reproduction technique are (a) the risk of a perceived coloration of the source signals and (b) the constraint for the listener to sit at the sweet spot.

2.2. Artificial reverberation

Digital reverberation and room simulation techniques can be employed to place the physical or virtual loudspeakers in a "virtual listening room" [2] [6] [7].

In the context of headphone virtualization, simulating the virtual loudspeakers within a virtual room provides a more natural, "exteriorized" listening experience (perhaps more convincing when the virtual room acoustics sound similar to the physical space where the listener is located). To achieve this effect, a subtle amount of artificial reverberation is sufficient. It has been reported that simulating a few selected early reflections can be more effective than a complete room simulation [4].

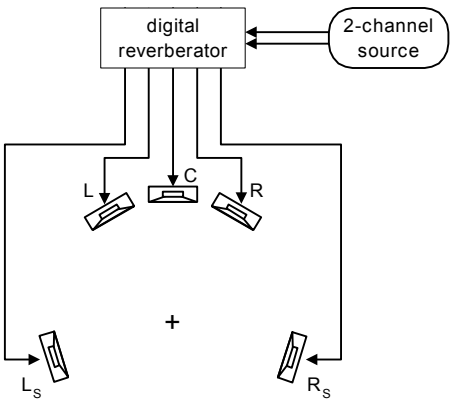

Fig. 2: 2-to-5 reverberation process.

The addition of multi-channel reverberation to a two-channel recording (Fig. 2) can be used in a multi-channel upmix system [5]. It may be argued that, in order to preserve the frontal stereo image, the artificial reverberation signal should only feed the surround channels. On a two-loudspeaker system, an effective stereo widening effect can be obtained by reproducing the artificial reverberation or reflections through virtual side loudspeakers, while leaving the "dry" signal path unprocessed [5].

In these scenarios, the artificial reverberation parameters (e. g. level, decay time) will usually not match the conditions in which the recording was produced. The stereo image may therefore be altered in a manner that is not consistent with the intention of the producer, particularly in terms of the distance of the primary sound events, the sense of environment, or the overall "color".

With the "virtual microphone" technique described in [8], reverberation responses corresponding to far-field microphone positions in the original recording venue can be derived and employed to compute surround-channel signals by convolution. However, the application of this method is in principle restricted to "live" recordings with detailed additional information on the original recording conditions and microphone techniques.

2.3. Fixed matrix

The basic "passive" 2-to-4 decoding matrix derives a center-channel signal C and a surround-channel signal S from the incoming L and R signals [9]:

$$C = 0.707\ (L + R),$$
$$S = 0.707\ (L - R). \tag{1}$$

The signal S is often called the "ambience" signal because subtracting R from L has the effect of eliminating or attenuating sounds panned at or near the center of the stereo image, while preserving the sounds that are panned to lateral positions or weakly correlated. Before feeding signal S to the surround loudspeakers, it is common to apply a low-pass filter and a delay of typically 10-15 ms in order to avoid localization of lateral sounds to the rear for listeners located away from the sweet spot (due to the "precedence effect" [10]).

A stereo widening effect may be achieved by adding this ambience signal to the original stereo signal as follows:

$$L' = L + a\,S,$$
$$R' = R - a\,S, \tag{2}$$

where the coefficient a controls the amount of additional ambience, and can be advantageously replaced by a filter that attenuates mid frequencies [11]. Equation (2) can be equivalently written:

$$L' = 0.707\ (C + w\,S),$$
$$R' = 0.707\ (C - w\,S), \tag{3}$$

where $w = 0.707\ (1 + 2a)$ controls the S/C ratio and can be interpreted as a relative image-width increase.

The signal C of equation (1) is intended for feeding the front center loudspeaker. The resulting basic 2-to-3 upmix scheme has several limitations, some of which may be addressed with the improved upmix matrices proposed by Gerzon [12] (these matrices will be further discussed in section 5 of this paper). Despite these improvements, fixed upmix matrices suffer from the following limitations:

- The frontal stereo image is narrower by roughly 25% on typical material [9] [12]. Wider spacing of the loudspeakers is necessary in order to preserve the original frontal stage width.
- The robustness of the sound image against lateral displacements of the listener is only moderately improved. Channel separation remains low, which causes image shifts towards the nearest loudspeaker because of the precedence effect.

2.4. Signal-dependent matrix

Signal-dependent (or "active") 2-to-4 matrix decoders were introduced in movie theaters in the late 1980's. They provide increased directional separation in multi-channel sound systems driven by 2-channel soundtracks [9].

