Frequency-Domain Two- to Three-Channel Upmix for Center Channel Derivation and Speech Enhancement

Earl Vickers[1]

[1] STMicroelectronics, Santa Clara, CA 95054
earl.vickers@st.com

ABSTRACT

Two- to three-channel audio upmix can be useful in a number of contexts. Adding a front center loudspeaker provides a more stable center image and an increase in dialogue clarity. Even in the absence of a physical center speaker, the ability to derive a center channel can facilitate speech enhancement by making it possible to boost or filter the dialogue, which is usually panned to the center. Two- to three-channel upmix can also be a first step in upmixing from two to five channels. We propose a frequency-domain upmix process using a vector-based signal decomposition, including methods for improving the selectivity of the center channel extraction. A geometric interpretation of the algorithm is provided. Unlike most existing frequency-domain upmix methods, the current algorithm does not perform an explicit primary/ambient decomposition. This reduces the complexity and improves the quality of the center channel derivation.

1. INTRODUCTION

While much of the currently available audio content uses a two-channel stereo format, there are many advantages to deriving a center channel signal, whether or not a physical center loudspeaker is available.

When there are only two front speakers, the phantom center tends to collapse toward the nearest speaker, due to the precedence effect. In addition, phantom center images can suffer from timbral modifications due to comb filtering. Adding a center speaker helps anchor the dialogue in the middle of the screen, providing a more stable center image, an enlarged sweet spot and improved dialogue clarity [1][2][3][4].

Relatively few televisions come with 5.1 speaker systems, but a growing number of widescreen TVs include a built-in center speaker. Furthermore, a two- to three-channel upmix can be the first step in a two to five upmix in which the surround channels may be synthesized or derived from other signals.

Even if no physical center speaker is present, center channel derivation makes it easier to enhance the intelligibility of the dialogue, which is usually panned to the center. Once we have isolated the center channel, we can boost it in proportion to the remaining channels, helping it to stand out from competing sounds such as music or sound effects, or we can filter the derived center channel to amplify the voice frequencies.

This paper is organized as follows. Section 2 provides a brief background on audio upmix methods and describes the mathematical notation that will be used. Section 3 defines the signal model. Section 4 presents the decomposition algorithm, including a geometric interpretation. Section 5 discusses methods for improving the selectivity of the center channel derivation. Section 6 presents analysis, discussion and preliminary perceptual evaluation. Section 7 provides a summary and conclusions.

2. BACKGROUND

2.1. Existing Algorithms

As discussed by Avendano and Jot [5], there are two main categories of two- to N-channel upmix algorithms: multichannel converters and ambience generators.

Multichannel converters, which include linear ("passive") and steered ("active") matrix methods, are used to derive additional loudspeaker signals in cases where there are more speakers than input channels. These methods are typically implemented in the time domain.

Michael Gerzon, who promoted the idea of three-channel stereo in the early 1990s [1], developed passive energy-preserving matrices for upmix to three or more channels, using frequency-dependent matrix coefficients [6]. While linear matrix methods are relatively inexpensive to implement, they reduce the width of the front image. In a two- to three-channel upmix, any signal intended for the center is also played through the left and right speakers; the channel separation between left and center, for example, is only 3 dB [7].

Matrix steering methods update the matrix coefficients dynamically and provide the ability to extract and boost a dominant source [8][9]. These methods are particularly useful for content such as movie soundtracks, in which one source may be of primary interest at any given time, but the signal-dependent gain changes may cause audible side effects with music [6].

Ambience generation methods attempt to extract or simulate the ambience of a recording. The term "ambience" refers to the components of a sound that create the impression of an acoustic environment, with sound coming from all around the listener but not from a specific place. Ambience may include room reverberation as well as other spatially distributed sounds such as applause, wind or rain [5][10]. The goal of the ambience extraction is to increase the sense of envelopment, typically using the rear speakers.

Ambience generation methods may extract the natural reverberation from the audio signal (for example, by taking the difference of the left and right inputs, which attenuates centered sounds and preserves those that are weakly correlated or panned to the sides [11]), or they may add artificial reverberation [5].

Recently, a number of researchers have developed frequency-domain upmix (and downmix) techniques for spatial audio coding and enhancement [5][12][13][14][15]. These methods typically perform spatial decomposition and extract the existing ambience. Thus, these are categorized as ambience generation methods, but they can also be thought of as frequency-domain steering methods, because they dynamically change the panning of each frequency subband based on the correlation between the left and right input signals.

In [5], Avendano and Jot presented frequency domain upmix techniques based on inter-channel coherence measures, non-linear mapping functions and panning coefficients. In [12], Faller used STFT-based critical band processing to extract the ambient and direct components using least-squares estimation. Merimaa et al [13] and Goodwin and Jot [14] discussed using Principal Components Analysis (PCA) for decomposing stereo signals.

In this paper, we present a frequency-domain upmix method that does not attempt an explicit ambience extraction. Instead, the focus is on extracting a center channel, improving its channel separation and maximizing its audio quality. Note that we are only attempting *spatial decomposition*, which involves re-panning (perhaps dynamically) from two channels to three or more. We are not attempting *source separation*, which involves explicitly recovering the original source signals [16].

2.2. Decomposition Framework

Audio signals tend to be more sparse when represented in the frequency domain, which makes it easier to analyze their spatial orientation and separate their components accordingly. Therefore, our upmix algorithm uses a time-frequency analysis-synthesis framework.

Currently the short-time Fourier transform (STFT) is used [17], with the Fourier transforms being implemented using the fast Fourier transform (FFT). Other time-frequency transforms, such as the Discrete Cosine Transform, wavelets, etc., could possibly be used instead. It may also be possible to group adjacent STFT subbands together to reduce computation or simulate the critical bands of the human hearing system.

2.3. Notation and Definitions

Each STFT subband will be treated as a vector in time, as follows:

$$\vec{X}_L[b,l] = \left[x_L[b,l], x_L[b,l-1], \ldots \right]^T \qquad (1)$$

$$\vec{X}_R[b,l] = \left[x_R[b,l], x_R[b,l-1], \ldots \right]^T , \qquad (2)$$

(after [14]), where channel vectors $\vec{X}_L$ and $\vec{X}_R$ represent the left and right channels of the stereo input signal, and $x_L[b,l]$ and $x_R[b,l]$ are the (complex) STFT representations of the left and right input channels for a pair of time-frequency tiles with subband index *b* and time index *l*. Henceforth, we will generally simplify the notation by dropping the *b* and *l* indices. For the signal model, the actual (or presumed) signal components will be denoted with calligraphic symbols (for example, $\vec{\mathcal{L}}$), and our estimates (output signals) will use the normal italic symbols (e.g., $\vec{L}$).

The norm (length or absolute value) of a vector such as $\vec{X}_L$ will be shown as

$$\left\| \vec{X}_L \right\| = \sqrt{\vec{X}_L \cdot \vec{X}_L} = \sqrt{\vec{X}_L{}^H \vec{X}_L} , \qquad (3)$$

where $\| \ \|$ denotes the vector magnitude (or square root of the autocorrelation), the dot denotes the dot product, and ^{H} denotes Hermitian transposition.

All operations will be performed independently on each STFT subband. In addition, we will generally choose to simplify the algorithm by performing operations independently on each STFT time frame, without regard to past inputs. This eliminates the need for a "forgetting factor," which can cause problems with transients.

3. SIGNAL MODEL

3.1. Left, Right and Center Components

Our goal is to decompose a stereo signal by first extracting any information common to the left and right inputs and routing that to the center output; any residual audio energy will be routed to the left or right outputs as appropriate.

To facilitate this goal, we will assume that our inputs were created using the following signal model:

$$\vec{X}_L = \vec{\mathcal{L}} + \sqrt{0.5}\vec{\mathcal{C}} \qquad (4)$$

$$\vec{X}_R = \vec{\mathcal{R}} + \sqrt{0.5}\vec{\mathcal{C}} \qquad (5)$$

where the (known) input signals $\vec{X}_L$ and $\vec{X}_R$ are composed of an equal-power stereo mix of unknown left, right and center components $\vec{\mathcal{L}}$, $\vec{\mathcal{R}}$ and $\vec{\mathcal{C}}$, respectively. The outputs of the upmix algorithm will be the corresponding signal estimates: $\vec{L}$, $\vec{R}$ and $\vec{C}$. Note that we have two equations in three unknowns; therefore, additional information is needed.

3.2. Primary and Ambient Source Signals

We will assume that components $\vec{\mathcal{L}}$, $\vec{\mathcal{R}}$ and $\vec{C}$ are in turn made up of the following (sub-component) source signals, as shown in Figure 1:

$$\vec{\mathcal{L}} = g_L \vec{P} + \vec{A}_L, \quad (6)$$

$$\vec{\mathcal{R}} = g_R \vec{P} + \vec{A}_R, \text{ and} \quad (7)$$

$$\vec{C} = g_C \vec{P}, \text{ with} \quad (8)$$

$$g_L g_R = 0, \quad (9)$$

where $\vec{A}_L$ and $\vec{A}_R$ are the left and right ambient sources, and $\vec{P}$ is a primary source that is pair-wise panned anywhere between left and center *or* between right and center (inclusive), using (time- and frequency-variant) gains g_L, g_R and g_C. (If desired, these gains can be regarded as transfer functions, to allow the possibility of decomposing convolutive mixes created using non-coincident microphone pairs or delay panning.)

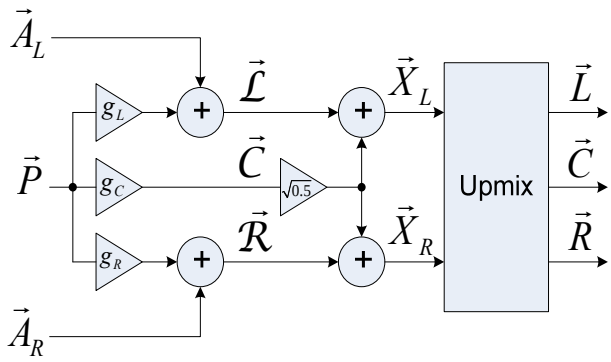

Figure 1. Signal model, with primary source $\vec{P}$, ambient sources $\vec{A}_L$ and $\vec{A}_R$, gains g_L, g_C and g_R, unknown components $\vec{\mathcal{L}}$, $\vec{C}$ and $\vec{\mathcal{R}}$, known input signals $\vec{X}_L$ and $\vec{X}_R$, and estimated (output) components $\vec{L}$, $\vec{C}$ and $\vec{R}$. We assume $g_L g_R = 0$.

