Interactive Enhancement of Stereo Recordings Using Time-Frequency Selective Panning

Maximo Cobos[1] and Jose J. Lopez[1]

[1] *Institute of Telecommunications and Multimedia Applications (iTEAM), Universidad Politecnica de Valencia, Valencia, 46022, Spain*

Correspondence should be addressed to Maximo Cobos (mcobos@iteam.upv.es)

ABSTRACT

Localization of sounds in physical space plays a very important role in multiple audio related disciplines, such as music, sound art or sound editing for audiovisual productions. The most well known technique for providing such spatial impression is stereo panning, which creates a virtual location of a sound source by distributing its energy between two independent channels during the mixing process. However, once all the sound events have been mixed, re-distributing source locations to widen the stereo image is not certainly an easy task. Motivated by this problem, we propose a source spatialization technique based on the time-frequency processing of the stereo mixture. The energy distribution over the stereo panorama is modified according to a non-linear warping function, providing a widening effect that enhances the stereo experience without degrading the sound quality and preserving the original conception of the mixing engineer.

1. INTRODUCTION

One of the most important cues in spatial perception of sound is localization. Generally, sound is perceived in all three dimensions, width, height and depth, which all are necessary to achieve a natural perception of sound [1]. The term *spatialization* denotes the use of the localization of sounds in physical space as a compositional element in music, in sound art, and in sound editing for audio recordings, film, and video [2]. In fact, positioning sound objects in a virtual space is a key aspect in audio mixing, which is the process by which a multitude of recorded sounds are combined into one or more channels, most commonly two-channel stereo [3].

Stereo reproduction [4] is based on the fact that, if a particular source is appropriately scaled and/or delayed between the left and right channels, it will appear to originate from an imaginary position (the so-called phantom sound source) along a line connecting both loudspeakers (the loudspeaker basis). The most frequently used technique to position sound objects in a virtual space is *amplitude panning* [5]. Mixing desks and software audio workstations operate by attenuating and panning audio signals independently before being added. Panning means to assign an stereo position to a particular sound signal by sending two differently scaled versions of it to the output left and right channels. Therefore, a sound image where different sound sources are distributed in a virtual space is generated, providing the listener with an enhanced spatial experience in the mix. However, once this process is finalized, there is little option for the listeners to modify the spatial impression of the recording and the sound imagery evoked is limited to the one contained in the final mix. To mitigate this problem, we propose a selective time-frequency panning technique with the aim of enhancing the spatial impression of common stereo sound material. According to a non-linear warping function, the virtual location of the sound sources contained in the mixture is modified interactively to widen the stereo scene, preserving the intention of the mixing engineer and the quality of the original recording.

The paper is structured as follows. Section 2 provides a brief review of stereo recording techniques, which will be helpful to clarify and characterize different types of stereo sound material. Section 3 explains some of the classical approaches to stereo enhancement and widening, which are quite related to the topic covered in this paper. Section 4 describes our proposed approach in detail, discussing a set of clarifying examples. In Section

5, we evaluate the performance of the proposed approach from a subjective perspective and, finally, the conclusions of this work are summarized in Section 6.

2. STEREO RECORDING

To date, stereophony is still the most common format for sound recording and reproduction. Although multichannel recordings for 5.1 reproduction systems have been widely available since the arrival of DVDs, they have not displaced the classic stereo format yet. The vast majority of CDs, compressed formats such as MP3 or AAC, FM radio broadcasts, as well as many analogue and digital TV broadcasts, are in stereo.

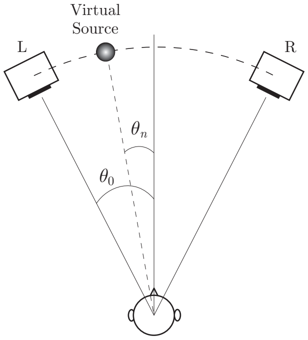


Fig. 1: Stereo reproduction set-up

The motivation of stereo recording and reproduction relies on the fact that the physical superposition of two loudspeakers enables the building of a *phantom source*, which is understood as a substitute sound source. This effect is called "*summing localization*", which is supposed to create binaural cues very similar to the ones created by real sources [4]. There are objections to this explanation. In his "*association model*", Theile [6] argues that the superposition of the loudspeaker signals does not create localization, but rather that the signals from the two loudspeakers give two different localization stimuli that merge together into a phantom source after a complex psychoacoustic process. The angle of incidence θ_n of the phantom sound source to the listener is commonly referred as the azimuth angle of the source, and it depends on the relative position of the listener to the loudspeakers, given by the loudspeaker base angle θ_0 (see Figure 1). The correct perception of the source direction only occurs if the listener is on the vertex completing an equilateral triangle with the loudspeakers ($\theta_0 = 30^\circ$). This point is called the *sweet spot* [7].