In the following decade, matrix steering technology migrated into multi-channel home video systems [9], and digital implementations were introduced in some high-end receivers [5]. These 2×4 matrix decoders are designed for processing two-channel soundtracks encoded with the standard phase-amplitude downmix matrix [9] [13]:

$$L_T = L + 0.707\ (C - \mathrm{j}\ S),$$

$$R_T = R + 0.707\ (C + \mathrm{j}\ S),$$

where j denotes an idealized frequency-independent 90º phase shift.

In these active decoders, the signal separation between adjacent channels is improved by making the matrix coefficients signal-dependent. The decoder continuously monitors the left/right and front/back ratios in order to determine a dominant direction and its degree of salience in the incoming two-channel signal. This is used to emphasize signals in the desired loudspeakers and eliminate leakage into unwanted loudspeakers, while preserving the overall energy balance of the different sound components.

At the end of the 1990's, several companies introduced improved decoder designs optimized for producing two or four separate surround channels [13] [14]. Unlike the previous generation, these new matrix decoders are also proposed as upmix processors for non-encoded music material.

A comparative subjective assessment of several commercial 2-to-5 upmix processors, including both passive and active matrices, is described in [15]. Expert listeners were asked to judge frontal image quality, spatial impression and overall preference with excerpts of non-encoded 2-channel recordings (including classical music, sports broadcast and radio drama). The original two-channel versions of the recordings were preferred by a majority of the subjects over the upmixed versions. The artifacts reported include reduced focus, width or stability of the frontal image, and unstable or unnatural spatial impression.

Recently, a new algorithm specifically designed for 2-to-5 upmix of music recordings was described in [16], with apparently no consideration for matrix-encoded source material. Unlike previous active matrix decoders, it uses a cross-correlation measure to determine the amount of signal fed to the surround channels, and a running principal component analysis (PCA) to identify the dominant signal component and its panning position over the front channels.

The fundamental limitation of matrix steering technology remains that it is not able to track multiple dominant sound events simultaneously. Directional separation can only be improved when dialogue or a single sound effect dominates the soundtrack. Therefore, significant leakage of direct-path components into the surround channels inevitably occurs with complex music mixes.

2.5. Frequency-domain analysis/synthesis

The ability of the auditory system to separate sound events is often referred to as the "cocktail-party effect" [10]. The primary mechanism responsible for this ability is equivalent to computing the cross-correlation of the left and right ear signals in each critical band, followed by correlation across frequency [10].

There is also evidence that the auditory system is able to separate late reverberation tails from direct-path information. For instance, the direct-to-reverberant energy ratio is recognized as the principal cue to the perceived distance of sound events [10]. This ability is consistent with the above auditory processing model, since a characteristic feature of binaural reverberation signals is that they are weakly correlated.

This auditory process takes advantage of two facts: (a) we have two acoustic captors and (b) the sound events that we experience are largely disjoint in the time-frequency domain. Such considerations have inspired the general principles of a frequency-domain framework for analyzing and processing 2-channel audio recordings [17] [18]. As described in more detail in sections 4 and 5 of this paper, this approach enables the following operations:

- Identifying and extracting signals representative of the reverberation and ambience information present in a recording, with minimal leakage of primary (direct-path) signal components. This has applications for upmix, stereo widening and headphone virtualization.

- Discriminating multiple sound sources in a two-channel mix according to their panning positions, and extracting or re-distributing them over two or more channels. This enables 2-to-N upmix with high channel separation, and the accurate reproduction of the original stereo image (if desired), irrespective of the target loudspeaker layout geometry.

3. SIGNAL MODEL

In this section, we define a general model for the recorded audio signal, specify its decomposition into ambient components and primary components, and discuss design goals for 2-to-5 upmix processes.

3.1. Recording techniques

Two-channel or multi-channel recording techniques can be roughly categorized into two main classes: "live" (or natural), and "studio" (or artificial) [15].

Live recording involves two or more spatially distributed microphones, each of which may capture several sound sources simultaneously, with their individual directional cues encoded as inter-channel amplitude or time delay differences between the microphone signals. Ambience information (including reverberation and spatially distributed sources such as wind or audience noise) is naturally included in the recording and exhibits a weak correlation between the different channels.

In studio recording, each source (or instrument) is individually captured or synthesized, directionally panned and added into the mix. A source may also feed an artificial reverberator producing two or more weakly correlated output channels that are added into the mix. The final mix may also include some natural components recorded with live multi-microphone techniques, as described above.

