# FREQUENCY DOMAIN TECHNIQUES FOR STEREO TO MULTICHANNEL UPMIX

**CARLOS AVENDANO AND JEAN-MARC JOT**

*Creative Advanced Technology Center*
*1500 Green Hills Road, Scotts Valley, California, USA.*
carlosa@atc.creative.com jmj@atc.creative.com

In this paper we propose a series of upmixing techniques for generating multichannel audio from stereo recordings. The techniques use a common analysis framework based on the comparison between the Short-Time Fourier Transforms of the left and right stereo signals. An inter-channel coherence measure is used to identify time-frequency regions consisting mostly of ambience components, which can then be weighed via a non-linear mapping function, and extracted to synthesize ambience signals. A similarity measure is used to identify the panning coefficients of the various sources in the mix in the time-frequency plane, and different mapping functions are applied to unmix (extract) one or more sources, and/or to re-pan the signals into an arbitrary number of channels. We illustrate the application of the various techniques in the design of a two-to-five channel upmix system.

## INTRODUCTION

The availability of multichannel audio recordings is still very limited nowadays. While recent movie soundtracks and some musical recordings are available in multichannel format, most music recordings are mixed in stereo. The playback of this material over a multichannel system poses a fundamental problem: stereo recordings are mixed with a very particular set up in mind, which consists of a pair of loudspeakers placed symmetrically in front of the listener. Thus, listening to this kind of material over a multi-speaker system (e.g. 5.1 surround) raises the question of what signals should be sent to the additional channels (e.g. surround and center channels). Unfortunately, the answer to this question depends strongly on individual preferences and no definite objective criteria exist.

In this paper we present a solution strategy based on deriving new signals from the stereo recording and mixing them according to a model of how multichannel audio is generally mixed. There are two main approaches for mixing multichannel audio [7]. One is the *direct/ambient* approach, in which the different sources (e.g. instruments) are panned among the front channels in a frontally oriented fashion as is commonly done with stereo mixes, and *ambience* signals are distributed among all channels. This mix creates the impression that the listener is in the audience, in front of the stage (best seat in the house). The second approach is the *in-the-band* approach, where the instrument and ambience signals are panned among all the loudspeakers, creating the impression that the listener is surrounded by the musicians.

Thus, the signal processing techniques necessary to implement this solution should be able to extract the signals of individual instruments as well as the ambience signals. In reality this is a very difficult task since no information about how the stereo mix was done is available in most cases. However, knowledge about how stereo audio is recorded and mixed can be used to derive a reasonable model of the stereo signal.

Based on this model, the techniques proposed in this paper are capable of achieving this signal extraction or segregation to a great extent. The basic idea is to compare the STFT's (Short-Time Fourier Transforms) of the left and right stereo signals in order to identify different components in the mix. An inter-channel coherence measure is used to identify ambience components, and an inter-channel similarity measure is used to identify the panning coefficients corresponding the various individual instruments in the mix. After identification of these components, a non-linear mapping function is used to weigh and extract the time-frequency regions of interest and an inverse STFT (via overlap-and-add) is used to synthesize the new signals.

The paper is organized as follows. First we give some background on upmixing techniques, followed by the description of a model of the stereo signal. We then present a mathematical framework from which we derive two measures or indices that are the basis of the proposed upmix techniques. We finally present the application of the techniques to the design of a two-to-five upmix system and finish with a discussion and conclusions.

## 1. BACKGROUND

The existing two-to-N channel upmix algorithms can be classified in two broad classes: ambience generation techniques which attempt to extract and/or synthesize the ambience of the recording and deliver it to the surround channels (or simply enhance the natural ambience), and multichannel converters that derive additional channels for playback in situations when there are more loudspeakers than program channels. In the latter case, the goal is to increase the listening area while preserving the original stereo image.

The ambience generation methods rely on combinations of the following methods:

- Applying artificial reverberation to the stereo signal. The resulting impression is essentially of listening to the original recording in a virtual listening room (see e.g. [3]). This artificial ambience information does not match the conditions in which the original recording was produced.

- Computing the difference of the original left and right signals. This provides a monaural signal whose content includes the desired ambience information and excludes any primary signal panned to the center of the original stereo image. However, the resulting ambience signal also contains unwanted leakage from any primary signals not panned to the center. This leakage can be partially reduced by use of *logic steering* techniques (see e.g. [5]).

- Deriving a stereo ambience signal from a mono signal (pseudostereophony). Two weakly correlated signals can be obtained by applying two allpass filters with different phase characteristics to a single audio signal [13].

- Applying a small delay (typically $5$ to $20$ ms) on the rear-channel signals to alleviate unwanted localization artifacts caused by any leakage of primary signals into the rear channels [5, 12]. This is an effective method for better preserving the frontal stereo image of the original recording, but it cannot correct the ambience information itself.

- Deriving room responses corresponding to *virtual microphone* positions so as to synthesize reverberation signals that match the acoustics of the original venue [9]. However, the application of this method is in principle restricted to live recordings for which detailed additional historical information is available on the original recording conditions and techniques. Also the method cannot reproduce other ambience components due to background noise in the original recording.

Multichannel converters can be categorized in the following two classes:

- Linear matrix converters, where the new signals are derived by scaling and adding and/or subtracting the left and right signals [10, 6]. Mainly used to create a two-to-three channel upmix, this method inevitably introduces unwanted artifacts and preservation of the stereo image is compromised.