Although there exist techniques involving phase/time differences between stereo channels, we will center our discussion on *intensity stereophony* [8]. Therefore, we assume a linear mixing model where N sound sources are multiplied by scalars before being added, resulting in the following stereo mixing model:

$$x_l(t) = \sum_{n=1}^{N} a_{ln} s_n(t), \tag{1}$$

$$x_r(t) = \sum_{n=1}^{N} a_{rn} s_n(t), \tag{2}$$

where $x_l(t)$ is the left stereo channel, $x_r(t)$ is the right stereo channel, $s_n(t)$ is the n-th source signal and, finally, a_{ln} and a_{rn} are the mixing coefficients for the n-th source in the mix.

In the following, a general distinction will be made between natural stereo mixtures and synthetic stereo mixtures. Natural mixing refers to recording situations in which the mixing parameters are determined by the relative positions of a set of sound sources and the microphones. In contrast, synthetic mixing consists of artificially combining a set of separated sound sources using a mixing desk or mixing software. Traditionally, natural techniques are preferred for classical music, whereas synthetic techniques are most common in popular genres, in which studio post-processing effects play a crucial role. Next, several well known techniques for stereo recording are briefly described.

2.1. Natural Mixing Techniques

Natural mixing methods involve a pair of directional microphones whose membranes are located at the same point. The directionality properties of the microphones is used to obtain intensity differences which are dependent on the actual directions of the recorded sound sources. The directionality or polar pattern of the microphone indicates its sensitivity to sound pressure as a function of the angle of arrival of the waves. The most common approaches using this principle are the XY and MS stereophony techniques. Other popular techniques not covered in this section can be easily found in the literature [9].

2.1.1. XY Recording

The set-up used in this technique is graphically represented on Figure 2(a). Both microphones are directional and the stereo effect is achieved by mutually rotating

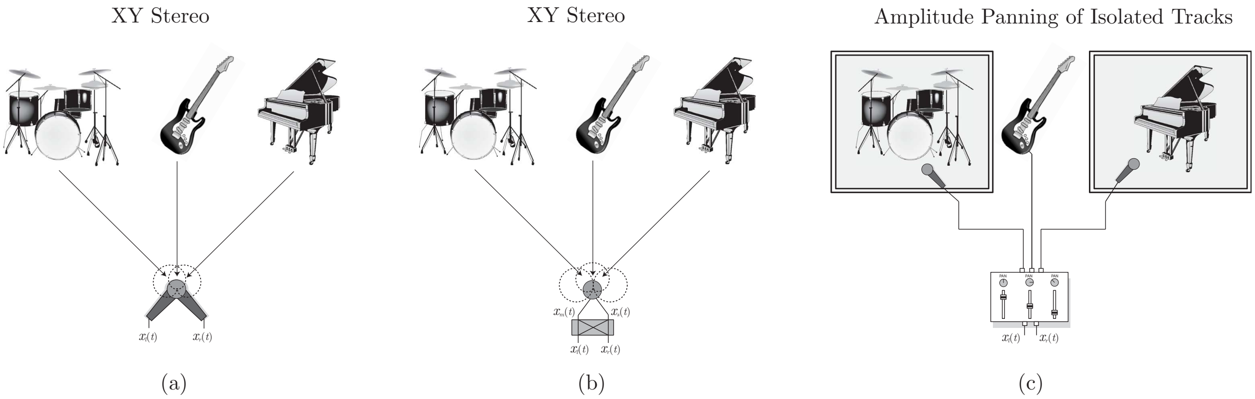


Fig. 2: Stereo recording techniques. (a) XY stereo. (b) MS Stereo. (c) Amplitude Panning

them to a certain angle, usually 90°. The direct sound waves coming from the different sources, arriving from different directions and distances, are picked up with different intensities dependent on the angle of impingement. The mixing coefficients are therefore those corresponding to the polar pattern of the microphones used: $a_{ln} = D_l(\theta_n)$ and $a_{rn} = D_r(\theta_n)$, being $D_l(\theta)$ and $D_r(\theta)$ the polar response of both microphones.