3.2. Ambient and primary signals

The resulting P-channel recorded signal, which we will represent below as a $1\times P$ column matrix $\boldsymbol{X} = [\, X_1\, X_2 \ldots X_P \,]^{\mathrm{T}}$, can be decomposed into "primary" and "ambient" components as follows:

$$\boldsymbol{X} = \sum_{i=1}^{N} \boldsymbol{D}_i\, S_i \;+\; \sum_{i=1}^{N} \boldsymbol{R}_i\, S_i \;+ \boldsymbol{B}, \qquad (4)$$

where we simplify notations by adopting a frequency-domain representation and omitting the frequency dependence. The scalar signals S_i ($i = 1 \ldots N$) represent the individual source signals. The $1\times P$ matrices $\boldsymbol{D}_i$ and $\boldsymbol{R}_i$ represent respectively the direct-path transfer function and the reverberation response, for each source S_i. The $1\times P$ matrix $\boldsymbol{B}$ represents the P-channel background noise signal contributing to the surrounding ambience (e.g. audience noises, wind, etc.).

Ambient signal components are characterized by the fact that they are weakly correlated and evenly distributed across the channels. A primary component is usually directional and does not necessarily contribute to more than one channel.

Primary signal components are typically direct-path arrivals or strong isolated reflections. If a primary signal feeds two or more channels, its contributions to these channels are strongly correlated and its direction is determined by inter-channel differences in the direct-path transfer function $\boldsymbol{D}_i$. In the simplest cases, such as multi-track studio mixing, only frequency-independent amplitude differences are introduced. With coincident microphone techniques, the amplitude differences are somewhat frequency dependent and may be approximated by minimum-phase filters. With non-coincident microphone techniques, inter-channel delays will also be introduced. In the case of matrix-encoded surround recordings, directional encoding may also involve frequency-independent inter-channel phase differences, due to the 90º phase shift introduced on the surround signal S [14].

3.3. Design goals for a 2-to-5 upmix process

In order to determine design criteria for a 2-to-5 upmix process, one is led to consider what use the producer of the original recording would have made of five recording channels. While a spatially even distribution of the ambient components is probably a natural choice in most cases, various strategies are used for the distribution of the primary components, including the *direct/ambient* approach and the *in-the-band* approach [19].

In the direct/ambient approach to 5-channel mixing, the primary signal components are panned across the three front channels, and the surround channels exclusively receive ambient components. This suggests an upmix strategy that seeks to reproduce the original frontal image as accurately as possible (while taking advantage of the center channel), and seeks to spread the ambient components of the original recording into the surround channels. This is the least obtrusive design strategy for a high-fidelity upmix system that does not require user adjustments.

In the in-the-band approach, the primary signals are distributed across all five channels, which may suggest that the listener is surrounded by musicians (or may provide increased spatial articulation in an abstract composition). This approach could be accommodated in an upmix processor by providing, for instance, a controllable "wrap-around" effect (spreading the original stereo image towards the surround loudspeakers by an adjustable factor).

In reality, as suggested in [19], a producer might prefer to exploit the added freedom offered by the five channels and radically reorganize the relative placement of the sources (perhaps introducing dynamic movements). If only a two-channel mix is

provided, multi-channel remixing would require a sophisticated tool for extracting (or "unmixing") and re-panning individual sources. Although this direction is not explored in the present paper, the techniques presented in [18] make the necessary signal processing realizable to a significant extent.

In the next two sections of this paper, we will discuss methods for enhancing the reproduction of the ambient components and of the directional components of a two-channel recording, with consideration of the design goals discussed above, in the various playback system configurations (five loudspeakers, two loudspeakers, or headphones).

4. AMBIENCE ENHANCEMENT

4.1. Time-domain matrices

Fig. 3 shows a passive 2-to-5 matrix surround decoder, where the surround channel is delayed and filtered in order to maintain the localization of primary sound events in the front. In order to improve the spatial impression, the all-pass filters produce two weakly correlated surround signals from the mono surround channel ([2], [20]). The main limitation of this design is that the non-centered primary signal components are allowed to contribute to the surround channels. It relies heavily on the surround-channel delay in order to maintain frontal localization, and therefore provides a spatial impression that depends strongly on the position of the listener relative to the surround loudspeakers.

An active upmix matrix as described in section 2.4 may be expected to reduce the leakage of primary components into the surround channels, but cannot eliminate it completely. When a dominant signal component is detected, the active steering logic can eliminate or reduce its contribution to the surround channels. However, leakage cannot be avoided for the concurrent primary signal components that are panned to different directions.

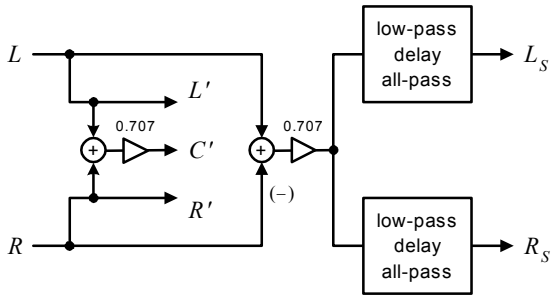

Fig. 3: passive 2-to-5 matrix surround decoder.