We are not interested in explicitly deriving the source signals $\vec{P}$, $\vec{A}_L$ and $\vec{A}_R$ or gains g_L, g_R and g_C. The point of equations (6-9) is simply to clarify the following assumptions:

- Each stereo pair of time/frequency input tiles $\vec{X}_L$ and $\vec{X}_R$ will contain only one significant primary source signal $\vec{P}$. In practice, there may be some overlap of multiple primary sources, but this assumption has proven useful.
- If primary source $\vec{P}$ is panned somewhat left of center (i.e., between the left and center components $\vec{\mathcal{L}}$ and $\vec{C}$), it will not be present in the right component $\vec{\mathcal{R}}$, and vice versa, since gains g_L and g_R cannot both be non-zero. To the extent that inputs $\vec{X}_L$ and $\vec{X}_R$ contain a common primary source, it should be regarded as coming from center component $\vec{C}$ instead of from $\vec{\mathcal{L}}$ and $\vec{\mathcal{R}}$. This will provide a useful constraint.
- We also assume that ambient sources $\vec{A}_L$ and $\vec{A}_R$ are uncorrelated.

4. DECOMPOSITION ALGORITHM

Since the ambient sources are uncorrelated, and since components $\vec{\mathcal{L}}$ and $\vec{\mathcal{R}}$ do not contain a common primary source $\vec{P}$, due to (9), the left and right components are uncorrelated and we can regard them as orthogonal.[1] Therefore

[1] Over a large number of STFT frames, the average absolute principal phase difference between the left and right components within a given sub-band will approach 90°, since the components are uncorrelated. (The phase differences will range from 0° to 180°, and the average phase difference will be 90°.) The components will not necessarily be orthogonal for a given frame, however, since the instantaneous phase differences are essentially random.

$$\vec{\mathcal{L}} \cdot \vec{\mathcal{R}} = 0 . \quad (10)$$

From (4) and (5), we can rewrite this as

$$\left(\vec{X}_L - \sqrt{0.5}\vec{C}\right) \cdot \left(\vec{X}_R - \sqrt{0.5}\vec{C}\right) = 0 , \quad (11)$$

which yields

$$\begin{aligned} &0.5\left\|\vec{C}\right\|^2 - \sqrt{0.5}\left\|\vec{C}\right\|\left\|\vec{X}_L + \vec{X}_R\right\|\cos(\theta) \\ &+\vec{X}_L \cdot \vec{X}_R = 0 \end{aligned} , \quad (12)$$

where θ is the angle between known $\vec{X}_L + \vec{X}_R$ and unknown $\vec{C}$.

In the absence of a better estimate, it may be reasonably assumed that $\theta \cong 0°$; i.e., that the angle of center component $\vec{C}$ is roughly equal to that of the sum of the left and right input vectors [2]:

$$\angle\vec{C} \approx \angle\left(\vec{X}_L + \vec{X}_R\right). \quad (13)$$

By adding equations (4) and (5), we can see that as $\left\|\vec{\mathcal{L}} + \vec{\mathcal{R}}\right\|$ approaches zero, the angle of $\vec{X}_L + \vec{X}_R$ will approach that of $\vec{C}$, in which case our angle estimate will be quite accurate. On the other hand, the larger the magnitude of $\left\|\vec{\mathcal{L}} + \vec{\mathcal{R}}\right\|$ compared to the magnitude of $\vec{C}$, the more incorrect our center component angle estimate will be, but the less it will matter, because the magnitude of $\vec{C}$ will be comparatively small.

In practice, good results are achieved by setting angle θ to zero, which yields

$$\begin{aligned} &0.5\left\|\vec{C}\right\|^2 - \sqrt{0.5}\left\|\vec{C}\right\|\left\|\vec{X}_L + \vec{X}_R\right\| \\ &+\vec{X}_L \cdot \vec{X}_R = 0 \end{aligned} , \quad (14)$$

which is quadratic in $\left\|\vec{C}\right\|$ [18]. After using the quadratic formula, we obtain

$$\begin{aligned} &\left\|\vec{C}\right\| = \sqrt{0.5}\left\|\vec{X}_L + \vec{X}_R\right\| \\ &\pm\sqrt{0.5\left\|\vec{X}_L + \vec{X}_R\right\|^2 - 2\vec{X}_L \cdot \vec{X}_R} \end{aligned} , \quad (15)$$

which simplifies to

$$\left\|\vec{C}\right\| = \sqrt{0.5}\left(\left\|\vec{X}_L + \vec{X}_R\right\| \pm \left\|\vec{X}_L - \vec{X}_R\right\|\right). \quad (16)$$

We will choose the negative sign to achieve the following minimum-energy solution:

$$\left\|\vec{C}\right\| = \sqrt{0.5}\left(\left\|\vec{X}_L + \vec{X}_R\right\| - \left\|\vec{X}_L - \vec{X}_R\right\|\right). \quad (17)$$

Since we have assumed (equation 13) that the angle of center component $\vec{C}$ is approximately equal to that of the sum of the left and right input vectors, we can estimate $\vec{C}$ by taking a unit vector in the direction of $\vec{X}_L + \vec{X}_R$ and scaling it by the magnitude estimate $\left\|\vec{C}\right\|$ from (17):

$$\vec{C} = \frac{\left(\vec{X}_L + \vec{X}_R\right)\left\|\vec{C}\right\|}{\left\|\vec{X}_L + \vec{X}_R\right\| + \varepsilon}, \quad (18)$$

In practice, it is possible to achieve a surprisingly good decomposition by operating independently on each stereo pair of time/frequency input tiles, without regard to past inputs, so long as the frame size is on the order of 4096 or 8192 samples. There will be some leakage of ambience into the center output, but this is not objectionable.

[2] If we solve (12) to obtain the general solution

$$\begin{aligned} &\left\|\vec{C}\right\| = \sqrt{0.5}\left\|\vec{X}_L + \vec{X}_R\right\|\cos(\theta) \\ &-\sqrt{0.5\left\|\vec{X}_L + \vec{X}_R\right\|^2 \cos^2(\theta) - 2\vec{X}_L \cdot \vec{X}_R} \end{aligned}$$

and minimize this with respect to θ, we find that $\left\|\vec{C}\right\|$ reaches a minimum when $\theta = 0°$.

where ε is a very small number intended to prevent division by zero.

Finally, from (4) and (5), we obtain estimated components $\vec{L}$ and $\vec{R}$:

$$\vec{L} = \vec{X}_L - \sqrt{0.5}\vec{C} \quad (19)$$

$$\vec{R} = \vec{X}_R - \sqrt{0.5}\vec{C} \quad (20)$$

A typical set of input and output vectors are shown in Figure 2. The similarity in angle and magnitude between inputs $\vec{X}_L$ and $\vec{X}_R$ results in a strong center output $\vec{C}$. Note that estimated left and right components $\vec{L}$ and $\vec{R}$ are orthogonal by construction (equation 10).

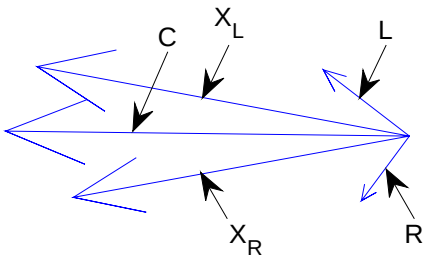

Figure 2. Typical set of left and right input vectors ($\vec{X}_L$ and $\vec{X}_R$) and left, right and center output vectors ($\vec{L}$, $\vec{R}$ and $\vec{C}$). $\vec{L}$ and $\vec{R}$ are perpendicular.

4.1. Geometric Interpretation

In equation (17), the estimated magnitude of center component $\vec{C}$ equals $\sqrt{0.5}$ times the difference between the magnitude of the sum of the left and right input vectors and the magnitude of their difference. This equation has a geometric interpretation as shown below.

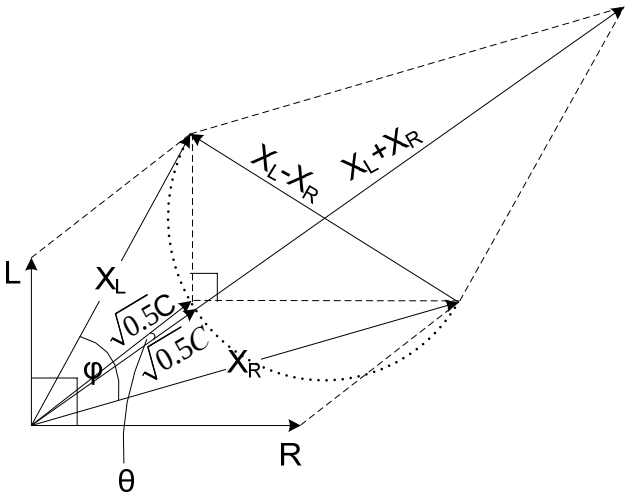

Figure 3. Geometric interpretation of the vector decomposition, depicting left and right inputs $\vec{X}_L$ and $\vec{X}_R$, components $\vec{\mathcal{L}}$, $\vec{\mathcal{R}}$ and $\sqrt{0.5}\vec{C}$, diagonal sum vector $\vec{X}_L + \vec{X}_R$, diagonal difference vector $\vec{X}_L - \vec{X}_R$, and center output $\sqrt{0.5}\vec{C}$. $\sqrt{0.5}\vec{C}$ must lie along the dotted semicircle.

In Figure 3 we see that left input $\vec{X}_L$ is a diagonal of a parallelogram that has components $\vec{\mathcal{L}}$ and $\sqrt{0.5}\vec{C}$ as two of its sides. In other words, $\vec{X}_L$ is composed of $\mathcal{L} + \sqrt{0.5}\vec{C}$, and similarly for the right channel, as given in (4) and (5). We also see that $\vec{X}_L + \vec{X}_R$ and $\vec{X}_L - \vec{X}_R$ are the diagonals of a parallelogram having two sides of length $\left\|\vec{X}_L\right\|$ and two sides of length $\left\|\vec{X}_R\right\|$. Furthermore, we see that, at least in this case, the angle of center component $\vec{C}$ is similar but not identical to that of $\vec{X}_L + \vec{X}_R$.

