Stereo Upmixing using Frequency Direction Enhancement and Ambience Extraction

Alexander Chiriboga
Department of Music Research, Schulich School of Music, McGill University
Montreal, Canada H3A 1E3
December 17, 2023

Abstract

This paper investigates a method for upmixing stereo audio to 5 channels following the Direct/Ambeince philosophy. Building upon Jot and Avendano's approach, the proposed method uses frequency direction enhancement to upmix the front channels and inter-channel coherence values for ambient extraction for the surround channels. This paper proposes an updated frequency-dependent repanning index, as the previous panning index is ill-defined for left-channel audio. A discussion on the theoretical foundations of stereophony and surround sound will be presented, along with the upmixing algorithm and MATLAB implementation. The performance of the proposed system is analyzed, highlighting its advantages and limitations compared to conventional stereo-to-surround upmixing techniques.

1. Introduction

Throughout history, sound has captivated us, whether as a means of communication or through creating artistic works. The ability to capture and replay sound, however, was impossible till the invention of the phonograph. As our technology has advanced, so has our ability to recreate aural experiences with greater realism. One of the most critical developments in spatial audio was stereophonic sound, which offered a new way to reproduce audio in three-dimensional space. Pioneered by Alan Blumlei in 1931, Stereophony developed a new way to capture and reproduce three-dimensional sound using two speakers.[1] This breakthrough could not recreate a sound field that enveloped the listener. To solve this, Dolby Laboratories implemented a 5-channel speaker setup known as surround sound, consisting of three speakers in the front and two in the back. The addition of rearward speakers helped create a sound field that could immerse the listener, while the additional center channel helped enlarge the listening sweet spot. Surround sound speakers are some of the most common sound systems today, with widespread availability of multichannel audio for movies and TV. However, despite the abundance of surround sound systems, most music media is still mixed and produced for stereo. The question of how to present stereo-mixed music in multi-channel format forms is central to designing an upmixing framework. This paper will discuss the direct/ambient approach proposed by Jot and Avendano, where the sources are panned in the front speakers, preserving the intent of the stereo mix. At the same time, the ambience is added to the surround speakers. This creates the perception of the listener being placed in the audience with a band in front of them. Under this model, the Short-Time Fourier Transform (STFT) between the left and right channels of the original audio are compared via inter-channel coherence. Using this measurement, we can perform source separation, allowing for the replanning of the sources and extraction of the ambience. This paper will review the theory of stereophony and surround sound, present the upmixing model proposed by Jot and Avendano, implement the model in MATLAB, and discuss the advantages and limitations of this method.

2. Stereophonic Reproduction

The most common method of spatial sound reproduction is stereophony. Stereophony can be defined as changing the amplitude and delay between the signals of two loudspeakers to create the perception of sound localization.[2] In stereophony, the loudspeakers are arranged in an equiangular triangle, with the listener positioned between the two loudspeakers in the space referred to as the "sweet spot." (see Figure 3)

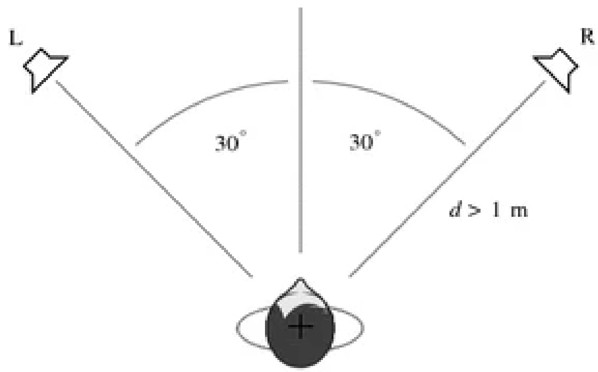

Figure 1: Standard stereo setup; the channels are termed left ('L') and right ('R'); the loudspeakers are at equal distance d from the listener.[3]

An auditory image can then be created through a psychoacoustic phenomenon known as summing localization,[4] which was later expanded into the association theory.[5] Summing localization and association theory describes how the auditory perception of multiple loudspeakers is translated into a single auditory event. In the figure above, when identical signals are sent to the two loudspeakers, a single sound source, known as the phantom source, is perceived to be located between the two loudspeakers. The relative position of this source can changed using amplitude panning (summed localization) and delay panning (the precedence effect), where the precedence effect describes how two time-delayed signals can be perceived as a single auditory event.[6] By increasing the amplitude of the signal sent to a given loudspeaker, the perceived location of a phantom source can be shifted towards the respective loudspeaker (amplitude panning). Inversely, delaying a loudspeaker signal shifts the perceived location of the phantom source away from the respective loudspeaker (delay panning). A consequence of amplitude and delay panning is the phantom sources can only be presented between the loudspeakers. This is especially problematic when trying to give a plausible presentation of reverb, where the lack of rearward speakers makes it hard to create the impression of being immersed in the room.