- Matrix steering methods which are basically dynamic linear matrix converters [5]. These methods are capable of detecting and extracting prominent sources in the mix such as dialogue, even if they are not panned to the center. Gains are dynamically computed and used to scale the left and right channels according to a dominance criterion. Thus a source (or sources) panned in the *primary* direction can be extracted. However, this technique is still limited to looking at a primary direction, which in the case of music might not be unique.

## 2. SIGNAL MODEL

Stereo recordings can be roughly categorized into two main classes: *studio* or artificial, and *live* or natural [12]. In *studio* recording, the different sources (or instruments) are individually recorded and then mixed into a single stereo signal. Stereo reverberation is then added artificially to the mix. In general, the left and right impulse responses of the reverberation processor are different (weakly correlated) to increase the perception of spaciousness [4]. *Live* recording involves a number of spatially distributed microphones to capture all sound sources. With stereo microphone techniques, the ambience is naturally included in the recording and presents a weak correlation between the left and right channels [12] [7]. Notice that the ambience is not only produced by reverberation, but can be introduced by other spatially distributed sources such as wind, audience noise, etc.

A signal model can be defined as follows: assume that there are $N$ sources $s_j(t), j = 1, ..., N$ convolved with room impulse responses (or artificial reverberation responses) $h_{ij}(t)$ to generate the left ($i = 1$) and right ($i = 2$) stereo channels respectively. The responses $h_{ij}(t)$ can be split into a direct-path or primary signal component and a reverberation component as $h_{ij}(t) = d_{ij}(t) + r_{ij}(t)$, and the stereo signal can be written as:

$$x_i(t) = \sum_{j=1}^{N} s_j(t) * d_{ij}(t) + \sum_{j=1}^{N} s_j(t) * r_{ij}(t) + n_i(t), \quad (1)$$

where $n_i(t)$ are background noise signals contributing to the surrounding ambience (e.g. audience noise, wind, etc.). In (1) the left term corresponds to the primary signals and the right term corresponds to the ambience signals. In general the ambience signals will have comparable levels, i.e. $||r_{1j}(t)||^2 \simeq ||r_{2j}(t)||^2$ and $||n_1(t)||^2 \simeq ||n_2(t)||^2$.

Notice that if there is a source $s_j(t)$ panned to the center (i.e. $d_{2j}(t) = d_{1j}(t)$) the difference of the left and right signals will eliminate its direct-path component and yield a signal that contains only the reverberation information corresponding to this source, i.e. $x_1(t) - x_2(t) = s_j(t) * (r_{1j}(t) - r_{2j}(t))$. However, notice that only the ambience of signals panned to the center can be extracted with this method, and the ambience component is monaural and consists of the difference between the stereo ambience signals (in fact this approach is commonly used to extract ambience information from stereo recordings, as described in Section 1, and we readily see the limitations).

## 3. FRAMEWORK

In this section we describe the mathematical framework on which the proposed up-mix algorithms are based. Our signal processing front-end consists of an STFT analysis stage. Let us first denote the STFT's of the channel signals $x_i(t)$ as $X_i(m,k)$, where $m$ is the time index and $k$ is the frequency index. We define the following statistical quantities:

$$\phi_{ij}(k) = \mathrm{E}\{X_i(m,k)X_j^*(m,k)\}, \qquad (2)$$

where E is the expectation operator with respect to $m$ and the superscript $*$ denotes complex conjugation. Audio signals are in general non-stationary. For this reason the statistics will change with time. To track the changes of the signal and to be able to implement a causal system we introduce a forgetting factor $\lambda$ in the computation of the cross-correlation functions in (2). Thus in practice these statistics are computed as short-time functions:

$$\begin{aligned}\phi_{ij}(m,k) = \; & (1-\lambda)\phi_{ij}(m-1,k) + \\ & \lambda X_i(m,k)X_j^*(m,k).\end{aligned} \qquad (3)$$

Using these statistical quantities we define the inter-channel short-time coherence function as:

$$\phi(m,k) = \frac{|\phi_{12}(m,k)|}{[\phi_{11}(m,k)\phi_{22}(m,k)]^{\frac{1}{2}}}. \qquad (4)$$

The coherence function above is real and bounded between zero and one, with values close to one in time-frequency regions of high inter-channel correlation, and values close to zero in regions of weak inter-channel correlation [1].

Another useful measure to compare the stereo signals is obtained from (3) by setting the forgetting factor to one. This similarity function is defined as:

$$\psi(m,k) = 2\frac{|\psi_{12}(m,k)|}{[\psi_{11}(m,k)+\psi_{22}(m,k)]}, \qquad (5)$$

where the scaling factor 2 is introduced for normalization purposes and $\psi_{ij}(m,k) = \phi_{ij}(m,k)|_{\lambda=1.}$ The similarity function above is real and bounded between zero and one, and its value will depend on the relative levels between the left and right signal components at each time frame. An additional measure that will be useful in the following sections is the partial similarity function:

$$\psi_i(m,k) = \frac{|\psi_{ij}(m,k)|}{\psi_{ii}(m,k)}. \qquad (6)$$

Next we describe how the functions in (4), (5) and (6) can be used to extract information from the stereo signal.

### 3.1. Ambience Index

In the time-frequency plane, the time correlation between the stereo channels at each frequency and time will be high in regions where the direct component is dominant, and low in regions dominated by the reverberation or ambience [2]. Given its properties, the coherence function in (4) can readily be used to identify the regions dominated by ambience in the stereo signal. We define the ambience index as:

$$\Phi(m,k) = 1 - \phi(m,k), \qquad (7)$$

where regions of low coherence will have values close to one, i.e. an indication of ambience, while high coherence values will have values close to zero, i.e primary components. An additional criterion to identify ambience components is that the left and right signals have to have comparable energies across a few frames. The reason is that the coherence will be low (or zero) for primary signals panned completely to one side. In Section 4.1 we show how this index is applied to extract and generate ambience signals.