The dashed lines connecting $\sqrt{0.5}\vec{C}$ to $\vec{X}_L$ and $\vec{X}_R$ are orthogonal, since they are constructed to be parallel to orthogonal components $\vec{\mathcal{L}}$ and $\vec{\mathcal{R}}$, respectively. Together with the diagonal vector $\vec{X}_L - \vec{X}_R$, these two lines form a right triangle. By the Pythagorean theorem [19],

$$\left\|\vec{X}_L-\sqrt{0.5}\vec{C}\right\|^2+\left\|\vec{X}_R-\sqrt{0.5}\vec{C}\right\|^2=\left\|\vec{X}_L-\vec{X}_R\right\|^2 \quad . \qquad (21)$$

This simplifies to equation (11) and merely reiterates that the dashed lines connecting $\sqrt{0.5}\vec{C}$ to $\vec{X}_L$ and $\vec{X}_R$ are orthogonal.

From the law of cosines, $\sqrt{0.5}\vec{C}$ is constrained to be at some point along a semicircle (shown as a dotted line) of diameter $0.5\left\|\vec{X}_L-\vec{X}_R\right\|$, centered around $0.5\left(\vec{X}_L+\vec{X}_R\right)$, at the intersection of the sum and difference vectors.[3] Therefore, $\sqrt{0.5}\vec{C}$ can be visualized geometrically according to

$$\sqrt{0.5}\left\|\vec{C}\right\|=0.5\left\|\vec{X}_L+\vec{X}_R\right\|-0.5\left\|\vec{X}_L-\vec{X}_R\right\| \qquad (22)$$

(from (17)), by applying this magnitude to the direction of the sum vector. $\sqrt{0.5}\vec{C}$ is the point where the sum vector intersects the dotted semicircle.

4.2. Geometric Interpretation of Phase and Magnitude Differences

4.2.1. Phase Differences

The phase difference φ between $\vec{X}_L$ and $\vec{X}_R$ is a useful indicator of how much primary content the left and right inputs may have in common. The smaller the value of φ, the more likely that both inputs contain significant amounts of the same primary source $\vec{P}$.

Figure 4 illustrates how the phase difference φ relates to the difference between the magnitudes of diagonals $\vec{X}_L+\vec{X}_R$ and $\vec{X}_L-\vec{X}_R$ in (17). Comparing Figures 4a through 4c, we see that as φ becomes smaller, the length of sum diagonal $\vec{X}_L+\vec{X}_R$ increases in relation that of difference diagonal $\vec{X}_L-\vec{X}_R$.

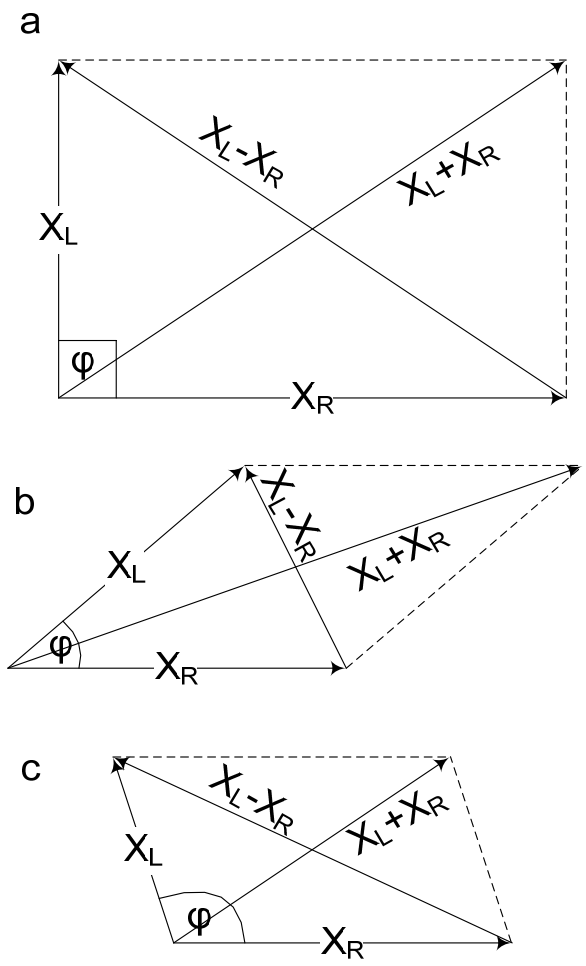

Figure 4. Graphical interpretation of various angles φ in relation to the sum and difference diagonal.

When φ < 90° (Figure 4b), the sum diagonal is larger than the difference diagonal, causing $\left\|\vec{C}\right\|$ to approach $\sqrt{2}$ times the minimum of $\left\|\vec{X}_L\right\|$ and $\left\|\vec{X}_R\right\|$ in equation (17) as φ approaches 0°. If the left and right inputs are identical, angle φ will equal 0° and $\left\|\vec{C}\right\|$ will equal $\sqrt{0.5}\left\|\vec{X}_L+\vec{X}_R\right\|=\sqrt{2}\left\|\vec{X}_L\right\|=\sqrt{2}\left\|\vec{X}_R\right\|$. In

[3] (The semicircle appears warped due to an optical illusion.)

this case, all of the input energy will be allocated to center output $\vec{C}$, as desired.[4]

When φ = 90° (Figure 4a), the two diagonals of the parallelogram ($\vec{X}_L + \vec{X}_R$ and $\vec{X}_L - \vec{X}_R$) will be of equal length, regardless of the relative levels of the left and right inputs. As a result, the magnitude of center output $\vec{C}$ will be zero (17). Therefore, if the input signals are uncorrelated, all of their energy will be sent to left and right outputs $\vec{L}$ and $\vec{R}$, and none to center output $\vec{C}$.

Finally, when φ > 90° (Figure 4c), the sum diagonal is smaller than the difference diagonal, causing $\left\|\vec{C}\right\|$ to approach $-\sqrt{2}$ times the minimum of $\left\|\vec{X}_L\right\|$ and $\left\|\vec{X}_R\right\|$ as φ approaches 180°. In other words, when inputs $\vec{X}_L$ and $\vec{X}_R$ are largely out of phase, the magnitude of center output $\vec{C}$ in (17) becomes negative. This suggests that our model did not take into account anti-phase inputs.

One option for dealing with this possibility is simply to keep the negative value of $\left\|\vec{C}\right\|$, despite the non-physical idea of a negative length. This will reverse the direction of the $\vec{C}$ vector in (18), which may cause a slight amount of energy gain (since the output vectors will be pointing in opposing directions) and create unwanted crosstalk from anti-phase left and right components into the center output. Other options are to set $\left\|\vec{C}\right\|$ to 0 whenever the estimated magnitude is negative, or to attenuate it by some arbitrary factor. These options can reduce the crosstalk but may cause "musical noise" artifacts. In practice, keeping the negative value of $\left\|\vec{C}\right\|$ seems to be the best option.

4.2.2. Magnitude Differences

The effect of input phase and magnitude differences on the magnitude of the center output $\vec{C}$ can be seen in Figure 5 for the case when $\left\|\vec{X}_L\right\| = 1$, for various values of $\left\|\vec{X}_R\right\|$ and φ, where φ is the phase difference between inputs $\vec{X}_L$ and $\vec{X}_R$.

The magnitude of the center output is partly a function of how much magnitude the two inputs have in common; according to (17), the center magnitude can be no more than $(\pm)\sqrt{2}$ times the length of the smaller of the two input vectors.

If one of the inputs, such as $\vec{X}_R$, equals zero in (17), the magnitude of $\vec{C}$ will equal 0; since there is no right channel input energy, all of the left input energy will be applied to the left output and none to the center. Note that this would not have been the case if we had selected the plus sign for the ± in equation (16).

When the left and right input magnitudes are identical (e.g., $\left\|\vec{X}_L\right\| = \left\|\vec{X}_R\right\| = 1$ in Figure 5), the magnitude of center output $\vec{C}$ varies almost linearly with the input phase difference φ, reaching a maximum when the input phases are equal.

[4] When the inputs are identical, the left and right outputs will be zero. The left output $\vec{L}$, for example, is given by:

$$\begin{aligned}
\vec{L} &= \vec{X}_L - \sqrt{0.5}\vec{C} \\
&= \vec{X}_L - \sqrt{0.5}\sqrt{0.5}\frac{\left\|\vec{X}_L + \vec{X}_R\right\|\left(\vec{X}_L + \vec{X}_R\right)}{\left\|\vec{X}_L + \vec{X}_R\right\|} \\
&= \vec{X}_L - 0.5\left(\vec{X}_L + \vec{X}_R\right) \\
&= 0.5\left(\vec{X}_L - \vec{X}_R\right) \\
&= 0,
\end{aligned}$$

since $\vec{X}_L = \vec{X}_R$.

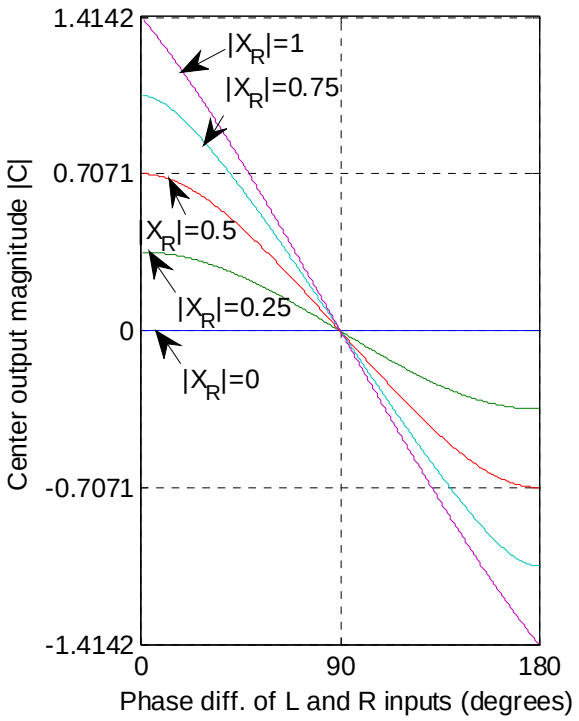

Figure 5. Magnitude $\left\|\vec{C}\right\|$ of the center output for various input phase differences φ and right input magnitudes $\left\|\vec{X}_R\right\|$, given $\left\|\vec{X}_L\right\| = 1$.