Whenever a source temporarily loses or gains the status of "dominant component", the amplitude of its contribution to the surround channels will tend to fluctuate. Such fluctuations, introduced by the steering logic, may cause listening fatigue and may perhaps explain some of the "unstable" spatial impressions reported in [15].

A more unexpected limitation of the design of Fig. 3 is its poor performance with source material that has been low-bit-rate encoded and decoded prior to upmixing. For instance, if a two-channel MP3-encoded sound file is decoded to linear PCM format before entering the upmix processor, a fluctuating timbre distortion can be heard in the surround channels. This very disturbing artifact may be attributed to the fact that the *L–R* signal difference reveals data-reduction effects which, in the incoming 2-channel signal, are masked by the center-panned components.

4.2. Frequency-domain upmix process

In order to provide a stable spatial impression over a wide listening area, it is necessary to minimize the leakage of primary (or direct-path) signal components into the surround channels. This requires a process capable of extracting two audio signals representative of the reverberation and surrounding ambience information present in a two-channel recording, excluding primary signal components. As mentioned in section 2.5, well known auditory processing mechanisms indicate that this is possible if the ambient components are disjoint from the primary components in the time-frequency domain, and if a criterion can be defined to distinguish the two in real time.

In the time-frequency plane, the correlation between the stereo channels will be high in regions where primary components are dominant, and low in regions dominated by ambient components. A similar rationale is used in the two-microphone speech de-reverberation algorithm described in [21]. In order to compute this inter-channel correlation criterion, we derive a left/right coherence index in the time-frequency plane as:

$$\phi(m,k) = \frac{\left| \phi_{LR}(m,k) \right|}{\sqrt{\phi_{LL}(m,k)\,\phi_{RR}(m,k)}}, \qquad (5)$$

where m is the discrete time index, k is the discrete frequency index, and $\phi_{ij}(m,k)$ denotes the running-time cross-correlation function of signals X_i and X_j.

The cross-correlation function $\phi_{ij}(m,k)$ can be computed as follows:

$$\phi_{ij}(m,k) = (1-\lambda)\,\phi_{ij}(m-1,k) + \lambda\, C_{ij}(m,k),$$

where λ is a forgetting factor and C_{ij} is the instantaneous cross-spectrum at time m, defined by:

$$C_{ij}(m,k) = X_i(m,k)\, X_j^*(m,k),$$

where $X_i(m,k)$ is the short-time Fourier transform (STFT) of the time-domain signal $x_i(k)$, and index i represents one of the two channels, L or R.

The coherence index $\phi(m,k)$ can readily be used to identify the regions dominated by ambient signal components in the two-channel signal, provided that the left and right signals have comparable energies across a few successive time frames. The *ambience index*, defined as:

$$\Phi(m,k) = 1 - \phi(m,k), \qquad (6)$$

can be directly used as a spectral modification function.

Fig. 4 shows the signal flow diagram of the ambience extraction algorithm (illustrated for only one of the sub-bands indexed by k). In typical recordings, some ambient signal components will partially overlap low-energy primary components from other sources. The reconstructed ambience signals will then include some amount of residual leakage from primary components. This undesired leakage can be reduced by mapping the ambience index with a non-linear soft-decision function $\Gamma(\Phi)$. This and additional algorithm details are described in [17], along with simulation results illustrating the effectiveness of the ambience extraction algorithm.

4.3. Applications

Fig. 5 shows a complete block diagram of the proposed frequency-domain 2-to-5 upmix processor. The system includes the ambience extraction algorithm of Fig. 4, as well as a 2-to-3 front-channel upmix algorithm to be described in section 5.2 (Fig. 6 or Fig. 7).

The extracted ambience signals are passed through left and right all-pass filters and delays, as in the 2-to-5 upmix matrix of Fig. 3. Here, the all-pass filters are introduced in order to avoid the formation of lateral phantom images (because the extracted ambience signals are, by nature, strongly correlated with the original ambient components reproduced through the front channels). The role of the delays and low-pass filters is to avoid frontal image alterations that could be caused by residual leakage of primary signal components into the surround channels (although the occurrence and amount of such leakage is considerably reduced here in comparison to the system of Fig. 3).

Improved stereo widening or headphone virtualization systems may be designed by combining the upmix scheme of Fig. 5 with the head-related virtualization techniques reviewed in section 2.1. In the stereo widening system, the extracted ambience signals are reproduced through virtual left and right surround loudspeakers. In the headphone virtualization system, the five output signals of the upmix processor are reproduced through five virtual loudspeakers located at the positions indicated in Fig. 1 (instead of using only two virtual front loudspeakers receiving the original L and R signals).