### 3.2. Panning Index

In this section we describe how the similarity function in (5) can be used to identify the various sources in the stereo mix based on their panning coefficients. This is accomplished via a panning index (applications where this index is useful are presented in Section 4). To derive this panning index let us first simplify our signal model and assume that only one amplitude-panned source $s_j(t)$ is present in the mix and that there are no ambience components. Thus, from the signal model in (1) the left and right signals can be written as $x_1(t) = (1-\alpha)s_j(t)$ and $x_2(t) = \alpha s_j(t)$ respectively, where $d_{10}(t) = (1-\alpha)$ and $d_{20}(t) = \alpha$. The similarity function (5) will have a value proportional to the panning coefficient $\alpha$ at those time/frequency regions where the source has energy, i.e.

[1] Notice that different values of $\lambda$ can be used in different frequency bands.

[2] A similar rationale is used by the two-microphone speech dereverberation algorithm described in [1].

$$\psi(m,k) = 2\frac{\alpha(1-\alpha)\left|S_j(m,k)S_j^*(m,k)\right|}{\alpha^2|S_j(m,k)|^2 + (1-\alpha)^2|S_j(m,k)|^2},$$

where $S_j(m,k)$ is the STFT of $s_j(t)$, which cancels out and yields

$$\psi(m,k) = 2\frac{\alpha-\alpha^2}{\alpha^2+(1-\alpha)^2}.$$

The similarity function for various values of $\alpha$ is plotted in the left panel of Figure 1. If the source is center-panned (i.e. $\alpha = 0.5$), then the function will attain its maximum value of one, and if the source is panned completely to either side, the function will attain its minimum value of zero. In other words the function is bounded.

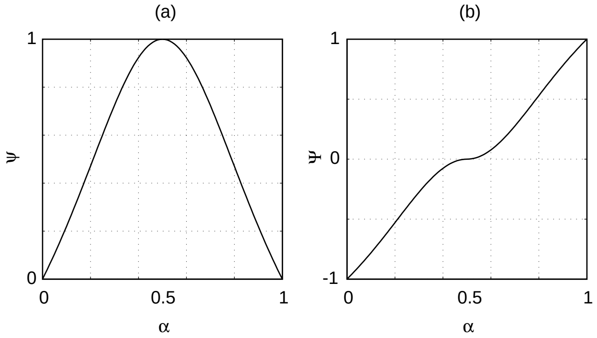

Figure 1: (a) Similarity function and (b) panning index as functions of panning coefficient $\alpha$.

Notice, however, that given the quadratic dependence on $\alpha$, the function (5) is multi-valued and symmetrical about $0.5$. That is, if a source is panned say at $\alpha = 0.2$, then the similarity function will have a value of $0.47$, but a source panned at $\alpha = 0.8$ will have the same similarity value. While this ambiguity might appear to be a disadvantage for source localization and segregation, it can easily resolved using the difference between the partial similarity measures in (6). The difference is computed simply as

$$\Delta(m,k) = \psi_1(m,k) - \psi_2(m,k), \tag{8}$$

and we notice that time-frequency regions with positive values of $\Delta(m,k)$ correspond to signals panned to the left (i.e. $\alpha < 0.5$), and negative values correspond to signals panned to the right (i.e. $\alpha > 0.5$). A value of $\Delta(m,k)$ equal to zero corresponds to non-overlapping time-frequency regions of signals panned to the center. Thus we can define an ambiguity-resolving function as

$$\widehat{\Delta}(m,k) = \begin{cases} 1 & \text{if} \quad \Delta(m,k) > 0 \\ 0 & \text{if} \quad \Delta(m,k) = 0 \\ -1 & \text{if} \quad \Delta(m,k) < 0 \end{cases} \tag{9}$$

Shifting and multiplying the similarity function by $\widehat{\Delta}(m,k)$ we obtain a new index, which is anti-symmetrical, still bounded but whose values now vary from one to minus one as a function of the panning coefficient, i.e.

$$\Psi(m,k) = \left[1-\psi(m,k)\right]\widehat{\Delta}(m,k), \tag{10}$$

This panning index is shown in the right panel of Figure 1. Notice that now there is a one-to-one relationship between panning coefficient $\alpha$ and panning index. In Section 4 we describe the several applications of the short-time panning index.

## 4. UPMIX TECHNIQUES

In this section we describe the application of the ambience and panning indices to the design of various upmixing techniques. Two classes of techniques are described, one of which extracts ambience information and synthesizes a stereo signal consisting mostly of the ambience in the recording. The other techniques use the panning index to identify, separate (unmix), and re-pan amplitude-panned sources.