4.3. Summary: Decomposition using Phase and Magnitude Differences

Table 1 gives a simplified overview of how the input phases and magnitudes affect our signal decomposition. If the phase difference φ of left and right input vectors $\vec{X}_L$ and $\vec{X}_R$ is close to 0°, these signals are likely to include a common primary component. If their magnitudes are also similar, we will assume they represent a single primary signal that has been panned toward the center.

The more dissimilar the magnitudes of the inputs, the more we will pan the signal toward the left or right (depending on which input is larger).

With ambient sources such as reverberation, the left and right input signals are uncorrelated and generally similar in magnitude, because the sound is reflecting from all directions. The center magnitude estimate for uncorrelated (orthogonal) inputs is zero, regardless of the relative magnitudes, so orthogonal input signals are panned to the left and right outputs, $\vec{L}$ and $\vec{R}$.

Phase Difference φ	**Mag. Difference**	**Component**	**Primary or Ambient**	$\|\vec{C}\|$ **(from (17))**
0°	0	Center	Primary	$=\sqrt{2}\|X_L\|$ $=\sqrt{2}\|X_R\|$
0°	large	Center and Left/Right	Primary	$=\sqrt{2}$ min($\|X_L\|$, $\|X_R\|$)
90°	0	Left/Right	Ambient	0
90°	large	Left/Right	–	0
180°	0	Left/Right	Anti-phase Primary / Ambient	$=-\sqrt{2}\|X_L\|$ $=-\sqrt{2}\|X_R\|$
180°	large	Left/Right	Anti-phase Primary	$=-\sqrt{2}$ min($\|X_L\|$, $\|X_R\|$)

Table 1. Using phase and magnitude differences to distinguish left, right and center components, and primary and ambient subcomponents.

5. IMPROVING THE CENTER SELECTIVITY

For the purpose of enhancing dialogue clarity, we would prefer to reserve the center output mostly for primary sources that were panned directly to the center.

The current algorithm is reasonably effective at keeping the center output free of sources that were hard-panned toward the left or right. However, when primary sources such as music or sound effects are panned off-center (e.g., somewhere between left and center), a significant amount of off-center content may end up in the center output channel. This result is correct according to our original signal model, which required that any common portion of the left and right inputs should be sent to the center output. However, this behavior may cause off-center music and sound effects to mask or compete with any dialogue that may be present.

We can improve the center channel separation by using various heuristic methods.

5.1. Geometric Mean Method

The following method extends the previous decomposition by allowing us to redirect off-center sounds away from the center output, toward the side outputs. We will begin by referring to the magnitudes of the sum and difference of the left and right inputs as ς and δ, respectively:

$$\begin{aligned}\varsigma &= \left\|\vec{X}_L + \vec{X}_R\right\| \\ \delta &= \left\|\vec{X}_L - \vec{X}_R\right\|\end{aligned} \tag{23}$$

(where δ is not to be confused with the "delta function"). Recall from (17) that the estimate of the center channel's magnitude is proportional to the difference between the magnitude of the sum of the left and right inputs and the magnitude of their difference, as follows:

$$\left\|\vec{C}\right\| = \sqrt{0.5}\left(\varsigma - \delta\right). \tag{24}$$

If we can find a controlled way to increase the value of δ, making it closer to the value of ς (assuming the magnitude of the difference is less than that of the sum), this will reduce the estimated center channel magnitude for off-center sounds, causing more of the energy to be panned toward the left and right outputs instead.

Begin by dividing δ by ς, so that the resulting normalized difference magnitude, δ_1, will usually be less than 1.0 when primary sources are present:

$$\delta_1 = \frac{\delta}{\varsigma} \ . \qquad (25)$$

Next, take the square root of the normalized difference magnitude:

$$\delta_2 = \sqrt{\delta_1} \ . \qquad (26)$$

The purpose of the square root operation is to move the value closer to 1.0, increasing the difference magnitude in the usual case in which δ was less than ς.

Finally, undo the normalization from (25) by multiplying by the sum magnitude:

$$\hat{\delta} = \delta_2 \ \varsigma \ . \qquad (27)$$

Combining (25-27), we have

$$\hat{\delta} = \varsigma \sqrt{\frac{\delta}{\varsigma}} \ , \text{ or, simplifying,} \qquad (28)$$

$$\hat{\delta} = \sqrt{\delta \ \varsigma} \ . \qquad (29)$$

Thus, our modified difference magnitude $\hat{\delta}$ is the geometric mean of the magnitudes of the actual difference and sum, which moves the difference magnitude halfway (in a geometric sense) toward the sum magnitude. Substituting this for δ in (24) yields

$$\left\| \vec{C} \right\| = \sqrt{0.5} \left(\varsigma - \sqrt{\delta \ \varsigma} \right). \qquad (30)$$

This new center magnitude estimate preserves a couple of desired characteristics of (24). First, as δ approaches zero, the center magnitude approaches $\sqrt{0.5}\varsigma$; thus, when the left and right inputs are identical, the output will be sent only to the center channel (as in Section 4.2.1). Second, as δ approaches ς, the center magnitude approaches zero; this ensures that orthogonal inputs will be panned only to the left and right outputs (again see Section 4.2.1).

However, when $0 < \delta < \varsigma$ (the usual case for a primary source panned off-center), equation (30) will reduce the estimated center magnitude, sending more of the off-center energy toward the left and right outputs. This will make it easier to isolate the center channel so we can increase the gain of center-panned dialogue relative to that of any off-center music and sound effects.

Figure 6 shows the effect of input phase and magnitude differences on the magnitude of the center output $\vec{C}$ for the "geometric mean" method, when the left input has unity magnitude. Comparing this to Figure 5, we see that when the input phase difference φ is zero (suggesting that the inputs have a common primary source), the center output magnitude is attenuated as the input magnitudes become more dissimilar. In other words, off-center sources will be panned less to the center output and more to the left and right sides, as desired.

Recall from Section 4.2.1 that when the magnitude of the difference of the inputs was greater than the magnitude of their sum ($\delta > \varsigma$), the resulting center magnitude estimate was negative. In Figure 6, we see that with the geometric mean method, anti-phase inputs (identical magnitudes and 180° phase difference) result in a center output magnitude of zero, instead of a negative value; this is because ς becomes zero in equation (30). Other magnitude and phase differences can still result in negative center magnitude estimates, but the negative center outputs are attenuated compared to those in the original (Figure 5).

Finally, observe from Figure 6 that when the input magnitudes are the same ($\left\| X_L \right\| = \left\| X_R \right\| = 1$), the center output magnitude drops off much more rapidly with increases in the input phase difference φ than was the case in the previous figure. This could help keep

unwanted ambient sources (having similar magnitudes and dissimilar phases) out of the center output channel.

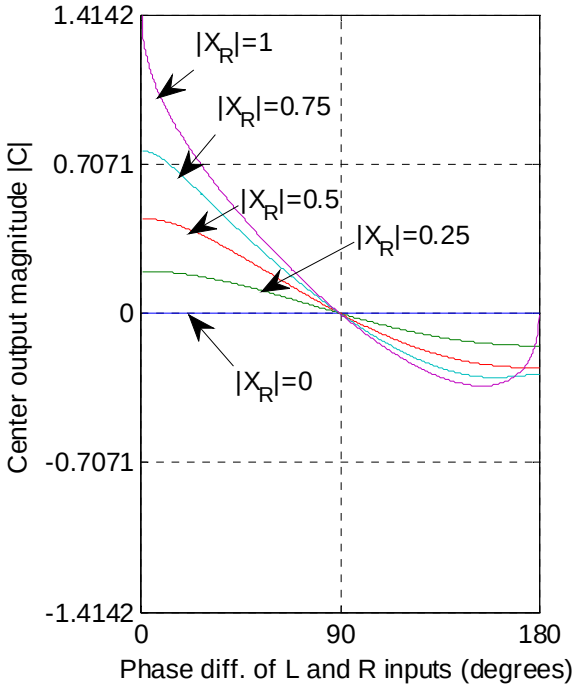

Figure 6. Magnitude $\left\|\vec{C}\right\|$ of the center output for various input phase differences φ and right input magnitudes $\left\|\vec{X}_R\right\|$, for the geometric mean method, given $\left\|\vec{X}_L\right\| = 1$.

For certain types of source signals (such as wide-band wind or water sounds), the geometric mean method can result in slight "musical noise" artifacts [5]. If desired, we can minimize unwanted effects by replacing (29) with the following equation:

$$\hat{\delta} = \sqrt{\delta\left((1-k)\delta + k\varsigma\right)}\ , \qquad (31)$$

where k is a parameter between zero and one, inclusive. The k parameter controls the extent to which the geometric mean method is applied. When $k = 0$, $\hat{\delta} = \delta$, yielding the original method; when $k = 1$, $\hat{\delta} = \sqrt{\delta\ \varsigma}$, as in (29), applying the full geometric mean method. When $0 < k < 1$, an intermediate amount of modification is applied, providing a way to achieve additional center channel selectivity without obvious artifacts. Substituting (31) for δ in (24) yields

$$\left\|\vec{C}\right\| = \sqrt{0.5}\left(\varsigma - \sqrt{\delta\left((1-k)\delta + k\varsigma\right)}\right). \qquad (32)$$

The geometric mean modification improves the isolation of the center channel, though it violates our original assumption that any signal common to the left and right inputs should be panned to the center. As a result, the left and right outputs, $\vec{L}$ and $\vec{R}$, will no longer be orthogonal after performing this modification.