2.1.2. MS Recording

The Mid/Side (MS) technique uses one bidirectional and one directional (or, alternatively, omnidirectional) microphone at the same place arranged such that the point of maximum directivity of the directional microphone lies at an angle of 90° from either bidirectional maximum (see Figure 2(b)). In this way, a central channel $x_m(t)$ and a side channel $x_s(t)$ are obtained, which are then transformed into the left/right channels by this simple operation

$$x_l(t) = \frac{1}{\sqrt{2}}\left(x_m(t) + x_s(t)\right), \tag{3}$$

$$x_r(t) = \frac{1}{\sqrt{2}}\left(x_m(t) - x_s(t)\right). \tag{4}$$

Obviously, the mixing parameters will depend again on the directional characteristics of the sensors. An advantage of the MS system is its total compatibility with mono reproduction: the middle signal directly corresponds to the mono signal, avoiding possible phase cancellations and level imbalances that can appear when adding two separated stereo channels.

2.2. Synthetic Mixing Techniques

Most electrical and electroacoustic instruments, as well as any other kind of electronic sound generators such as synthesizers or computers running synthesis software, can be directly connected to a mixing unit that creates a stereo mixture by adding the channels corresponding to the different sound sources. A synthetic mixture can also be generated by adding artificially the signals obtained by recording isolated instruments, as it is often done with singers. The stereo effect is then achieved by using amplitude panning, which is next described.

To reproduce the image of a sound source at some angle θ_n with $|\theta_n| < \theta_0$, amplitude panning is applied (Figure 2(c)). With this technique, the same driving signal is fed to the right and left channels, but with different weighting factors a_{ln} and a_{rn}. These are selected such that the superposition of the sound fields of both loudspeakers makes the listener perceive a single sound source (phantom source) at the desired angle θ_n. The functional dependency of the weighting factors on θ_n is called *panning law*.

2.2.1. Sine Law

The sine panning law is based on the model of a locally plane wave [10]. The weighting factors are determined such that the wave field in the close vicinity of the listener can be approximated by a plane wave with a normal vector in the desired position. The resulting panning law is given by

$$\frac{\sin\theta_n}{\sin\theta_0} = \frac{a_{rn} - a_{ln}}{a_{rn} + a_{ln}}. \tag{5}$$

For a complete explanation of the mathematical derivation of the above formula, we refer the reader to [11].

2.2.2. Tangent Law

The tangent panning law is based on a projection of the phantom source at angle θ_n to a coordinate system represented by unit vectors in the direction of the right and left loudspeakers. The gain factors are found as the corresponding projection coefficients:

$$\frac{\tan\theta_n}{\tan\theta_0} = \frac{a_{rn} - a_{ln}}{a_{rn} + a_{ln}}. \tag{6}$$

A common way of assigning the panning coefficients in audio mixers and computer music software is by means of a panning parameter $p_n \in [0,1]$, which is controlled by the user to adjust the virtual location of a given sound source in the stereo panorama. Usually, the coefficients follow an energy-preserving law:

$$a_{ln}(p_n) = \cos(p_n \cdot \pi/2), \tag{7}$$

$$a_{rn}(p_n) = \sin(p_n \cdot \pi/2), \tag{8}$$

so that $a_{ln}^2 + a_{rn}^2 = 1$. Therefore, the equivalence between between the panning parameter p and the source mixing ratio is given by:

$$p_n = (2/\pi)\arctan(a_{nl}/a_{rl}). \tag{9}$$

Note that the ratio a_{ln}/a_{rn} goes from 0 to $+\infty$ depending on the value of p_n, positioning a given source between $\theta_n = \theta_0$ and $\theta_n = -\theta_0$. Figure 3 shows the angle θ_n of a phantom source depending on the relation of the panning parameter p_n according to the sine and tangent panning laws and a loudspeaker angle $\theta_0 = 30°$. Obviously, there are only minor differences between the panning laws.