### 4.1. Ambience Extraction and Synthesis

We have previously described the ambience extraction algorithm in [2]. In this section we briefly summarize its operation. The main idea is to weigh the right and left STFT's of the input signal by the ambience index and apply an inverse STFT to these magnitude-modified transforms. A more general form is to weigh the transforms with a non-linear function of the ambience index, i.e.:

$$A_i(m,k) = X_i(m,k)\Gamma\left[\Phi(m,k)\right], \tag{11}$$

where $A_i(m,k)$ are the modified, or ambience transforms. The behavior of the non-linear function $\Gamma$ that we desire is one in which regions with large ambience index values are not modified and regions with small ambience index are heavily attenuated to remove the primary signal component. Additionally, the function should be smooth to avoid artifacts. One function that presents this behavior is the hyperbolic tangent. Thus we define $\Gamma$ as:

$$\Gamma\left[\Phi\right] = \left(\frac{\mu_1-\mu_0}{2}\right)\tanh\{\sigma\pi\widehat{\Phi}\} + \left(\frac{\mu_1+\mu_0}{2}\right) \tag{12}$$

with $\widehat{\Phi} = \Phi - \Phi_o$, where $\Phi_0$ is a threshold, the parameters $\mu_1$ and $\mu_0$ define the range of the output, and $\sigma$ controls the slope of the function. In Figure 2 we show the function $\Gamma$ for different values of the parameters. In general the value of $\mu_1$ is set to one since we do not wish to enhance the non-coherent regions (though this could be useful in other contexts). The value of $\mu_0$ determines the

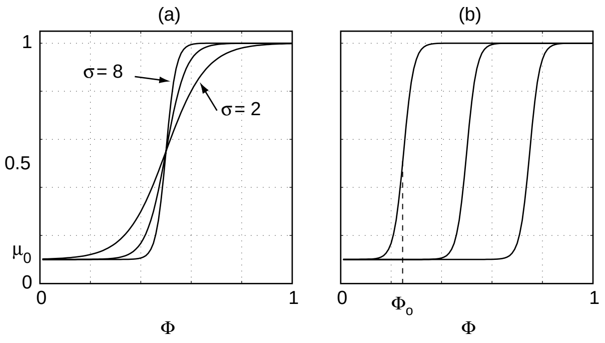

Figure 2: Mapping function for the ambience index (a) as a function of the parameter $\sigma$ with $\Phi_0 = 0.5$. (b) As a function of the offset $\Phi_0$, with $\sigma = 4$. In both cases $\mu_0 = 0.1$.

floor of the function and it is important that this parameter be set to a small value greater than zero to minimize spectral-subtraction-like artifacts at the output.

A block diagram of the ambience signal extraction algorithm is shown in Figure 3. The inputs to the system are the left and right channel signals of the stereo recording, which are first transformed into the short time frequency domain. The ambience index is mapped to generate the multiplication coefficients that modify the short-time transforms. After modification, the time domain ambience signals $a_1(t)$ and $a_2(t)$ are synthesized by applying the inverse short-time transform via the overlap-and-add method. In Section 5 we describe an application of this technique in the context of a direct/ambient upmixing system. For more details and simulation results of the technique see [2].

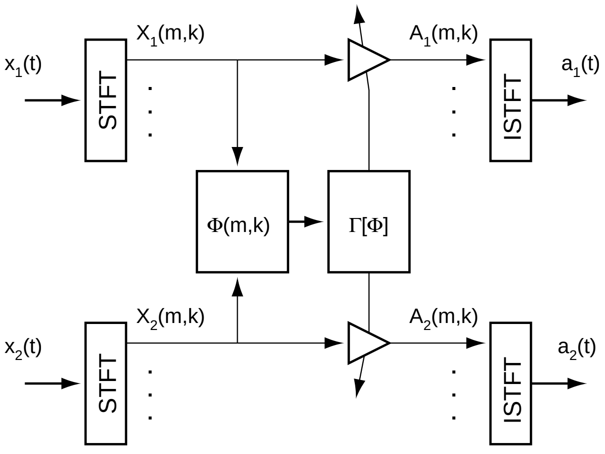

Figure 3: Block diagram of the ambience extraction system. Only the processing for one frequency band is shown.

### 4.2. Identification of Amplitude-Panned Sources

In this section we describe an application of the panning index to the derivation of a new signal representation, which allows us to identify and localize the different sources in the stereo mix. The basic idea is to compute the short-time panning index $\Psi(m,k)$ and produce an energy histogram at each time frame by integrating the energy within frequency regions with the same or similar panning index value. Thus, this time-pan representation or *panogram* will indicate the distribution of energy as a function of panning index (or coefficient) and time. In this new domain it is in principle possible to identify prominent sources in the mix and their panning coefficients, or positions in the stereo image if the loudspeaker layout is known.

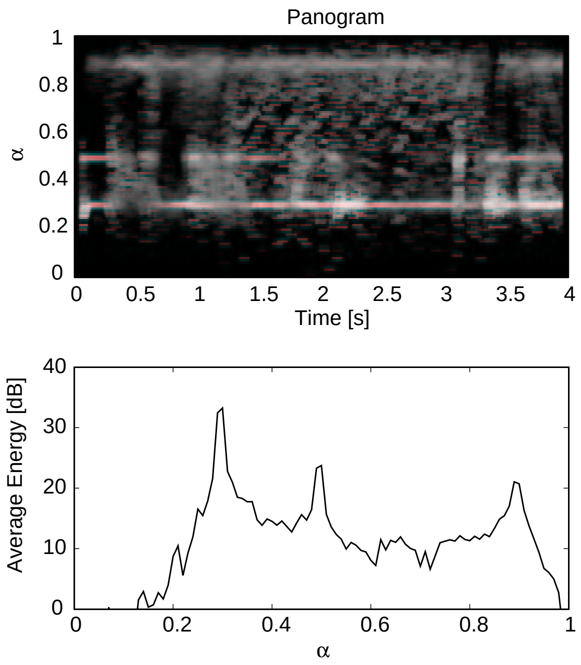

Figure 4: Top panel: Example of a panogram. The gray scale indicates the energy level, with bright regions having higher value. Bottom panel: Time average of the panogram (over 4 seconds). The peaks correspond to panning regions of high energy, where amplitude-panned signals are likely to exist.