5.2. Magnitude Similarity Method

This section presents an alternate method, based on magnitude similarity, of improving the center selectivity by panning off-center content toward the side speakers, as follows:

$$m = \frac{\min\left(\left\|\vec{X}_L\right\|, \left\|\vec{X}_R\right\|\right)}{\max\left(\left\|\vec{X}_L\right\|, \left\|\vec{X}_R\right\|, \varepsilon\right)}, \text{ and} \qquad (33)$$

$$\left\|\vec{C}\right\| = m\left\|\vec{C}\right\|, \qquad (34)$$

where m is a measure of similarity between the magnitudes of the left and right inputs. Equation (33) is equivalent to the following equation,

$$m = 1 - \frac{\left\|\ \left\|\vec{X}_L\right\| - \left\|\vec{X}_R\right\|\ \right\|}{\max\left(\left\|\vec{X}_L\right\|, \left\|\vec{X}_R\right\|, \varepsilon\right)}, \qquad (35)$$

except in the case where both input magnitudes are zero (in which case the value of m is irrelevant). In either (33) or (35), m equals one when the inputs have identical non-zero magnitudes (i.e., maximum magnitude similarity); m equals zero if exactly one of the inputs has zero magnitude; and $0 < m < 1$ when the input magnitudes are non-zero and non-identical.

When the inputs are panned off-center (i.e., the left and right signals have different magnitudes), the center output magnitude is attenuated with the magnitude similarity method, as may be seen by comparing Figure 7 to Figure 5.

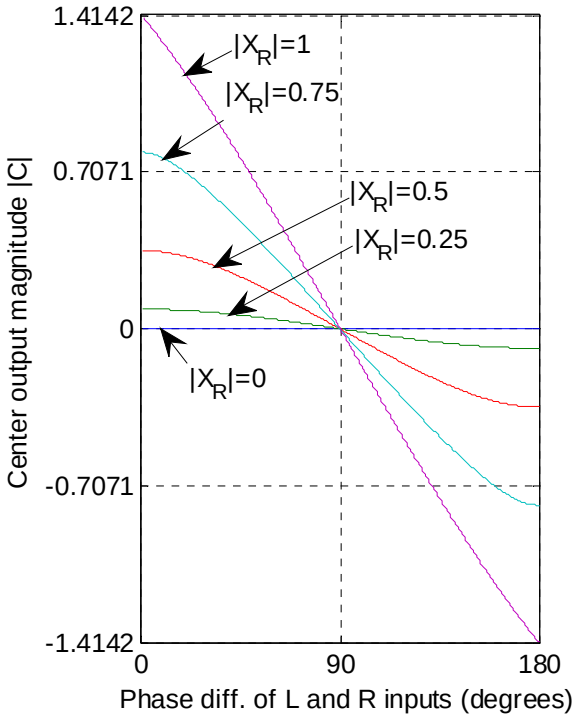

Figure 7. Magnitude $\left\|\vec{C}\right\|$ of the center output for various input phase differences φ and right input magnitudes $\left\|\vec{X}_R\right\|$, for the magnitude similarity method, given $\left\|\vec{X}_L\right\| = 1$.

In order to limit the "musical noise" artifact, it may be useful to limit m to a range such as [0.1, 0.9]. Additional center channel selectivity can be achieved by raising m to a power greater than one, such as 2.0; reduced selectivity (and presumably reduced artifacts) can be achieved by raising m to a power less than one.

If desired, the magnitude similarity m can be smoothed as follows,

$$\hat{m} = \sin\left(\frac{\pi}{2} m\right), \tag{36}$$

to remove slope discontinuities from the similarity function.

6. ANALYSIS AND DISCUSSION

6.1. Channel Separation

Figure 8 illustrates the channel separation, using the first ninety seconds of the song "Stairway to Heaven." In the stereo input (seen in the top two graphs), the acoustic guitar is panned to the left, recorders are panned to the right, and the voice is panned to the center. In addition, there is a fair amount of reverberation.

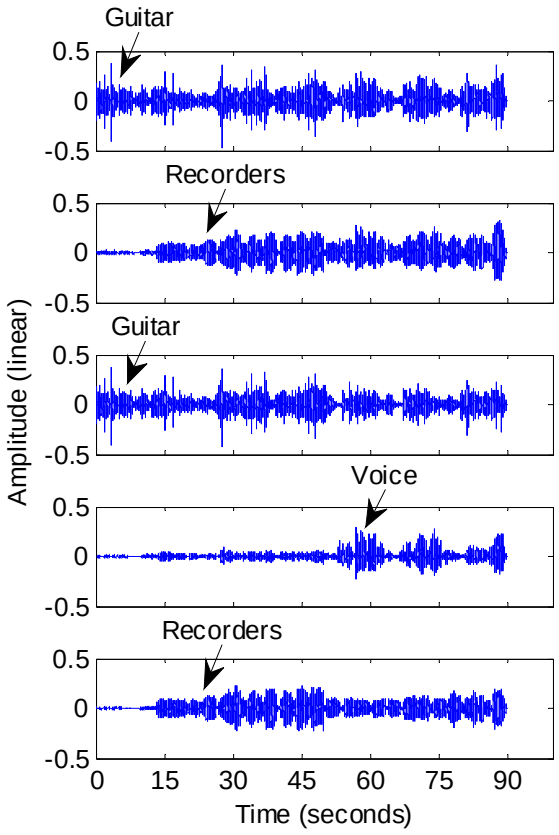

Figure 8. Channel separation. Top row: left input (guitar and voice); second row: right input (recorders and voice); third row: left output (guitar); fourth row: center output (voice); bottom row: right output (recorders).

We can see that very little of the acoustic guitar input is present in the center and right output channels (the bottom two graphs). The center output (fourth row) has some reverberation and/or crosstalk, but the onset of the voice is much more apparent than would be seen, for example, by summing the left and right inputs.

As mentioned in section 2.1, time-domain matrix upmix methods typically have only 3 dB of separation between, for example, the left and center output channels. With the current upmix method, Figure 9 shows that a signal panned to hard left has no center output gain, and a signal panned to the center has no left or right output gain. Therefore, the channel separation, as defined in [7], is infinite (assuming no inter-source interference or reverberation) for sources panned to hard left, hard right or center.

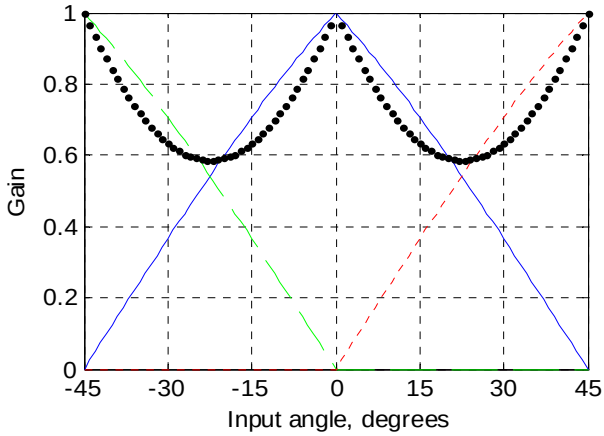

Figure 9. Panning gains and power gain: left output gain (light dashed line); center output gain (solid line); right output gain (dotted line); power gain (heavy dotted line). (The diagonal lines representing the left, right and center gains are not exactly linear.)

Figure 10 shows the center channel isolation (defined here as $\|C\| / \max(\|L\|, \|R\|, eps)$, expressed in dB), for sources panned to various directions. The solid line represents the current method, while the dashed line represents time-domain matrix methods.

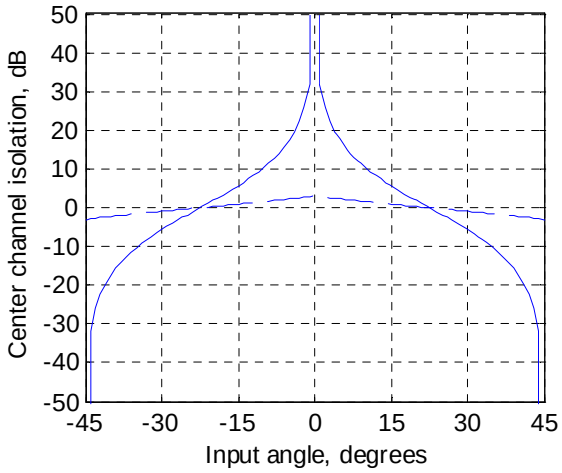

Figure 10. Center channel isolation for current upmix method (solid line) and time-domain matrix upmix (dashed line), as a function of the panning angle.

6.2. Power Gain and Panning Gain of Sources Panned in Various Directions

In Figure 9, the heavy dotted line shows that the current algorithm has unity power gain for inputs panned to hard-left, hard-right, and center. (This would not have been true if we had used other constants instead of $\sqrt{0.5}$ in (4) and (5).)

The current algorithm is not energy preserving, however, because it has approximately 2.3 dB of power loss around ±23°.

6.2.1. Energy Normalization

Power complementarity is considered a desirable property, because it guarantees a flat total radiated power response. If a power-complementary (energy-preserving) version of the algorithm is desired (e.g., for center channel derivation without speech enhancement), it can be obtained simply by normalizing each output time-frequency tile by the quotient, q, of the corresponding input and output energies, as follows:

$$q = \frac{\sqrt{\vec{X}_L{}^H \vec{X}_L + \vec{X}_R{}^H \vec{X}_R}}{\sqrt{L^H L + R^H R + C^H C + \varepsilon}}, \quad (37)$$

$$\vec{L} = q\vec{L}, \quad (38)$$

$$\vec{R} = q\vec{R}, \text{ and} \quad (39)$$

$$\vec{C} = q\vec{C}. \quad (40)$$

This normalization will not affect the perceived panning directions, because the same gain is applied to each component. The resulting (flat) power gain is shown by the heavy dotted line at the top of Figure 11.

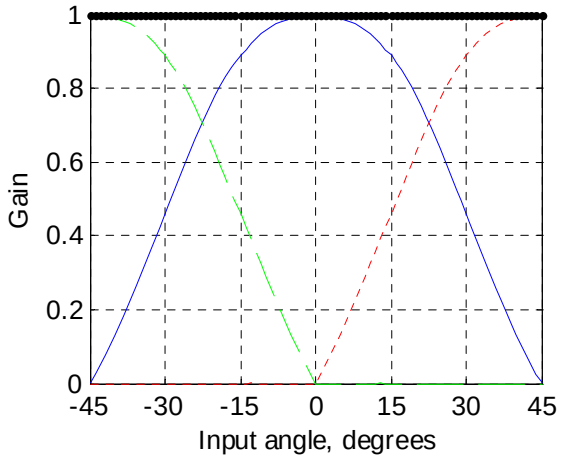

Figure 11. Panning gains and power gain for the energy-normalized version of the algorithm: left output gain (light dashed line); center output gain (solid line); right output gain (dotted line); power gain (heavy dotted line).