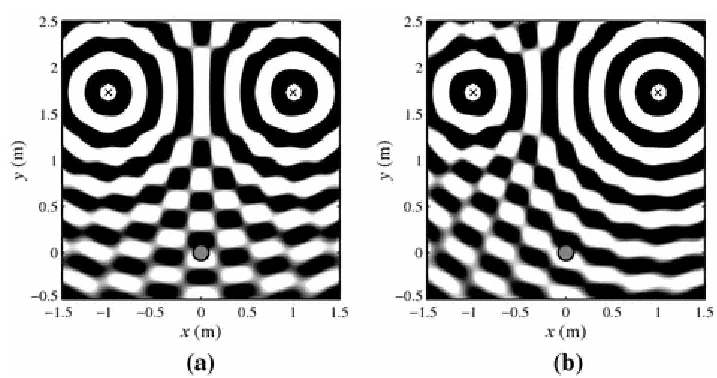

Figure 2: Sound field generated by two omnidirectional loudspeakers (represented by x) driven with a 1000hz signal. A grey disk represents the listener (located in the sweet spot). a) Both loudspeakers are driven with the same amplitudes. b) Right loudspeaker is driven with 6 dB higher amplitude[3]

While stereophony can recreate an accurate auditory experience, its generated sound fields are often inaccurate. Because stereophony relies on amplitude and delay panning from to create an auditory event, the listener's position heavily influences the perception of phantom sources. Because of this position-dependent panning, it's typically not possible to create the same perception in a location outside the sweet spot. For example, in the figure above, we see the generated sound field of a 1000hz signal using stereophony. If adequately synthesized, we should observe a 1000 hz plane wave propagating toward the listener. While the plane wave is correctly synthesized in the "sweet spot," listeners sufficiently far outside the sweet spot will perceive a positionally altered phantom source.

Arhens found that “When the listening position is chosen such that the relative timing between the loudspeakers is altered by significantly more than 1 ms, then the precedence effect can take over and the spatial composition of the presented scene collapses completely into the closest loudspeaker.” [3] This can be seen in the figure above; when sufficiently far outside the sweet spot, the wavefront can be observed as coming from the loudspeaker closest to the listener. As a result, the perception of the sound field is significantly altered if the listener’s position is outside the sweet spot.

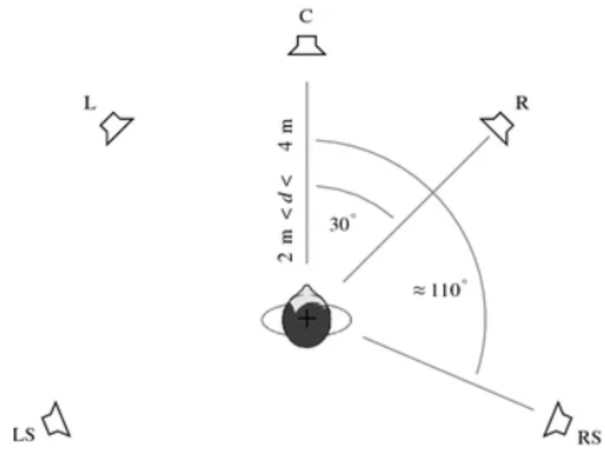

Figure 3: 5.0 surround sound setup; the front loudspeakers are termed left (‘L’), right (‘R’), and center (‘C’) while the rear loudspeakers are termed left surround (‘LS’) and right surround (‘RS’). [3]