To illustrate this technique we generated a stereo mix by amplitude-panning three sources, a voice signal $s_1(t)$, an acoustic guitar $s_2(t)$ and a trumpet $s_3(t)$ with the following panning coefficients:

$$x_1(t) = \quad 0.5s_1(t) + 0.7s_2(t) + 0.1s_3(t)$$

$$x_2(t) = \quad 0.5s_1(t) + 0.3s_2(t) + 0.9s_3(t).$$

In Figure 4 we show the panogram representation for this stereo signal. The histogram was computed by integrating the energy in both stereo signals for each panning coefficient value between 0 and 1 in 0.01 increments. The gray level indicates the average energy as a function of time (frame) and panning coefficient. Notice that this panogram shows three stripes with high energy at values of $0.3$, $0.5$ and $0.9$, which correspond to the panning coefficients of the sources in the mix. The stripes are not continuous in time due to silence intervals, which is a feature that could potentially be used for segmentation. The bottom panel of the figure shows the time average of the histogram as a function of $\alpha$, and we notice that three prominent peaks indicate the presence of the three amplitude-panned signals. The relative levels of the sources could also be obtained from this representation.

The previous example represents an ideal situation where there are no ambience components. However, ambience tends to smear the panogram representation, thus better results are obtained by reducing or weighing down the ambience components in the mix using the method in Section 4.1 prior to the computation of the panning index.

### 4.3. Source Unmix and Synthesis

Here we describe a method for extracting one or more audio streams from a two-channel signal by selecting directions in the stereo image. If multiple panned signals are present in the mix and if we assume that the signals do not overlap significantly in the time-frequency domain, then the $\Psi(m,k)$ will have different values in different time-frequency regions corresponding to the panning coefficients of the signals that dominate those regions. Thus, the signals can be separated by selecting time-frequency regions where $\Psi(m,k)$ has a given value and using these regions to synthesize a time domain signal. For example, to extract the center-panned signal(s) we find all time-frequency regions for which the panning index is zero and then synthesize a time domain function using only these components. The same procedure can be applied to signals panned to other locations by selecting different panning index values.

To avoid artifacts due to abrupt modifications of the STFT and to account for possible time-frequency overlap between signal components, we apply a narrow window centered around the panning index value corresponding to the desired panning coefficient. The width of the window is determined based on the desired trade-off between separation and distortion (a wider window will produce smoother transitions but will also allow signal components panned near zero to pass). Thus an unmixing function can be defined as a Gaussian window function, i.e.:

$$\Theta[\Psi] = \nu + (1-\nu)e^{-\frac{1}{2\xi}(\Psi-\Psi_0)^2} \qquad (13)$$

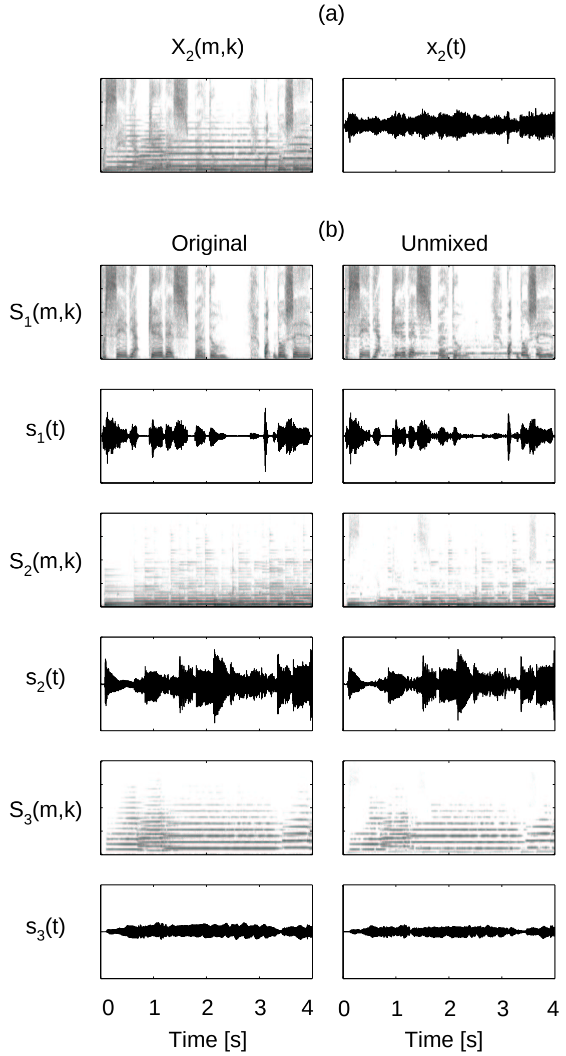

Figure 5: (a) Spectrogram and waveform of the right-channel signal. (b) Spectrograms and waveforms of the original signals (left) and unmixed signals (right). Spectrograms have a range of $0-8$ kHz, bottom to top.

where $\Psi_0$ is the desired panning index value, $\xi$ controls the width of the window, and $\nu$ is a floor value necessary to avoid setting STFT values to zero which would create spectral-subtraction-like artifacts. To unmix the desired source we simply multiply the input STFT's by (13) and add the left and right components to obtain a new STFT as:

$$S_u(m,k) = \Theta[\Psi(m,k)](X_1(m,k) + X_2(m,k)),$$

and we finally apply an inverse STFT to the new transform $S_u(m,k)$ to obtain the time domain signal. To illustrate the operation of this technique we apply the algorithm to separate the three sources in the example of Section 4.2. We applied three windows centered at