6.2.2. Gains for the Geometric Mean Method

Figure 12 shows the gains for the energy-normalized version of the algorithm, as in Figure 11, but using the "geometric mean" method to pan off-center sources away from the center and toward both side channels. Notice that the center channel panning gain (solid line) is more selective than its counterpart in Figure 11.

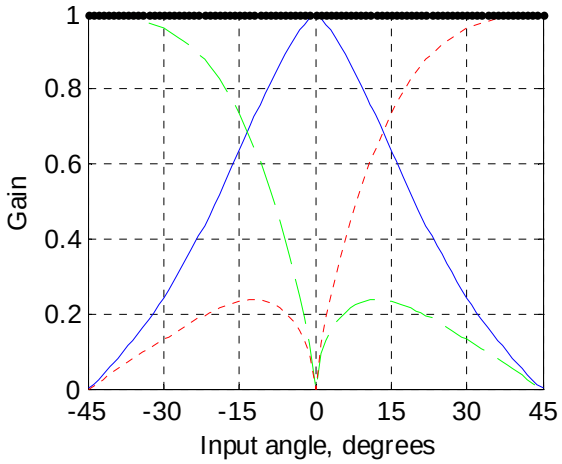

Figure 12. Panning gains and power gain for the energy-normalized version of the algorithm, using the geometric mean method to redirect some of the off-center audio toward the left and right outputs: left output gain (light dashed line); center output gain (solid line); right output gain (dotted line); power gain (heavy dotted line).

6.3. Effect on Perceived Source Directions and Width

Ideally, we would like to preserve the perceived source directions and width of the original signal. The overall perceived width is partly a function of the apparent position of each panned source, and partly a function of the overall center vs. side channel energies. We will discuss each of these factors separately.

6.3.1. Apparent Source Directions

If we pan a primary input source in various directions and upmix to three channels, the current algorithm preserves the apparent source direction of the original two-channel mix according to the tangent law [21], as seen in Figure 13.

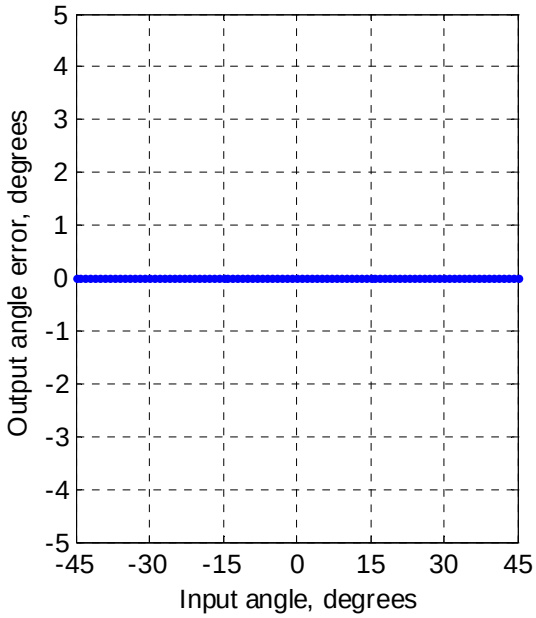

Figure 13. Zero panning error (heavy dotted line) with current upmix algorithm.

This can be shown as follows, assuming that the center speaker is positioned at 90° (directly in front) and the left and right speakers are positioned at 45° to either side. First, define unit vectors in the left, right and center speaker directions, as follows

$$\begin{aligned} U_L &= \sqrt{0.5}\left(-1+i\right) \\ U_R &= \sqrt{0.5}\left(1+i\right) \quad , \\ U_C &= i \end{aligned} \tag{41}$$

where $i = \sqrt{-1}$. Next, apply the magnitudes of our left, right and center output signals to the corresponding speaker direction unit vectors, and take the sum, S, of the resulting speaker vectors:

$$S = \left\|\vec{L}\right\| U_L + \left\|\vec{R}\right\| U_R + \left\|\vec{C}\right\| U_C \ . \tag{42}$$

Assuming our original input and output vectors all have the same phase, i.e.,

$$\angle\vec{L} = \angle\vec{R} = \angle\vec{C} = \angle\vec{X}_L = \angle\vec{X}_R \ , \tag{43}$$

since we are dealing with a single primary source, we can combine equations (19), (20), (24) and (42) as follows:

$$\begin{aligned} S &= \left(\left\|\vec{X}_L\right\| - 0.5(\varsigma - \delta)\right)U_L \\ &+ \left(\left\|\vec{X}_R\right\| - 0.5(\varsigma - \delta)\right)U_R \\ &+ \sqrt{0.5}\,(\varsigma - \delta)U_C \end{aligned} \quad . \qquad (44)$$

This simplifies to

$$S = \left\|\vec{X}_L\right\|U_L + \left\|\vec{X}_R\right\|U_R \ . \qquad (45)$$

Taking the angle of both sides, we have

$$\angle S = \angle\left(\left\|\vec{X}_L\right\|U_L + \left\|\vec{X}_R\right\|U_R\right) . \qquad (46)$$

Therefore, the apparent angle of the sum of the left, right and center speaker vectors equals the apparent angle of the left and right input signals, applied to speakers at 90° ± 45°. (These speaker vectors should not be confused with the input and output signal vectors, where the angles were phase angles, not speaker directions.)

The preservation of the apparent angles is illustrated in Figure 14, which shows that the vector sum of left and right inputs with magnitudes $\left\|\vec{X}_L\right\|$ and $\left\|\vec{X}_R\right\|$ and spatial directions of 135° and 45° equals the vector sum of left and center outputs with magnitudes $\left\|\vec{L}\right\|$ and $\left\|\vec{C}\right\|$ and directions of 135° and 90°, respectively. (The right output $\vec{R}$ equals zero since any energy common to $\vec{X}_L$ and $\vec{X}_R$ ends up in $\vec{C}$.)

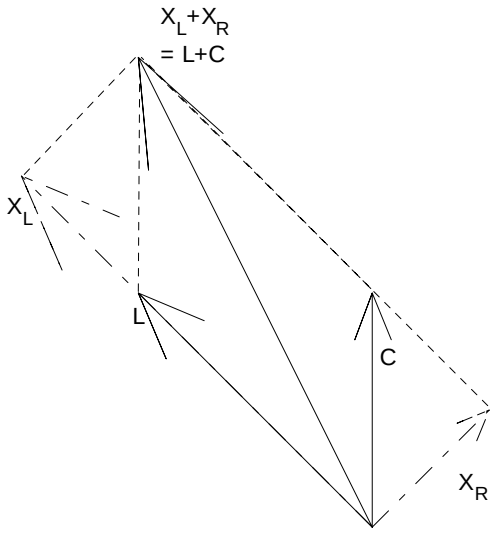

Figure 14. Preservation of apparent source direction. This example shows inputs $X_L = 3\left(\sqrt{0.5}\,(-1+i)\right)$ and $X_R = 1\left(\sqrt{0.5}\,(1+i)\right)$ (dash-dotted arrows) and outputs $L = 2\left(\sqrt{0.5}(-1+i)\right)$ and $C = 2i\sqrt{0.5}$ (solid arrows). The sum of the inputs, $X_L + X_R$, equals the sum of the outputs, $L + C = 2\sqrt{0.5}\,(-1+2i)$ (solid arrow). Dotted lines indicate the vector addition.

Thus, this method preserves the apparent position of each amplitude-panned source. (This would not have been the case if we had derived the algorithm from a signal model that used other constants, such as 0.5 or 1.0, instead of $\sqrt{0.5}$ in equations (4) and (5).)

The modified versions of the algorithm, using the geometric mean, magnitude similarity and energy normalization methods, are also direction-preserving.

6.3.2. Weighting of Center vs. Sides

Even though the original algorithm preserves the apparent direction of each source, it does not preserve the energy in each direction; the off-center power loss was illustrated in Figure 9. As a result, the algorithm does not necessarily preserve the perceived width of the overall signal. However, the energy-normalized method in Section 6.2.1 does preserve the perceived width.

If the center output channel is amplified – for example, in an effect to enhance dialogue intelligibility – the additional center weighting will modify the apparent source directions and reduce the perceived stereo width. Boosting the relative level of the center channel lowers the "second moment" (moment of inertia) of the energy distribution. Even if all of the individual sources were panned in the correct directions, comparatively little energy would arrive from the sides, leading to an impression of a narrowed overall width.

It may be possible to compensate for the width reduction by applying "stereo widening" techniques such as those discussed in [20], but some of these techniques may boost the side energy relative to the center energy, thus negating some of the dialogue clarity benefit of our center boost.

6.4. Using 2-to-3 Channel Upmix for Voice Enhancement

As mentioned, in movies and related content, the dialogue is usually panned to the center. Once we have performed the two- to three-channel upmix, we can enhance the voice by applying an amplitude gain to the extracted center channel (after deriving $\vec{L}$ and $\vec{R}$).

We can also enhance dialogue intelligibility by performing filtering to pass the voice frequencies (approximately 100-8000 Hz) in the center channel and attenuate other frequencies. The filtering can be applied to the time-domain output, but it may be more efficient to apply the filtering directly in the STFT domain, taking care to minimize any time aliasing by smoothing the gain changes from one subband to the next.

For example, for STFT bins below a low voice cutoff frequency f_L (e.g., 150 Hz), we could apply a frequency-dependent gain $g_V(b)$ as follows:

$$g_V(b) = 10^{\frac{G(b)}{20}}, \text{ where} \tag{47}$$

$$G(b) = \frac{s_V \log\left(\frac{f(b)}{f_L}\right)}{\log(2)}, \text{ and} \tag{48}$$

$$f(b) = \frac{b f_S}{N}, \tag{49}$$

where b is the bin index for bins below low cutoff bin $b_L = floor(f_L N / f_S)$, $G(b)$ is the gain of bin b expressed in dB, N is the FFT size, f_S is the sampling rate in Hz, and s_V is the desired filter rolloff (e.g., 12 dB/octave). (The equations will be similar for rolloffs above a high cutoff frequency, but with a negative value of s_V.)