Therefore, the reproduction of a two-channel recording over five loudspeakers, two loudspeakers or headphones can be enhanced with natural sounding, enveloping reverberation and ambience information exclusively derived from audio cues originally present in the recording, with minimal alteration of the stereo image formed by the primary signal components.

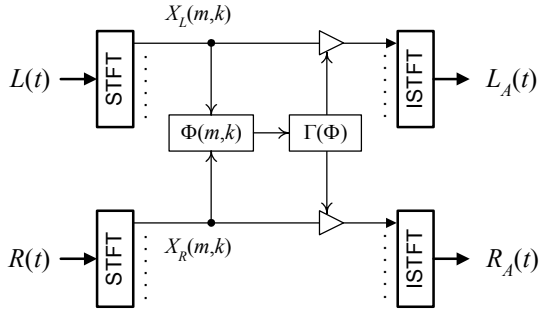

Fig. 4: ambience extraction process.

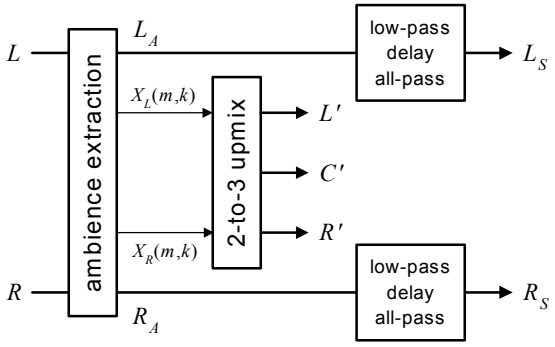

Fig. 5: block diagram of 2-to-5 upmix system.

5. DIRECTIONAL ENHANCEMENT

5.1. Time-domain matrices

Feeding the center loudspeaker with signal C defined in equation (1) yields the original 2×3 upmix matrix attributed to Klipsch [9]:

$$\begin{pmatrix} L' \\ C' \\ R' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0.707 & 0.707 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} L \\ R \end{pmatrix}. \quad (7)$$

The factor 0.707 has the effect of equalizing the energy of the three channels when L and R are uncorrelated and of equal energy. This matrix amplifies total signal energy by a factor 3/2 on average (1.76 dB). Its main benefit is that center-panned sources are reproduced more naturally than by phantom imaging over two loudspeakers [9]. However, the addition of the center loudspeaker also has the effect of narrowing the sound image by approximately 25% with typical stereo material [9]. This matrix also affects the balance of the recording: center-panned sounds are boosted by 1.25 dB relative to sounds panned to the sides. Gerzon addressed this issue by defining a general class of energy preserving M-to-N upmix decoders [12]. In the 2-to-3 case, these decoders are defined by:

$$\begin{aligned} C' &= C \cos\theta, \\ L' &= 0.707\ (C \sin\theta + wS), \\ R' &= 0.707\ (C \sin\theta - wS), \end{aligned} \quad (8)$$

where C and S are defined in equation (1) and θ is an angle between 0° and 90°. The gain factor w applied to the difference signal S is provided for optional width control, as in the stereo widening scheme defined by equation (3). However, the decoder is not energy-preserving when $w > 1$: the energy of signals panned left or right is increased by a factor $(1 + w^2) / 2$ relative to the energy of center-panned components. The requirement to preserve the balance of sources in the mix imposes $w = 1$ and leaves only θ as a free variable for optimization.

The optimal value for θ is determined with the help of the vector based psychoacoustic localization theory developed by Gerzon (reviewed in [12]). The purpose of the optimization is to ensure that amplitude-panned sounds in the original stereo signal will be perceived at their original intended position, with an improvement in image stability and sharpness. However, all the solutions proposed in [12] require that the left and right loudspeakers be positioned more widely when the center loudspeaker is added.

A typical value for θ in equation (8) is 50°, which yields the following 2×3 upmix matrix:

$$\begin{pmatrix} L' \\ C' \\ R' \end{pmatrix} = \begin{pmatrix} 0.8830 & -0.117 \\ 0.4545 & 0.4545 \\ -0.117 & 0.8830 \end{pmatrix} \begin{pmatrix} L \\ R \end{pmatrix}. \quad (9)$$

With such fixed (passive) 2-to-3 upmix matrices, centered sounds will not be eliminated from the left and right loudspeakers, and sounds panned to a side will not be removed from the opposite loudspeaker. As a result, despite the addition of a center loudspeaker, the perceived localization of both centered voices and spatialized music is still significantly altered for listeners seating on either side of the sweet spot.