As mentioned in the introduction, Dolby Laboratories proposed a five-speaker setup that promised to enlarge the sweet spot by introducing a center channel and creating higher immersion by allowing for rearward panned sources. Under this scheme, left and right loudspeakers are used similarly to two-channel Stereophony, except that panning is instead performed between the left/right speaker and the center speaker. The center channel also presents centrally panned audio, allowing for better source localization of centrally panned content to a listener off-axis, effectively creating a more prominent sweet spot. The Surround left and Surround right loudspeakers are often used to present decorrelated signals such as reverberation to enhance spatial perception and are occasionally used to present rearward sources. When presenting rearward content, the signal is sent to only one loudspeaker, where the source is localized at the loudspeaker’s position. [1] These two approaches, using surround speakers for reverberation or synthesizing rearward sources, demonstrate the two potential philosophies that can be used in stereo upmixing.

3. Stereo Upmixing

A P-channel recorded signal, $\mathbf{X} = [X_1 X_2 \ldots X_P]^T$ can be decomposed into primary and ambient components

$$\mathbf{X} = \sum_{i=1}^{N} \mathbf{D}_i \mathbf{S}_i + \sum_{i=1}^{N} \mathbf{R}_i S_i + \mathbf{B}$$

where S_i represents the source signals, $\mathbf{D}_i$ and $\mathbf{R}_i$ represent the spatial transfer function and reverberation response, and $\mathbf{B}$ represents a source-independent background noise. Ambient components are characterized as weakly correlated components that contribute evenly across all channels. The tail end of the reverberation and the background noise are the main contributors to the ambient signal. Schroeder proposed a more statistical definition in 1987, calling ambience an ”exponentially decaying, ergodic and stochastic process (that is) normally distributed with a mean of zero. [7] Ergodic describes the property of systems that can be statistically described from a sufficiently large set of points. The primary signal, meanwhile, is composed of contributions from the direct path of the source signal and the early reflections. Using these two components, two 5.1 mixing philosophies, direct/ambience, and in-the-band, can be described.

The direct/ambience approach seeks to emulate the experience of a live concert where the listener is placed in the audience. In this mixing philosophy, the primary components are panned across the front three speakers, and the ambient components are sent to the surround speakers. The in-the-band approach, meanwhile, seeks to have primary and ambient components contribute to all channels, creating the perception of the listener being placed inside the band. The direct/ambient philosophy is preferred because it seeks to reproduce the original stereo image as accurately as possible, meaning that it preserves the intent of the original stereo production. The most basic upmix of this direct ambient approach is described in the figure below.

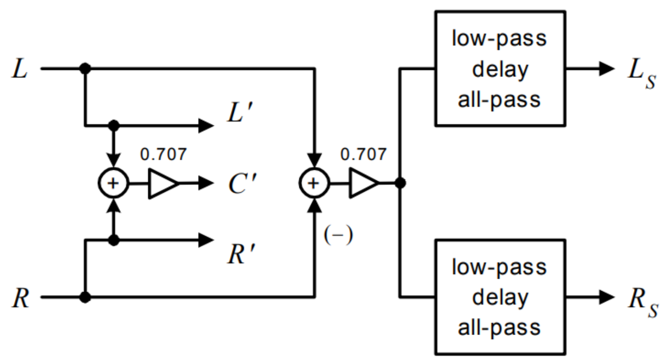

Figure 4: Passive 2-5 upmix matrix[8]

As can be seen in Figure 4 above, this upmix constitutes a direct/ambient approach where the ambient signal S is defined by subtraction of the left channel from the right channels where

$$S = 0.707(L - R).$$

By subtracting R from L, centrally panned components are removed while laterally panned or weakly correlated components are preserved. We can additionally define the center channel as the sum of the left and right channels, where the 2-3 upmix scheme is defined by the following matrix

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

Adding Low Pass filters and all pass delays helps prevent the surround signal from shifting the auditory image. A drawback of the passive matrix is that the stereo image is narrowed by 25%, meaning speakers need to be placed further apart to preserve the original stereo image. Additionally, this system suffers from low channel separation. Consequently, the precedence effect dictates that the source appears from the loudspeaker closest to the listener. Jot and Avendano proposed an improved upmix process that would remove all the center-panned components from the left and right channels and direct them to the center channel.[8] Additionally, they proposed that higher channel separation could be achieved by repanning sources through two adjacent speakers at a time. This is achieved by only panning over the center and the left channel or the center and the right channels through a technique known as frequency direction enhancement.