Instead of simply attenuating any non-voice frequencies in the center output, we can instead redirect those frequencies to the side channels by applying the gains g_V to our center magnitude estimate $\left\|\vec{C}\right\|$:

$$\left\|\vec{C}[b,l]\right\| = g_V(b)\left\|\vec{C}[b,l]\right\|. \tag{50}$$

The reduction in center channel gain at the non-voice frequencies will result in an increase in left and right output gains at those frequencies, as shown in Figure 15, due to equations (19-20). After deriving the left and right output signals, we can amplify the center channel output if desired, to reduce masking of the voice by left and right outputs in the vocal frequency range. A variety of advanced speech detection and enhancement methods can also be applied to the derived center channel [22].

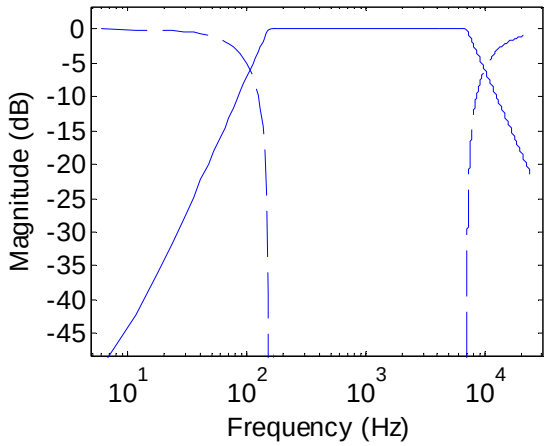

Figure 15. Redirecting non-voice center channel frequencies to the side outputs. The center output magnitude response (shown with 3 dB attenuation) is represented by the solid line, and the left (and right) side response by the dashed lines. In this example, the low frequency cutoff is 150 Hz., the high cutoff is 7 kHz, and the filter slope is -12 dB per octave.

6.5. Obtaining Additional Front Outputs

For multi-speaker systems such as television "soundbars," it may be useful to derive five or more front channels from a two-channel input. We can extract additional front channels by performing the algorithm repeatedly on adjacent pairs of output signals.

We will continue to use the assumption that any signal common to two speakers should be sent to the new, in-between speaker. An upmix from two to five front channels is performed as shown in Figure 16.

1. Decompose inputs $\vec{X}_L$ and $\vec{X}_R$ into outputs $\vec{L}$, $\vec{C}$ and $\vec{R}$ using (17-20), as before.

2. Then treat outputs $\vec{L}$ and $\vec{C}$ as inputs $\vec{X}_L$ and $\vec{X}_R$, and decompose them into ("left," "center," and "right") outputs $\vec{Y}_1$, $\vec{Y}_2$ and $\vec{Y}_{3a}$ using (17-20).

3. Treat outputs $\vec{C}$ and $\vec{R}$ (from step 1) as inputs $\vec{X}_L$ and $\vec{X}_R$, and decompose them into ("left," "center," and "right") outputs $\vec{Y}_{3b}$, $\vec{Y}_4$ and $\vec{Y}_5$ using (17-20).

4. Set $\vec{Y}_3 = 0.5\left(\vec{Y}_{3a} + \vec{Y}_{3b}\right)$.

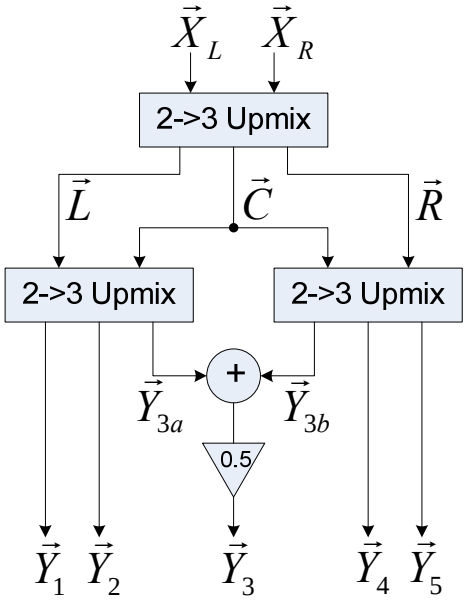

Figure 16. Block diagram of a two- to five-channel upmix comprising three two- to three-channel upmixes.

The resulting outputs, from left to right, are $\vec{Y}_1$, $\vec{Y}_2$, $\vec{Y}_3$, $\vec{Y}_4$, and $\vec{Y}_5$ (left, left-center, center, right-center, and right). The five-channel output from a two-channel equal-power sine sweep is illustrated in Figure 17.

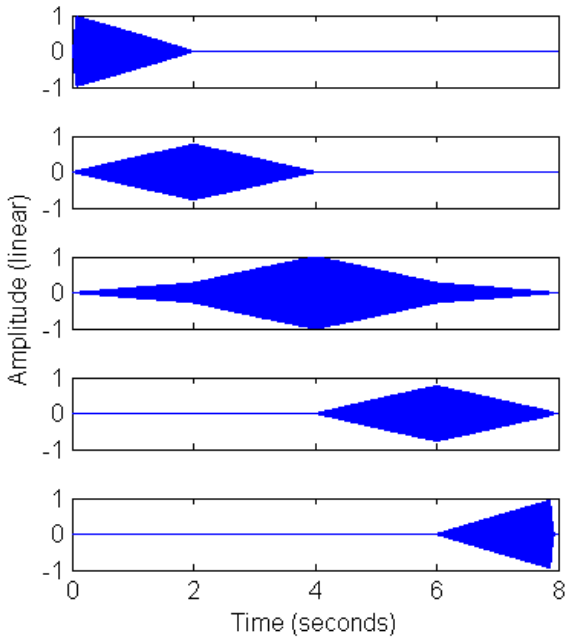

Figure 17. Sine wave with equal-power panning from left to right, upmixed from two to five front channels.

A playback system with multiple front speakers, such as a soundbar, may suffer from comb filtering or phase cancellation issues. The above method minimizes this problem because most of the inter-speaker correlation involves speakers that are immediately adjacent; since the adjacent speakers are relatively close together, any phase cancellations are likely to be in the mid- to high-frequency range. The decorrelation method presented in [2] may be used to address these phase cancellations.

6.6. Ambience Extraction

In typical stereo recordings, the left and right channels usually have similar ambience levels. Merimaa et al [13] discuss an "equal levels" decomposition method based on the assumption that the left and right ambience levels are equal.

The present method does not explicitly extract the ambience or require the left and right channels to have equal ambience levels. However, by selecting the angle of estimated center component $\vec{C}$ to equal that of the sum of the left and right input vectors (13), the algorithm automatically avoids grossly unequal ambience levels.

After our two- to three-channel upmix, any ambience will be contained primarily in the left and right output channels, since the center output consists mostly of signals that were common between the left and right inputs. If desired, we can then extract left and right ambience (surround) channels from the left and right outputs.

To the extent that a given pair of left and right output vectors has similar magnitudes, the vectors probably consist mostly of ambience, since a primary source present in both the left and right inputs would have been sent to the center output instead. Therefore, we can extract left and right surround signals from the left and right outputs using our magnitude similarity measure, as follows:

$$m = \frac{\min\left(\left\|\vec{L}\right\|, \left\|\vec{R}\right\|\right)}{\max\left(\left\|\vec{L}\right\|, \left\|\vec{R}\right\|, \varepsilon\right)}, \tag{51}$$

$$\vec{L}_S = m\vec{L}, \tag{52}$$

$$\vec{R}_S = m\vec{R}, \tag{53}$$

where m is a measure of similarity between the magnitudes of the left and right outputs, and L_S and R_S are the left and right surround outputs, respectively.[5] After extracting the left and right surround channels, we subtract them from the left and right outputs, respectively, to get the final left and right output signals:

$$\vec{L} = \vec{L} - \vec{L}_S, \text{ and} \tag{54}$$

$$\vec{R} = \vec{R} - \vec{R}_S. \tag{55}$$

As before, a sine function can be used to remove slope discontinuities from the magnitude similarity function:

[5] Note that m in (50) is based on the magnitudes of the left and right *output* vectors, unlike the magnitude similarity function in (33), which was based on the magnitudes of the left and right *input* vectors.

$$\hat{m} = \sin\left(\frac{\pi}{2} m\right). \quad (56)$$

As the difference between the left and right output magnitudes approaches zero, *m* will approach one, signifying that the left and right output channels consist primarily of ambience; as a result, a portion of the left and right outputs will be redirected to the corresponding surround channels. If the left and right output magnitudes are very different (e.g., if one of them is zero), *m* will approach zero, and none of the left and right output energy will be redirected to the surround channels.

6.7. Perceptual Evaluation

6.7.1. Two-Channel Downmix

A common usage scenario will be to upmix to three channels, boost or filter the center channel for speech enhancement, and downmix back to two channels for systems having two loudspeakers. As part of our perceptual evaluation, we would like to know that, in the absence of center channel speech enhancement, the resulting downmix sounds similar to the original signal.

When mixed back to two channels using an equal-power mixing matrix, the result sounds virtually identical to the input signal. If energy normalization is used (Section 6.2.1), the result preserves the apparent width of the input signal as well as the relative energies of sources panned to different directions.

The downmix to two channels can be done in the frequency domain, eliminating the need to perform inverse FFTs on the center channel.

6.7.2. Artifacts

Frequency-domain audio processes, such as encoding and enhancement, may produce certain artifacts, often described as “musical noise” or “watery sound.”

Such problems are rarely apparent when all of the outputs are played simultaneously at similar levels. The algorithm linearly decomposes each stereo pair of time-frequency input tiles and re-pans each component to one or more of the output channels, so any artifacts in one output channel tend to be “filled in” or masked by complementary information in the other channels.

However, artifacts can occasionally be heard with specific types of material if a single output channel, such as the center channel, is boosted too far or heard in isolation. For example, when listening to the center output by itself, slow piano arpeggios may sound a bit “mushy” – not quite as crisp and distinct. However, the output mix is generally acceptable with a center boost of six to twelve dB relative to the left and right output channels.