With an active upmix matrix, it is possible to resolve this issue and achieve exact preservation of the localization of a single amplitude-panned source, without repositioning the left and right loudspeakers [14] [16]. However, the steering logic cannot discriminate several concurrent sources panned in different directions. With music recordings, there are often passages where no dominant primary component stands out in the mix. In such passages, an active upmix matrix will perform similarly to a passive matrix, and will therefore narrow the stereo image. As a given source gains or loses the status of dominant component, the panning positions of other sound sources present in the mix may fluctuate slightly, which can cause the stereo image to appear less stable than with a passive upmix matrix.

5.2. Frequency-domain source separation

Ideally, a 2-to-3 upmix matrix should direct all center-panned signal components to the center channel and remove them from the left and right channels (although some listeners may prefer to allow for an adjustable residual amount of center-to-side leakage). A similar problem is the removal of the singer's voice from a recording, useful in applications such as karaoke. A frequency-domain voice suppression algorithm was developed by Laroche for this purpose [22]. Its basic assumptions are that the voice is center-panned and that it is disjoint from the other signal components in the time-frequency domain.

More generally, an N-to-M upmix scheme should be able to redistribute any primary signal component, irrespective of its position, over the two loudspeakers closest to that position, with adequate panning weights. This process can be called *pairwise upmix.*

Source separation and upmix techniques for pairwise upmix were proposed in [18]. The signal separation relies on the computation of a *panning index*, function of time and frequency. A primary signal component corresponds to a cluster of locations that have the same panning index in the time-frequency plane. This source separation technique is similar in philosophy to the blind separation technique described in [23]. Once identified in this manner, a primary signal component may be attenuated, redistributed over multiple different channels, or selected to reconstruct a separate time-domain signal.

The panning index is derived by comparing the left and right signals in the time-frequency plane. We define the left-right similarity measure:

$$\varphi(m,k) = \frac{2\, X_L(m,k)\, X_R^*(m,k)}{|X_L(m,k)|^2 + |X_R(m,k)|^2}, \qquad (10)$$

where $X_L(m,k)$ and $X_R(m,k)$ denote the short-time Fourier transforms of the two signals. The magnitude of $\varphi(m,k)$ is bounded within [0, 1] and reaches 1 when the two signals have the same magnitude. The phase of $\varphi(m,k)$ is equal to the inter-channel phase difference.

As illustrated in Fig. 6, the center-channel signal is extracted from the recording by selecting the time-frequency bins that correspond to a similarity measure of 1 and synthesizing a signal by inverse STFT. This signal is subtracted form the left and right channels so that the three-channel presentation remains spatially undistinguishable form the two-channel presentation for a listener located at the sweet spot. Experience has shown that while the synthetic center channel might include artifacts (due to STFT processing) that are audible when it is heard alone, the overall presentation is of very high fidelity.

In order to implement a 2-to-3 pairwise upmix system, we must modify the system of Fig. 6 so that signal components that are panned to the left in the original stereo image be redistributed over the left and center channels only, with appropriate weights, and similarly for the right side (Fig. 7). For this purpose, we define a panning index as follows.

If we assume that a single amplitude-panned source is present in the mix with no reverberation or ambience, the signal model (4) yields:

$$\varphi(m,k) = \frac{2\, d_L\, d_R}{d_L^{\,2} + d_R^{\,2}}, \qquad (11)$$

where d_L and d_R are the left and right panning coefficients.

Following previous authors (e. g. [9], [14], [16]), we can define a panning index as follows:

$$\alpha = \arctan(d_L / d_R). \qquad (12)$$

With this definition, the panning index α varies between 0 (right) and $\pi/2$ (left), and is equal to $\pi/4$ for centered sounds. This definition is consistent with the conventional sine/cosine panning law (Fig. 8). The similarity measure (10) is then converted into a panning index value by:

$$\alpha(m,k) = \frac{\arcsin \varphi(m,k)}{2}, \qquad (13)$$

where the left/right ambiguity resulting from the definition of $\varphi(m,k)$ can be resolved e. g. by the sign of the difference $|X_L(m,k)| - |X_R(m,k)|$.

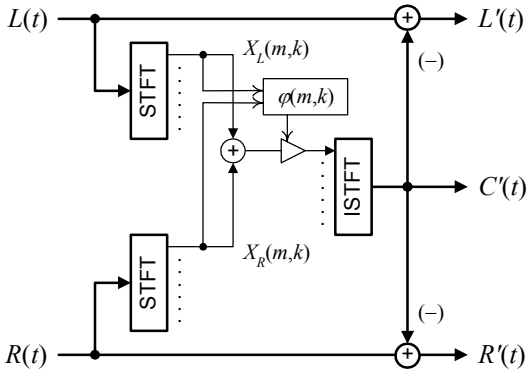

Fig. 6: center-channel extraction process.