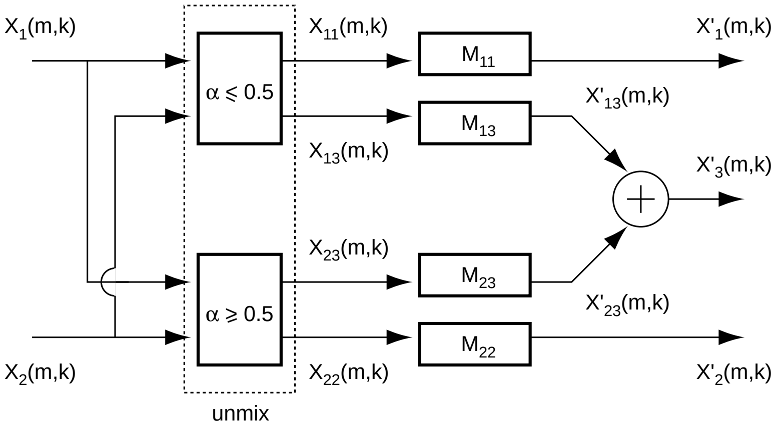

Figure 6: Two-to-three upmix system based on amplitude re-panning.

$\Psi_0 = 0, -0.27$ and $0.78$ to extract the speech, guitar and horn signals respectively. In this case we know the panning coefficients of the signals that we wish to separate, which corresponds to applications where we wish to extract or separate a signal(s) at a given location(s). Other applications may require identification of prominent sources as we described in Section 4.2.

The results of the simulation are shown in Figure 5, where side-by-side comparisons between the recovered and original waveforms and spectrograms illustrate the effectiveness of the techniques in this simplified scenario. Notice that there is some residual distortion due to the time-frequency overlap, which is inevitable. The algorithms have been tested with real audio signals and the results are equally promising.

### 4.4. Source Re-Panning

There are certain applications where we would be interested in re-panning the sources in the stereo mix without necessarily having to extract or localize them as we did in previous sections. For example, modifying arbitrarily the stage layout by altering the stereo image, or dynamically repositioning arbitrary instruments located in certain positions on the stage, etc. For this we could simply use the panning index to assign a panning value to every time-frequency component in the mix, and use a function to map these panning values into others as desired. Multiplying the STFT's by the ratio of the new and old panning values would result in the desired modification of the stereo image.

One important application of this technique is the two-to-N channel conversion, where the number of loudspeakers is larger than the number of program channels. In this application, the goal is to introduce additional loudspeakers between the stereo pair to widen the listening area [10]. A typical example is the two-to-three channel upmix necessary to deliver stereo content over a 5.1 system, where a center loudspeaker is placed symmetrically between the stereo pair. Matrix converters attempt to send the center-panned signals to the center channel to stabilize its image, but fail to improve stability of side-panned sources. In fact, the stereo image is affected adversely in many cases. We briefly describe a technique to re-pan the stereo material over three channels that overcomes some of these limitations.

The main idea is to generate two new signal pairs from the stereo recording as shown in Figure 6. The first pair $X_1'(m,k)$ and $X_{13}'(m,k)$ is played over the left and center channels, and the second $X_{23}'(m,k)$ and $X_2'(m,k)$ is played over the center and right channels respectively. The first pair will only contain sources panned to the left and center, and the second will have sources panned to the right and center. Thus the center channel will consist of a sum of signals, i.e. $X_3'(m,k) = X_{13}'(m,k) + X_{23}'(m,k)$. However, to preserve the stereo image, the signals will have to go through a previous re-panning stage where a mapping function on the panning index is used to re-pan the signals according to the original separation of the loudspeakers and the new configuration.

The separation stage can easily be accomplished with the help of the panning index, i.e. identifying STFT regions for which $\alpha \leq 0$ and $\alpha \geq 0$. This process produces four signals, $X_{11}(m,k)$ which contains the components panned to the left in the left channel, $X_{13}(m,k)$ which contains the components panned to the left and center in the right channel, $X_{23}(m,k)$ with components panned to the right and center in the left channel and $X_{22}(m,k)$ with components panned to right in the right channel. These four signals are then modified to generate the output signals $X_1'(m,k)$ , $X_2'(m,k)$ , and $X_3'(m,k)$ according to mapping functions which are derived in the following way:

If we assume a source panned such that its image corresponds to an angle $\theta_o$ with a loudspeaker separation of $\theta$

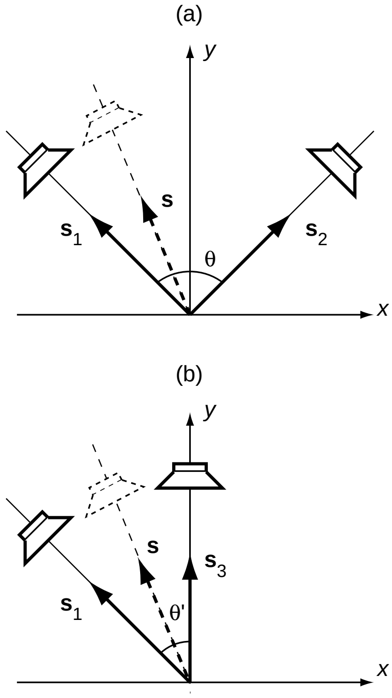

Figure 7: Coordinate system for loudspeaker layout. (a) Original layout, (b) left-front and center loudspeaker layout.