With certain types of source material, artifacts can become slightly more noticeable when the “geometric mean” (Section 5.1) or “magnitude similarity” (Section 5.2) modifications are used. However, results are usually quite acceptable.

6.7.3. Preliminary Listening Tests

The algorithm has been tested using various problematic audio signals, including solo piano, ocean sounds, and music and voice recordings. The energy-normalized version of the algorithm was compared informally to a couple of third-party, apparently frequency-domain, upmix algorithms. Each of the third-party algorithms was designed to extract two rear (ambience) channels in addition to left, right and center, and neither was intended for the purpose of boosting the center channel.

One of the third-party algorithms displayed excellent center channel separation, but when the center channel was heard by itself (not the intended usage), significant “watery sound” or “musical noise” artifacts were heard. (In its intended usage with all loudspeakers active at similar levels, these artifacts were not noticed.) The other third-party algorithm did not have obvious center channel artifacts, but it had significantly less center channel separation.

Overall, the current algorithm seems relatively robust and effective, possibly because it is less ambitious in scope than ambience-extraction methods, since (with the exception of the extension in Section 6.6) it does not attempt to upmix the input into center, side and surround components. The lack of obvious center channel artifacts is particularly important for our goal of boosting the center channel to enhance dialogue clarity.

It is possible that when multiple stages of signal decomposition are performed, the outputs of later stages may suffer in quality compared to the earlier outputs. If this is true, then for speech enhancement it may be

advantageous to extract the center channel before extracting the side and surround channels.

7. SUMMARY AND CONCLUSIONS

We have presented an algorithm for upmixing two-channel stereo to a three-channel format, deriving a center channel for use with a physical center speaker or to facilitate speech enhancement. The flow chart in Figure 18 summarizes the basic algorithm.

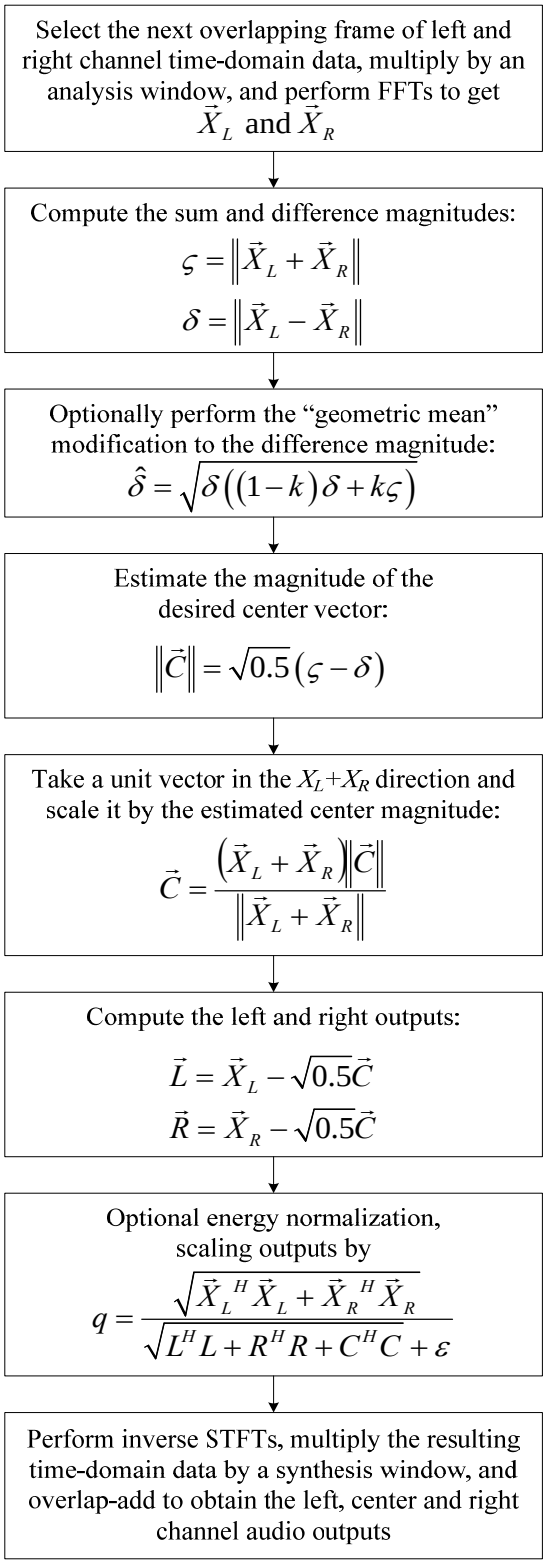

Figure 18. Flow chart of the algorithm.

The algorithm can also be used as a first step in upmixing to five or more front channels (such as for soundbar speaker systems) or to front and surround channels (such as for 5.1 systems). In addition, it can perform center channel vocal removal for karaoke.

In a couple of earlier studies, perceptual evaluations have shown that listeners often prefer the original stereo signal to the outputs of various upmix algorithms, despite the increased spatial separation [23][24]. The current algorithm is intended to address some of the limitations of previous techniques. Formal listening tests are needed to help validate the usefulness of this algorithm with a wide range of source materials.

8. ACKNOWLEDGEMENTS

This work was supported by STMicroelectronics, Inc.

9. REFERENCES

[1] Francis Rumsey, *Spatial Audio*, Oxford, UK, Focal Press, 2001.

[2] Earl Vickers, "Fixing the Phantom Center: Diffusing Acoustical Crosstalk," presented at the AES 127th Convention, New York, NY, 2009 October 9-12.

[3] Tomlinson Holman, "New Factors in Sound for Cinema and Television," *J. Audio Eng. Soc.*, vol. 39, no. 7/8, pp. 529-539, 1991 July/August.

[4] Ben Shirley, Paul Kendrick, and Claire Churchill, "The Effect of Stereo Crosstalk on Intelligibility: Comparison of a Phantom Stereo Image and a Central Loudspeaker Source," *J. Audio Eng. Soc.*, vol. 55, no. 10, pp. 852-863, 2007 October.

[5] Carlos Avendano and Jean-Marc Jot, "A Frequency-Domain Approach to Multichannel Upmix," *J. Audio Eng. Soc.*, vol. 52, No. 7/8, 2004 July/August.

[6] Michael Gerzon, "Optimum Reproduction Matrices for Multispeaker Stereo," *J. Audio Eng. Soc.*, vol. 40, pp. 571-589, 1992 July/August.

[7] David Griesinger, "Multichannel Matrix Surround Decoders for Two-Eared Listeners," presented at the AES 101st Convention, Los Angeles, CA, paper 4402, 1996 November 8-11.

[8] Kenneth Gundry, "A New Active Matrix Decoder for Surround Sound," presented at the AES 19th Int. Conf. on Surround Sound – Techniques, Technology, and Perception, Schloss Elmau, Germany, paper 1905, 2001 June.

[9] Roy Irwan and Ronald Aarts, "Two-to-Five Channel Sound Processing," *J. Audio Eng. Soc.*, vol. 50, pp. 914-926, 2002 November.

[10] Andreas Walther, Christian Uhle, and Sascha Disch, "Using Transient Suppression in Blind Multi-channel Upmix Algorithms," presented at the AES 122nd Convention, Vienna Austria, paper 6990, 2007 May 5-8.

[11] Jean-Marc Jot and Carlos Avendano, "Spatial Enhancement of Audio Recordings," presented at the AES 23rd International Conference, Copenhagen, Denmark, 2003 May 23-25.

[12] Christof Faller, "Multiple-Loudspeaker Playback of Stereo Signals," *J. Audio Eng. Soc.*, vol. 54., no. 11, pp. 1051-1064, 2006 November.

[13] Juha Merimaa, Michael M. Goodwin, and Jean-Marc Jot, "Correlation-Based Ambience Extraction from Stereo Recordings," presented at the AES 123rd Convention, New York, NY, paper 7282, 2007 October 5-8.

[14] Michael M. Goodwin and Jean-Marc Jot, "Primary-ambient decomposition and vector-based localization for spatial audio coding and enhancement," *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, Honolulu, HI, USA, 2007 April.

[15] Michael M. Goodwin, "Geometric Signal Decompositions for Spatial Audio Enhancement," *Proc. IEEE International*

Conference on Acoustics, Speech, and Signal Processing, 2008 April.

[16] Aki Härmä and Christof Faller, "Spatial Decomposition of Time-Frequency Regions: Subbands or Sinusoids," presented at the AES 116th Convention, Berlin, Germany, 2004 May 8-11.

[17] Earl Vickers, Jian-Lung (Larry) Wu, Praveen Gobichettipalayam Krishnan, and Ravirala Narayana Karthik Sadanandam, "Frequency Domain Artificial Reverberation using Spectral Magnitude Decay," presented at the AES 121st Convention, San Francisco, CA, paper 6926, 2006 October 7.

[18] Avery Lee, "The 'Center Cut' Algorithm," http://www.virtualdub.org/blog/pivot/entry.php?id=102, 2006 May 21.

[19] Pythagoras, ca. 550 BC.

[20] Martin Walsh, Jean-Marc Jot, "Loudspeaker-Based 3-D Audio System Design using the M-S Shuffler Matrix," presented at the AES 121st Convention, San Francisco, CA, paper 6949, 2006 October 7.

[21] V. Pulkki and M. Karjalainen, "Localization of Amplitude-Panned Virtual Sources, I: Stereophonic Panning," *J. Audio Eng. Soc.*, vol. 49, pp. 739-752.

[22] Christian Uhle, Oliver Hellmuth, and Jan Weigel, "Speech Enhancement of Movie Sound," presented at the AES 125th Convention, San Francisco, CA, paper 7628, 2008 October 2-5.

[23] Francis Rumsey, "Controlled Subjective Assessments of Two-to-Five-Channel Surround Sound Processing Algorithms," *J. Audio Eng. Soc.*, vol. 47, no. 7/8, pp. 563-582, 1999 July/August.

[24] Thomas Sporer, Andreas Walther, Judith Liebetrau, Sebastian Bube, Christian Fabris, Thomas Hohberger, and Anja Köhler, "Perceptual Evaluation of Algorithms for Blind Up-mix," presented at the AES 121st Convention, San Francisco, CA, paper 6915, 2006 October 5-8.