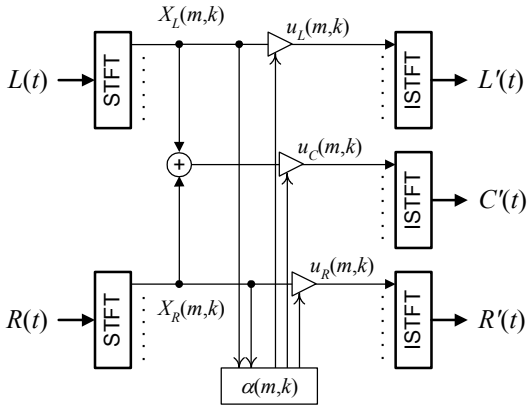

Fig. 7: 2-to-3 pairwise upmix process.

A 3-channel panning index can be defined by doubling α, as suggested in [16], so that $\sin(2\alpha)$ is equal to the center-channel panning weight and $|\cos(2\alpha)|$ is the left or right panning weight, depending on the sign of $\cos(2\alpha)$, as shown in Fig. 8. The 3-channel re-panning weights in Fig. 7 are then computed as follows (and also plotted in Fig. 8):

$$u_C = \frac{\sin 2\alpha}{\sin\alpha + \cos\alpha},$$
$$u_L = -\min\left(0, \frac{\cos 2\alpha}{\sin\alpha}\right), \quad (14)$$
$$u_R = \max\left(0, \frac{\cos 2\alpha}{\cos\alpha}\right),$$

where, for clarity, the dependence on m and k is omitted.

In [18] the re-panning coefficients are computed differently, according to the psychoacoustic localization theory of [12], which yields different optimal pairwise panning laws for low frequencies and high frequencies [24]. The panning laws and re-panning weights can be defined for different layouts or numbers of channels. It is possible to "warp" the stereo image by applying a mapping function to the panning index (allowing for a "wrap around" effect).

The frequency-domain pairwise upmix technique described here is limited to handling recordings using amplitude panning or coincident microphone techniques. In order to achieve compatibility with non-coincident microphone techniques, it would be necessary to incorporate the inter-channel time differences in the definition of the panning index. Methods for estimating inter-channel delays in the time-frequency plane are suggested in [21] and [23].

5.3. Applications

The center-channel extraction process (Fig. 6) is effective for anchoring centered sounds in the center loudspeaker. For a listener located at the sweet spot, the original two-channel stereo image is faithfully reproduced over three channels, irrespective of the recording technique used. However, with lateral displacements of the listener, non-centered sounds will shift towards the closest loudspeaker. The pairwise upmix process (Fig. 7) removes left-panned sounds from the right loudspeaker and vice versa, and takes full advantage of the three channels to stabilize the stereo image for off-center listeners. At the sweet spot, the original stereo image is faithfully reproduced for studio recordings (amplitude panned) or coincident stereo recordings.

In headphone reproduction, pairwise upmix is not necessary. Center-channel extraction is sufficient because the listener is always at the sweet spot of the virtual loudspeaker system. The main improvement is a more convincing frontal localization of centered signals, because these are virtualized via the frontal HRTF instead of a phantom image produced through the left and right virtual loudspeakers.

6. CONCLUSION

The main focus of this paper is the playback of two-channel recordings over five or two loudspeakers or over headphones. We have reviewed several existing virtualization and matrixing techniques, and shown that some of their fundamental limitations may be addressed by a frequency-domain signal processing framework motivated by auditory models. Surround ambience enhancement is achieved by use of an ambience extraction process that minimizes unwanted leakage of primary (direct-path) signal components. For directional enhancement, the framework allows for high channel separation with multiple concurrent audio sources. The overall result, verified in informal listening tests, is an enhanced listening experience with minimal unwanted alterations of the original stereo image. However, further work is needed for extending the pairwise 2-to-3 upmix technique to signals mixed with inter-channel delay or phase differences, and for a better understanding of the effects of signal overlap in the time-frequency plane. Future work should also include a more formal experimental assessment of the performance of the proposed techniques.

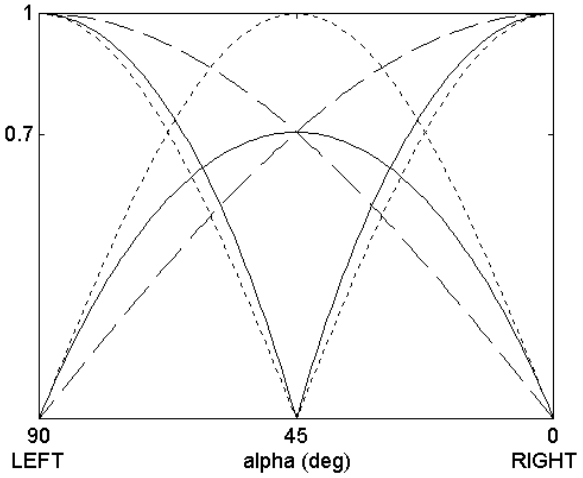

Fig. 8: 3-channel pairwise re-panning coefficients vs. panning index α (solid), 2-channel panning law (dashed), and 3-channel panning law (dotted).