3.1 Frequency Direction Enhancement

As discussed above, an ideal 2-3 upmixing matrix should remove all centrally panned components from the left and right audio and place them in the center channel. We can measure how closely related two channels are using a Short-Time Fourier Transform (STFT) and a left-right similarity measure

$$\varphi(m, k) = \frac{2C_{LR}}{C_{LL} + C_{RR}}.$$

C_{ij} is the instantaneous cross-spectrum, which tells us how closely related the STFT values of the left and right channels are at a given frequency bin and time window where $C_{ij} = X_i(m, k)X_j^*(m, k)$. The panning index value α, defined from [0 1], can then be derived from the left-right similarity measure, where

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

where the ambiguity of the left vs right panning can be resolved by solving for the higher amplitude of the two STFT values. If $|X_L(m,k)| > |X_R(m,k)|$ a signal is considered to be panned left and if $|X_L(m,k)| < |X_R(m,k)|$ the signal is considered to be panned right. We can then derive the new 3-channel panning index where the center channel magnitude is dictated by a magnitude of $\sin(2\alpha)$, and the left and right channels have a magnitude of either 0 or $\cos(2\alpha)$.

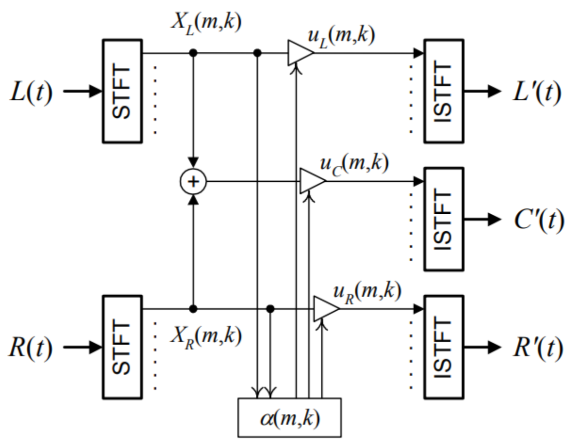

Figure 5: Frequency direction enhancement 2-3 upmix matrix[8]

$$u_c = \frac{\sin(2\alpha)}{\sin(\alpha)+\cos(\alpha)}, \quad u_l = -\min\left(0, \frac{\cos(2\alpha)}{\sin(\alpha)}\right), \quad \text{and} \quad u_r = \max\left(0, \frac{\cos(2\alpha)}{\cos(\alpha)}\right)$$

are the new panning index coefficients for the center, left, and right channels, respectively. A signal flow diagram, shown in Figure 5 above, demonstrates how the STFT, panning index value, and new panning index alter the signal to upmix stereo audio for a three-channel presentation. These minimum and maximum functions in u_l and u_r are essential to ensure we are only panning between two sources. This helps create a high channel separation by removing right-panned components from the left speaker and vice versa. Additionally, the left and right panning weights are zero for centrally panned content (when $\alpha = \pi/4$), and the only contribution is from the center channel. This fulfills both of the design criteria discussed in the passive upmix scheme.

3.2 Ambience Extraction

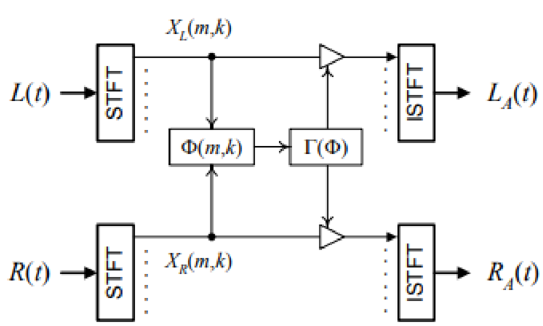

Figure 6: Ambeince Extraction matrix[8]

The ambience can also be found using the instantaneous cross-spectrum $C_{ij} = X_i(m,k)X_j^*(m,k)$, where the left/right coherence index is defined as

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

where $\phi_{ij}(m,k)$ is

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

and lambda is the forgetting factor, which tells us how much to weigh the previous sample. Since the left/right coherence index tells us how closely related two channels are, and the ambience by definition has lose correlation between stereo channels, we must subtract the left/right coherence index from one to get the Ambience index

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

To smooth out any artifacts, a soft decision function $\Gamma(\Phi)$ is applied to the Ambience index

$$\Gamma(\Phi) = \frac{\mu_1 - \mu_0}{2}\tanh\left[\sigma\pi(\Phi - \Phi_0)\right] + \frac{\mu_1 + \mu_0}{2}$$

where μ_1, μ_0 define the range of the output (the upper and lower thresholds), and σ controls the slope of the function. This method allows for a more accurate method of ambience presentation since both the left and the right channel present slightly different content.