(see Figure 7), then the modification of its panning values will have to be such that the apparent angle will not be modified when the angle between loudspeakers decreases by half $\theta'$ (assuming a symmetrical layout). In two-dimensional Cartesian coordinates, the apparent location of the source $\mathbf{s} = [x, y]^T$ is determined by the panning coefficient vector $\mathbf{a} = [a_1 a_2]^T$ ($a_1 = 1 - \alpha$ and $a_2 = \alpha$), and the positions of the loudspeakers relative to the listener, which are defined by vectors $\mathbf{s}_1 = [x_1, y_1]^T$ and $\mathbf{s}_2 = [x_2, y_2]^T$. At low frequencies ($f < 700$ Hz) the apparent location is obtained by vector addition as [8]:

$$\mathbf{s} = [\mathbf{s}_1 \mathbf{s}_2]^T \mathbf{a}. \tag{14}$$

At high frequencies $(f > 700Hz)$ the apparent location of the source is determined by adding the intensity vectors generated by each loudspeaker (as opposed to amplitude vectors). The location vector is computed as:

$$\mathbf{s} = [\mathbf{s}_1 \mathbf{s}_2]^T \mathbf{p}. \tag{15}$$

where $\mathbf{p} = [a_1^2 a_2^2]^T$. Notice that there is a discrepancy in the perceived location in different frequency ranges [11]. To re-pan this source between the left and center channels would be equivalent to compensating for the reduction in angular separation between left and right loudspeakers (see Figure 7(b)). Thus, to preserve the image we need to recalculate the panning gains. If the vector for the center loudspeaker is $\mathbf{s}_3 = [x_3, y_3]^T$, then the new coefficients $\mathbf{a}^{'}$ are found by solving the following equations:

$$[\mathbf{s}_1 \mathbf{s}_2]^T \mathbf{a} = [\mathbf{s}_1 \mathbf{s}_3]^T \mathbf{a}^{'}. \tag{16}$$

at low frequencies and

$$[\mathbf{s}_1 \mathbf{s}_2]^T \mathbf{p} = [\mathbf{s}_1 \mathbf{s}_3]^T \mathbf{p}^{'}. \tag{17}$$

at high frequencies, where $\mathbf{p}^{'}$ is the new intensity vector. Thus to preserve the source image in the new loudspeaker configuration we simply multiply the signal by the ratio of the new and old coefficients. The same process can be applied if the virtual source is panned to the right hemisphere, with the only difference that the right terms in (16) and (17) will involve the right and center vectors.
The re-panning strategy outlined above can be used to modify the STFT's as shown in Figure 6, where four modification functions $M_{11}$, $M_{13}$,$M_{23}$, and $M_{22}$ are used. For example, assuming an angle $\theta = 60°$ the desired panning coefficients for each output STFT region as function of the original coefficient $\alpha$ are shown in the upper panels of Figure 8. Notice the difference between high and low frequencies. The modification functions are simply obtained by computing the ratio between the desired gains and the input gains. In the bottom panels of Figure 8 we show the corresponding modification functions. The gains and modifications for the right-center are valid for values of $\alpha = 0.5$ and are mirror images of the functions shown in Figure 8.

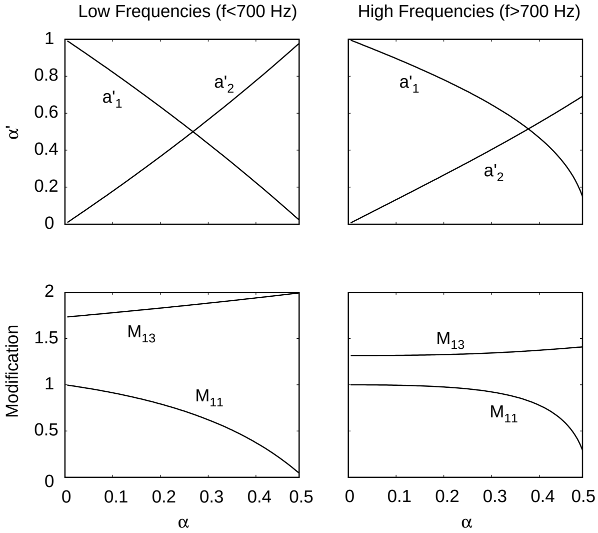

Figure 8: Target coefficients (top panel) and modification functions (bottom panel) for amplitude re-panning. Both low frequency (left) and high frequency (right) functions are shown. Loudspeaker separation is 60°.

The application described above can be generalized to include any number of loudspeakers. In essence, the technique is capable of *warping* the spatial distribution of the sources in the mix, thus it can be applied to any arbitrary loudspeaker configuration.

## 5. TWO-TO-FIVE CHANNEL UPMIX

In this section we describe the application of the various upmix algorithms to the design of a direct/ambient two-to-five channel upmix system. The idea is to extract the ambience signals from the stereo recording using the technique in Section 4.1, and use them to create the rear or surround signals. To derive the front channels we can use the unmix and re-panning techniques of Section 4.

The proposed system is shown in Figure 9. The surround signals are generated by first extracting the ambience signals and applying a de-correlating all-pass filter $G(z)$ such as [13]. The reason for doing this is that we are extracting the ambience from the front channels, thus the surround channels will be somewhat correlated with the front channels. This correlation might create undesired phantom images to the sides of the listener. To avoid delocalization due to the precedence effect and to simulate rooms of different sizes, the rear channels are delayed by some amount $D$.