7. ACKNOWLEDGMENTS

The authors would like to thank their colleagues Luke Dahl, Adrian Jost, Veronique Larcher, Jean Laroche and Alan Seefeldt for their contributions to the investigation of techniques covered in this paper, and Mark Dolson for supporting this research.

Some of the methods described in this paper are covered by issued patents or pending patent applications.

8. REFERENCES

[1] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*, Academic Press, 1994.

[2] D. Rocchesso, “Spatial Effects,” in *DAFX Digital Audio Effects*, John Wiley & Sons, 2002

[3] J.-M. Jot, V. Larcher, O. Warusfel, “Digital signal processing issues in the context of binaural and transaural stereophony,” presented at the AES 98th convention, preprint 3980, 1995.

[4] P. Rubak, “Headphone Signal Processing System for Out-of-the-head Localization,” presented at the AES 90th convention, preprint 3063, 1991.

[5] D. Griesinger, "Theory and Design of a Digital Audio signal Processor for Home Use," *J. Audio Eng. Soc.*, vol. 37, no. 1/2, p. 40, 1989.

[6] W. G. Gardner, “Reverberation algorithms,” in *Applications of Signal Processing to Audio and Acoustics*, Kluwer Academic, 1998.

[7] J.-M. Jot, “Efficient Models for Reverberation and Distance Rendering in Computer Music and Virtual Audio Reality," *Proc. International Computer Music Conference*, 1997.

[8] C. Kyriakakis and A. Mouchtaris, "Virtual Microphones for Multichannel Audio Applications," *Proc. IEEE ICME 2000*, vol. 1, pp. 11-14, August 2000.

[9] S. Julstrom, "A High-Performance Surround Sound Process for Home Video," *J. Audio Eng. Soc.*, vol. 35, no. 7/8, p. 536, 1987.

[10] J. Blauert, *Spatial Hearing*, MIT Press, 1983.

[11] A. Klayman, “Stereo Enhancement System,” United States Patent no. 5,661,808, August 1997.

[12] M. A. Gerzon, "Optimum Reproduction Matrices for Multispeaker Stereo," *J. Audio Eng. Soc.*, vol. 40, no. 7, p. 571, 1992.

[13] R. Dressler, "Dolby Surround Pro Logic II Decoder: Principles of Operation,” www.dolby.com/tech/l.wh.0007.PLIIops.pdf.

[14] D. Griesinger, "Multichannel Matrix Surround Decoders for Two-Eared Listeners," presented at the AES 101st convention, preprint 4402, 1996.

[15] F. Rumsey, "Controlled Subjective Assessment of Two-to-Five Channel Sound Processing Algorithms," *J. Audio Eng. Soc.*, vol. 47, no. 7/8, p. 563, 1999.

[16] R. Irwan, R. M. Aarts, "Two-to-Five Channel Sound Processing," *J. Audio Eng. Soc.*, vol. 50, no. 11, p. 914, 2002.

[17] C. Avendano, J.-M. Jot, “Ambience Extraction and Synthesis from Stereo Signals for Multi-channel Audio Up-mix,” *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2002.

[18] C. Avendano and J.-M. Jot, "Frequency-domain Techniques for Stereo to Multichannel Upmix." *Proc. AES 22nd International Conference*, June 2002.

[19] T. Holman, “Mixing the Sound,” *Surround Professional*, p. 35, June 2001.

[20] M. Schroeder, "An Artificial Stereophonic Effect Obtained from a Single Audio Signal," *J. Audio Eng. Soc.*, vol. 6, no. 2, p. 74, 1958.

[21] J. Allen, D. A. Berkeley, J. Blauert, “Multi-microphone Signal-processing Technique to Remove Room Reverberation from Speech Signals,” *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912-915, 1977.

[22] J. Laroche, "Process for Removing Voice from Stereo Recordings," United States Patent no. 6,405,163, September 1999.

[23] A. Jourjine, S. Richard, O. Yilmaz, "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, p. 2985, April 2000.

[24] J.-M. Jot, V. Larcher, J.-M. Pernaux, “A Comparative Study of 3-D Audio Encoding and Rendering Techniques,” *Proc. AES 16th International Conference*, April 1999.