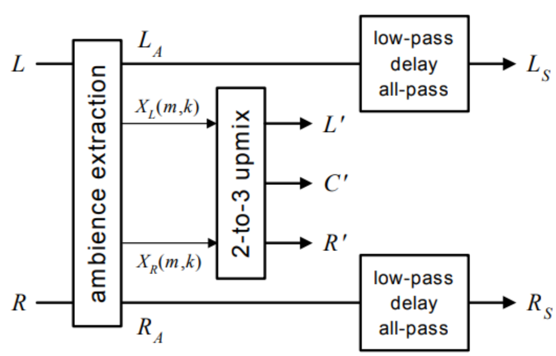

Figure 7: Frequency-dependent repanning and ambience extraction 2-5 upmix matrix [8]

Similar to the passive system, Low Pass filters and all pass delays are implemented to prevent the surround signal from shifting the auditory image. This can be seen in Figure 7 above, where the entire upmix matrix scheme is shown. This upmix scheme is effective

4. Discussion

The previous section discussed the theory and design goals for upmixing stereo content to 5-channel audio, focusing on source separation and repanning for the front channels and ambience extraction for the rear channels. While this upmixing job did a good job of preserving the stereo image, the method shown above has a couple of discrepancies and limitations.

The panning index uses the left-right similarity measure $\varphi(m,k)$ based on the instantaneous cross-spectrum C_{ij} between the left and right stereo channels. The panning index assumes that the panning index α spans the range $[0\ \pi/2]$, where fully panned left is $\pi/2$, the center channel is $\pi/4$ and fully panned right is 0.

What seems to be overlooked is that the left-right similarity measure is defined in the range [0 1]. When computing the range of the panning index value α we see it is only defined from $[0\ \pi/4]$. Under this assumption, all centrally panned content is sent to the center channel, and all the left or right-panned content is sent to the right channel, while the left channel is always zero. To resolve this issue, the panning index for the left and right weights was modified to be

$$u_l = \begin{cases} \frac{\cos(2\alpha - \pi)}{\sin(\alpha - \frac{\pi}{2})} & \text{if } |X_L| - |X_R| > 0 \\ 0 & \text{if } |X_L| - |X_R| \leq 0 \end{cases} \quad \text{and} \quad u_r = \begin{cases} \frac{\cos(2\alpha)}{\sin(\alpha)} & \text{if } |X_R| - |X_L| > 0 \\ 0 & \text{if } |X_R| - |X_L| \leq 0 \end{cases}.$$

This panning index ensures that the left channel is defined in the repanning scheme and prevents panning between the left and right channels. As discussed in our upmix design goals, the panning should only involve the center and left or center and right channels. This panning index was implemented in the MATLAB code (see below), where the interchannel coherence was used instead of comparing the STFT values to determine the panning.[9] C_{LL} and C_{RR} can also be defined as $|X_L|^2$ and $|X_R|^2$.

```
if C_LL-C_RR > 0
    u_l(k, m) =  cos(2*alpha(k, m)-pi)/sin(alpha(k, m)-pi/2);
else
    u_l(k, m) = 0;
end

if C_LL-C_RR < 0
    u_r(k, m) =  cos(2*alpha(k, m))/cos(alpha(k, m));
else
    u_r(k, m) = 0;
end
```

When testing mono audio sent to the left and silent right channels, issues with dividing by zero were noticed. The following MATLAB script, which uses if statements to prevent ambiguity when dividing 0/0, circumvents this issue. This issue was especially problematic for the left/right coherence index $\phi(m, k)$ and the panning index value α. Two different ways of dealing with this issue were applied here. The entire MATLAB script is included in Appendix A.

```
if sqrt(phi_LS(k, m)*phi_RS(k, m)) == 0
    Phi(k, m) = 1 - abs(phi_LR(k, m))/(sqrt(phi_LS(k, m)*phi_RS(k, m)
        +0.001));
else
    Phi(k, m) = 1 - abs(phi_LR(k, m))/(sqrt(phi_LS(k, m)*phi_RS(k, m))
        );
end

f C_LR == 0
    alpha(k, m) = 0;
else
    alpha(k, m) =  abs(asin( 2 * C_LR / (C_LL + C_RR)) / 2);
end
```