3. STEREO ENHANCEMENT

Since the widespread introduction of stereophonic sound reproduction in the 1950s, multiple methods have been proposed to allow manipulation of the spatial characteristics of sound recordings [12]. Stereo enhancement refers to processing stereophonic music or sound in such a way as to add spaciousness to the stereo sound field. The purpose of stereo enhancement is to widen the stereo sound field, thereby immersing the listener in a cleaner, richer sound experience, significantly improving the quality, depth and feel of the sound played. From a practical point of view, stereo enhancement is intended to spread the stereo field into a 180 degree arc in front of the listener in the case of headphone reproduction. When using a stereo loudspeaker set-up, the aim is to spread the stereo field into the arc going from $-\theta_0$ to $+\theta_0$.

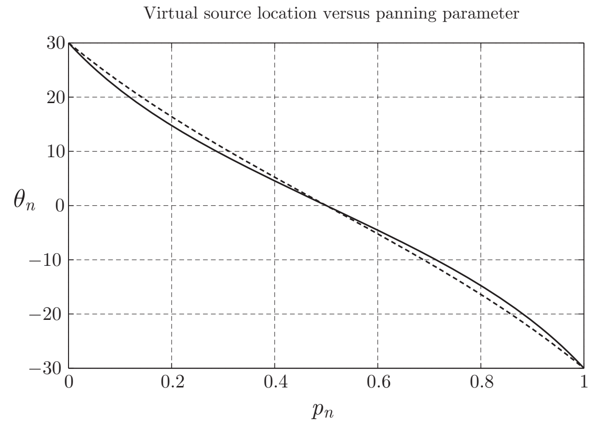


Fig. 3: Location θ_n of a phantom source versus the relation of the panning parameter p_n for the sine (solid line) and tangent (dashed line) panning laws and for $\theta_0 = 30°$

Stereo enhancement can be applied to listening situations with loudspeakers as well as headphones. Although the field of stereo enhancement has relatively few instances of scientific literature as compared to source positioning using binaural synthesis [1], there is a huge amount of patents related to enhancement of stereophonic recordings.

3.1. Classical Approach

Many enhancement schemes make use of the channel difference signal and/or the sum signal in order to emphasize the difference between the left and right signals [13][14][15]. For example, omitting the time dependence for the shake of clarity, if the signals x_l and x_r contain a substantial common component, it is possible to express them as follows:

$$x_l = x_m + x_{ls}, \tag{10}$$

$$x_r = x_m + x_{rs}, \tag{11}$$

where x_m is the common signal and x_{ls} and x_{rs} are the left-only and right-only components. In this situation, the channel difference signal can be written as

$$x_{\mathrm{dif}} = x_l - x_r = x_{ls} - x_{rs}, \tag{12}$$

so adding x_{dif} to x_l gives $x_m + 2x_{ls} - x_{rs}$, which boosts the proportion of x_{ls} in the composite left signal. Similarly, subtracting x_{dif} from x_r performs the same operation on the right channel. Furthermore, the presence of the inverted components ($-x_{rs}$ in the left output and $-x_{ls}$ in the right output) also serves to give a broadened spatial impression to the resulting stereo sound field [12]. However, these simple methods do not account for the real spatial properties of individual sources in the mix and, therefore, the spatialization effect is not applied selectively according to the actual spatial configuration of the instruments present in the recording.

3.2. Up-mixing and Related Approaches

Closely related to stereo enhancement are those techniques aimed at audio up-mixing. While recent movie soundtracks and some musical recordings are available in multichannel format, most music recordings are mixed in stereo. The playback of this material over a multichannel system poses a fundamental problem: stereo recordings are mixed with a very particular set up in mind, which consists of a pair of loudspeakers placed symmetrically in front of the listener. Thus, listening to this kind of material over a multichannel sound system raises the question of what signals should be sent to the additional channels. In this context, audio systems aimed at solving the up-mixing problem are widely used today, for example, Dolby Pro Logic II [16].