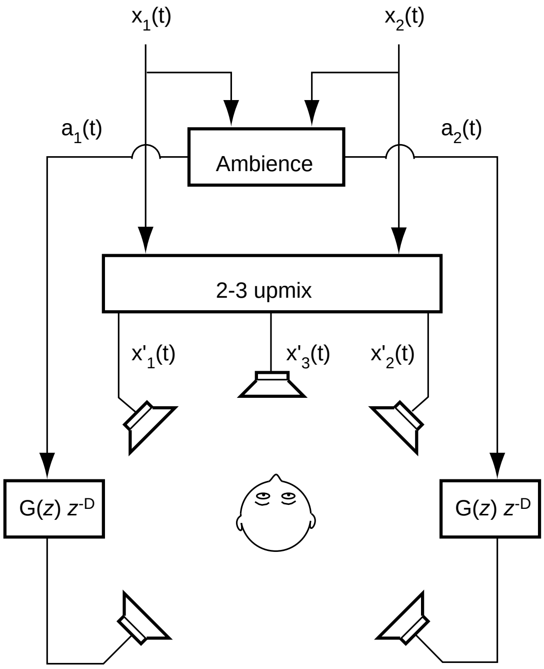

Figure 9: Block diagram of the two-to-five upmix system.

The simplest configuration to generate the front channels is to derive the center channel using the source unmix technique in Section 4.3 to extract the center-panned signal and subtracting this signal from the left and right channels. The residual signals are sent to the left-front and right-front channels. This system is capable of producing a stable center channel for off-axis listening, and it preserves the stereo image of the original recording when the listener is at the sweet spot. However, side-panned sources will still collapse if the listener moves off-axis. A system based on the re-panning technique in Section 4.4 can be designed to improve the image stability for off-center sources.

The system has been tested with a variety of audio material. While distortion is sometimes audible when the signals are played individually, the simultaneous playback of the five signals masks the distortion and creates the desired envelopment in the sound field with very high fidelity.

## 6. DISCUSSION

The techniques presented in this paper work mainly for studio mixes that use amplitude panning methods. While the performance of the ambience extraction algorithm does not degrade significantly in live mixes, the other methods based on the panning index suffer performance degradation. When sources are panned in delay (e.g. using non-coincident microphone pairs), the relationships between STFT's are not as straightforward as for the scalar-difference case found in amplitude panning. Current work is underway to derive a delay-panning index that will help us to address this problem more effectively. Another topic of further research is the perceptual assessment of the techniques. While we have conducted informal listening tests that confirm the effectiveness of our approach, there are inevitable errors introduced by the inherent signal overlap in the time-frequency plane and the STFT modifications themselves. We expect that formal evaluations will help us to understand the effects of these errors and will help us to find ways of improving the results.

## 7. CONCLUSIONS

We have presented a common frequency-domain framework to compare the signals in a stereo recording/mix. From this comparison we derived the ambience and panning indices, which are the basis of a series of techniques aimed at upmixing stereo material into a multichannel signal. We described their application in the design of a two-to-five upmix system based on the direct/ambient technique for mixing multichannel audio. While, the techniques presented are very effective, there is still room for improvement, mainly in the case of the live mix.

## ACKNOWLEDGEMENT

The unmix approach presented in this paper was inspired by the unpublished work of Jean Laroche on the extraction of voice from stereo recordings.

## REFERENCES

[1] J. Allen, D.A. Berkeley and J. Blauert, ”Multi-microphone Signal-Processing Technique to Remove Room Reverberation from Speech Signals.” *Journal of the Acoustical Society of America*, Vol. 62, No.4, pp. 912-915, October 1977.

[2] C. Avendano and J. M. Jot, ”Ambience Extraction and Synthesis from Stereo Signals for Multichannel Audio Upmix,” to appear in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP’02*, Orlando, Fl, May 2002.

[3] D.R. Begault, *3-D Sound for Virtual Reality and Multimedia*, pp. 226-229, Academic Press, Cambridge, 1994.

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

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

[6] M. A. Gerzon, ”Optimum Reproduction Matrices for Multispeaker Stereo.” *AES 90th Convention*, 1991.

[7] T. Holman, ”Mixing the Sound.” *Surround Magazine*, pp. 35-37, June 2001.

[8] 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*, Rovaniemi, Finland 1999.

[9] C. Kyriakakis and A. Mouchtaris, ”Virtual Microphones for Multichannel Audio Applications.” In Proc. *IEEE ICME 2000*, Vol. 1, pp. 11 - 14, August 2000.

[10] M. T. Miles, ”An Optimum Linear-Matrix Stereo Imaging System.” *AES 101st Convention*, 1996, preprint 4364(J-4).

[11] V. Pulkki and M. Karjalainen, ”Localization of Amplitude-Panned Virtual Sources I: Stereophonic Panning.” *Journal of the Audio Engineering Society*, Vol. 49, No. 9, pp. 739-752, September 2001.

[12] F. Rumsey, ”Controlled Subjective Assessment of Two-to-Five Channel Surround Processing Algorithms.” *Journal of the Audio Engineering Society*, Vol. 47, No. 7/8, pp. 563-582, 1999.

[13] M. Schroeder, ”An Artificial Stereophonic Effect Obtained from Single Audio Signal.” *Journal of the Audio Engineering Society*, Vol. 6, pp. 74-79, 1958.