The ambiance extraction process identifies weakly correlated components between the left and right channels by comparing STFT values with the instantaneous cross-spectrum C_{ij}. This method assumes that only the ambience component of the signal has a loose correlation between channels. This assumption is valid for reverberation or distant background noise but incorrectly extracts primary components when a signal is present in only one channel. This is because the ambience extraction uses the coherence index between the channels. When this coherence index is low (as in the case of a primary signal present only in one channel), the method assumes the primary signal is an ambience component. This then causes the incorrect perception of another source behind the listener.

5. Conclusion

A method for upmixing stereo audio to 5 channels was presented to enhance the spatial immersion of stereo-mixed music in a surround sound system. Using frequency-dependent repanning and ambience extraction preserves the original stereo image and the integrity of the mix while creating increased immersion through ambience reproduction over the surround speakers. The misidentification of primary signals with low inter-channel coherence as ambience components was discussed. Further research is needed to improve this ambience extraction process in the edge case where primary signals are sent to only one channel. Overall, this approach offers a significant step towards creating more immersive listening experiences for stereo recordings in surround sound systems.

References

[1] F. Rumsey. *Spatial Audio*. Focal Press, 2001.

[2] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, revised edition, 1997.

[3] Jens Ahrens. *Analytic Methods of Sound Field Synthesis*. Springer, 2012.

[4] H. Warncke. Die grundlagen der raumbezüglichen stereophonischen Übertragung im tonfilm (the fundamentals of room-related stereophonic reproduction in sound films). *Akustische Zeitschrift*, 6:174–188, 1941.

[5] G. Theile. *On the localisation in the superimposed soundfield*. PhD thesis, Technische Universität Berlin, 1980.

[6] H. Haas. Uber den einfluss eines einfachechos auf die horsamkeit von sprache (on the influence of a simple echo on the comprehensibility of speech). *Acustica*, 1:49–58, 1951.

[7] M. R. Schroeder. Statistical parameters of the frequency response curves of large rooms. *Journal of the Audio Engineering Society*, 35(5):299–305, 1987.

[8] J.-M. Jot and C. Avendano. Spatial enhancement of audio recordings. In *23rd International Conference, Audio Engineering Society*, May 2003.

[9] MATLAB. *MATLAB R2023b*. MathWorks, 2023. https://www.mathworks.com/products/matlab.html.