Dolby began introducing the multichannel to stereo down-mix feature in its codecs in order to respond to the requirements of backwards compatibility. Additionally, Dolby Pro Logic II includes the up-mix from stereo back to 5.1 multichannel format. In the down-mix, the original source audio signals are encoded into two program channels, that can be played back as stereo. The left and right stereo signals, called left-total and right-total, or *Lt* and *Rt*, are assembled by adding to the left and right multichannel signals (L and L) the center channel signal (C) as well as the corresponding surround channel signal (LS or RS), both attenuated by 3 dB. The phases of the surround channel signals are additionally shifted by 90 degrees and they are added with opposite signs. Similarly, the up-mix is carried out using the following decoding matrix:

$$\begin{bmatrix} L \\ R \\ C \\ LS \\ RS \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \mathrm{j}\frac{1}{\sqrt{\frac{3}{2}}} & -\mathrm{j}\frac{1}{\sqrt{3}} \\ \mathrm{j}\frac{1}{\sqrt{3}} & -\mathrm{j}\frac{1}{\sqrt{\frac{3}{2}}} \end{bmatrix} \cdot \begin{bmatrix} Lt \\ Rt \end{bmatrix}, \qquad (13)$$

The LFE channel signal is derived by low-pass filtering the sum of Lt and Rt signals.The LFE channel signal is derived by low-pass filtering the sum of *Lt* and *Rt* signals.

Alternatives to simple matrix-based up-mixing systems have also been proposed, which share many similarities with the stereo enhancement method proposed in this paper. For example, Avendano and Jot developed more advanced frequency domain techniques for the up-mix of stereo recordings into multichannel audio [17]. Aiming at a natural and generic multichannel audio mix, their method takes into account both the apparent directions of individual sound sources, and the ambient sound consisting of diffuse sound, reverberation and noise. The method compares the STFT of the left and right stereo signals and identifies a set of components for the up-mix using a time-frequency analysis/synthesis approach.

In fact, the analysis of components in this domain settles the basis of several sound processing systems aimed at estimating and/or modifying the spatial characteristics of multichannel audio signals. For example, Binaural Cue Coding (BCC) [18][19] is a technique capable of providing multichannel audio at low data rates. In the BCC transmitter, the original stereophonic signal is down-mixed to a mono signal that is afterwards compressed by an audio encoder. At the receiver side, the BCC synthesizer generates the stereophonic output from the compressed mono signal with the help of some side information, which is needed to restore the original spatial localization cues.

Another related technique is Directional Audio Coding (DirAC) [20]. In a first analysis stage, DirAC uses typically a B-format microphone to capture the spatial properties of the sound recorded in a given environment. In a second stage, the analyzed spatial features are employed to reproduce the recorded sound again by means of an arbitrary loudspeaker set-up [21].

Note that, despite sharing many similarities in the processing, the application covered in this paper is quite different from the above techniques. While the above techniques are mainly oriented to preserving the spatial characteristics of sound and reducing the required data rate of multichannel audio signals, the method presented in this paper is intended to modifying the spatial characteristics of a stereo recording with a simple parametric approach.

4. TIME-FREQUENCY SELECTIVE PANNING

The classical approach to stereo enhancement usually followed by commercial devices has the advantage of being extremely simple, so it has been widely deployed in many audio reproduction systems for years. However, the high computational power of today's portable devices makes it possible to think of new processing schemes for enhanced stereo reproduction. Small digital cameras, PDAs, iPods or mobile phones have sufficient power to incorporate sophisticated spatialization features, which provides the user with the possibility of widening interactively the stereo scene of any sound track in its reproduction.

In the following subsections, we describe in detail our proposed approach to stereo spatialization, which is based on the time-frequency processing of the stereo input. First, the left and right channels are transformed into the time-frequency domain using the *Short-Time Fourier Transform* (STFT), which provides a sparse representation of the signal useful for the subsequent processing. Afterwards, amplitude differences between the stereo channels are analyzed in this domain, which allows to observe the energy distribution of the processed sound over the stereo panorama. Finally, the energy of the sound is re-distributed according to a parameter of a non-linear warping function, which can be modified by the listener interactively.

4.1. Time-Frequency Transformation

Consider the stereo linear mixing model given by Equations (1) and (2). It is here useful to rewrite the model in the time-frequency domain considering the STFT of the left and right signals. This transform divides a time domain signal into a series of small overlapping pieces; each of these pieces is windowed and then individually Fourier transformed [22]. Mathematically, the STFT of a signal $s(t)$ is given by:

$$S(k,m) = \sum_{l=0}^{L-1} w(l)s(l+mH)e^{-j\omega_k l}, \qquad (14)$$

where $w(l)$ is the analysis window, L is the window length, m is the time frame index, and H is the hop size, i.e., the spacing in samples between consecutive applications of the sliding window. Due to the linearity of the STFT, the stereo mixing model can be expressed as

$$X_l(k,m) = \sum_{n=1}^{N} a_{ln}S_n(k,m), \qquad (15)$$

$$X_r(k,m) = \sum_{n=1}^{N} a_{rn}S_n(k,m), \qquad (16)$$

where $S_n(k,m)$ is the STFT of the n-th source signal.