A. MATLAB Implementation

```
clc
clear

[stereo_signal, fs] = audioread('Stereo Test.mp3');

L = stereo_signal(:, 1); % Left channel
R = stereo_signal(:, 2); % Right channel



% Parameters for STFT
window_size = 1024;
overlap = window_size / 2;
nfft = window_size;

% Compute the STFT for both channels
X_L = stft(L, 'Window', hamming(window_size, 'periodic'), 'OverlapLength', overlap, 'FFTLength', nfft);
X_R = stft(R, 'Window', hamming(window_size, 'periodic'), 'OverlapLength', overlap, 'FFTLength', nfft);

% Presets
lambda = 0.85;  % Forgetting factor
sigma = 8; %slope
mu_0 = 0.1;
phi_0 = 0.15;

%Intialize vectors
phi_LR = zeros(size(X_L));
phi_LS = zeros(size(X_L));
phi_RS = zeros(size(X_L));
Phi = zeros(size(X_L));
u_c = zeros(size(X_L));
u_l = zeros(size(X_L));
u_r = zeros(size(X_L));
alpha = zeros(size(X_L));



% Compute the cross-correlation and coherence index

for m = 1:size(X_L, 2)
    for k = 1:size(X_L, 1)
        C_LR = X_L(k, m) * conj(X_R(k, m));
        C_LL = X_L(k, m) * conj(X_L(k, m));
        C_RR = X_R(k, m) * conj(X_R(k, m));

        if m > 1
            phi_LR(k, m) = (1 - lambda) * phi_LR(k, m-1) + lambda * C_LR;
            phi_LS(k, m) = (1 - lambda) * phi_LS(k, m-1) + lambda * C_LL;
            phi_RS(k, m) = (1 - lambda) * phi_RS(k, m-1) + lambda * C_RR;
```

```matlab
        else
            phi_LR(k, m) = lambda * C_LR;
            phi_LS(k, m) = lambda * C_LL;
            phi_RS(k, m) = lambda * C_RR;
        end

        if sqrt(phi_LS(k, m)*phi_RS(k, m)) == 0
            Phi(k, m) = 1 - abs(phi_LR(k, m))/(sqrt(phi_LS(k, m)*phi_RS(k, m)+0.001));
        else
            Phi(k, m) = 1 - abs(phi_LR(k, m))/(sqrt(phi_LS(k, m)*phi_RS(k, m)));
        end

        if C_LR == 0
             alpha(k, m) = 0;
        else
            alpha(k, m) =  abs(asin( 2 * C_LR / (C_LL + C_RR)) / 2);
        end
        u_c(k, m) = sin(2*alpha(k, m))/(sin(alpha(k, m))+cos(alpha(k, m)));

        if C_LL-C_RR > 0
            u_l(k, m) =  cos(2*alpha(k, m)-pi)/sin(alpha(k, m)-pi/2);
        else
            u_l(k, m) = 0;
        end

        if C_LL-C_RR < 0
            u_r(k, m) =  cos(2*alpha(k, m))/cos(alpha(k, m));
        else
            u_r(k, m) = 0;
        end
    end
end

% Apply the soft-decision function
Gamma = (1-mu_0)/2*tanh(sigma*pi*(Phi-phi_0))+(1+mu_0)/2;

% Reconstruct the signals
X_nL = X_L.* u_l;
X_C = X_L .* u_c + X_R .* u_c;
X_nR = X_R.* u_r;
X_LS = X_L .* Gamma;
X_RS = X_R .* Gamma;

% Inverse STFT
L_upmix = real(istft(X_nL, 'Window', hamming(window_size, 'periodic'), 'OverlapLength', overlap, 'FFTLength', nfft));
C_upmix = real(istft(X_C, 'Window', hamming(window_size, 'periodic'), 'OverlapLength', overlap, 'FFTLength', nfft));
R_upmix = real(istft(X_nR, 'Window', hamming(window_size, 'periodic'), 'OverlapLength', overlap, 'FFTLength', nfft));
LS = real(istft(X_LS, 'Window', hamming(window_size, 'periodic'), 'OverlapLength', overlap, 'FFTLength', nfft));
```

```
RS = real(istft(X_RS, 'Window', hamming(window_size, 'periodic'), '
    OverlapLength', overlap, 'FFTLength', nfft));

% All-pass delay and low pass
delay_time = 0.015; % Delay in seconds (5-20 ms)

% Compute the delay in terms of samples
delay_samples = round(delay_time * fs);

% Compute the all-pass filter coefficients
a = exp(-2 * pi * delay_samples / fs);
b = [1 -a];
a_filter = [1 -a];

% Delay ambience to decorrelate stereo image
LS_delay = filter(b, a_filter, LS);
RS_delay = filter(b, a_filter, RS);

% Low-pass filter
cutoff = 120; % Low-pass cutoff frequency in Hz
order = 4;   % Filter order

% Design a Butterworth low-pass filter at 120 Hz
[b_lp, a_lp] = butter(order, cutoff / (fs / 2), 'low');

% Apply low-pass filter to LS and RS signals
LS_upmix = filter(b_lp, a_lp, LS_delay);
RS_upmix = filter(b_lp, a_lp, RS_delay);


stereo(:,1) = L_upmix;
stereo(:,2) = R_upmix;
rear(:,1) = LS_delay;
rear(:,2) = RS_delay;

% Find the maximum absolute value in the audio signal
max_val = max(abs(C_upmix(:)));

% Normalize the audio signal if necessary
if max_val > 1
    C_upmix = C_upmix / max_val;
end

audiowrite("LR.mp3", stereo, fs)
audiowrite("LSRS.mp3", rear, fs)
audiowrite("center.mp3", C_upmix, fs)
```