Signal decompositions or additive expansions have been shown to provide a sparse representation of audio signals, which is a desirable feature for many audio processing systems. A signal is said to be sparse if most of its components are zero or close to zero. The sparsifying properties of time-frequency transformations have been successfully applied in the context of applications where multiple sound sources are typically active, such as audio source separation [23][24][25] and sound source localization [26][27]. Therefore, since audio signals are sparse in the time-frequency domain, it is common to assume that the sources are *W-disjoint Orthogonal* (WDO), i.e. any time-frequency point is only occupied by one of the sources in the mixture [27]:

$$S_n(k,m)S_{n'}(k,m) = 0, \quad \forall(k,m), \forall n \neq n'. \qquad (17)$$

Although this assumption seems to be very strong, many works have shown the WDO properties of audio signals, including speech and music sources [28][29]. In our spatialization context, we also take profit of these properties, which will allow us to redistribute the sources in the stereo panorama without the need for separating them from the mixture or localizing them individually. Figure 4(a) shows the magnitude STFT of a mixture of four music instruments panned at different directions, specified in Table I. It can be easily observed the sparse structure of the mixture, since most of the transformation coefficients are below -30 dB.

4.2. Analysis of the Stereo Panorama

The analysis of the magnitude differences between the left and right channels over the time-frequency domain provides us with very useful data that represents the sound distribution in terms of virtual spatial azimuth, i.e. the composition of the stereo panorama in the mix. In

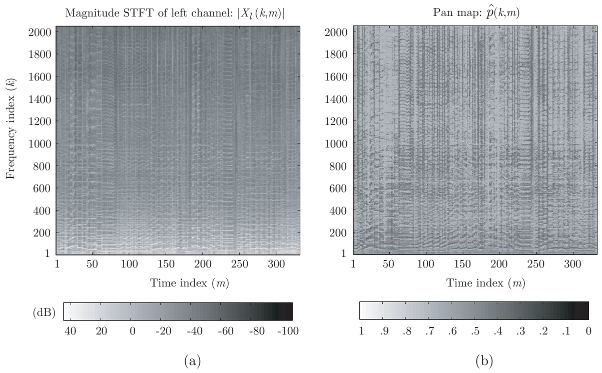


Fig. 4: Mixture of four music sources and its pan map. (a) Magnitude STFT of the left channel. (b) Time-frequency pan map calculated from the STFT of the input channels.

Table 1: Details on the example stereo mixture

Example Mixture	
Signal length	15.5 s
Sample frequency	44.1 kHz
Sources	Panning (p)
Sax	0.4
Guitar	0.5
Accordeon	0.6
Violin	0.7
STFT Processing	
Window type	Hann
Window length	4096
Overlap	50 %

fact, we call a bounded panorama map (or pan map [30]) to the time-frequency representation of the spatial distribution of the sources in the virtual stereo space, which can be calculated as:

$$\hat{p}(k,m) = \frac{2}{\pi}\arctan\left(\frac{|X_r(k,m)|}{|X_l(k,m)|}\right). \qquad (18)$$

Under the WDO assumption and considering the mixing model given by Equations (15) and (16), the ratio $\frac{|X_r(k,m)|}{|X_l(k,m)|}$ in Equation (18) is equivalent to $\frac{a_{ln}(\theta_n)}{a_{rn}(\theta_n)}$, i.e. the mixing ratio of the dominant source at time-frequency point (k,m). Since $\hat{p}$ is bounded between 0 (source totally panned to the left channel) and 1 (source totally panned to the right channel), it becomes clear that the pan map represents the estimations of the panning parameter (see Section 2.2.2) in the time-frequency plane. The pan map of the example mixture is shown in Figure 4(b). A histogram showing the distribution of estimates in the stereo panorama is shown in Figure 5(a).

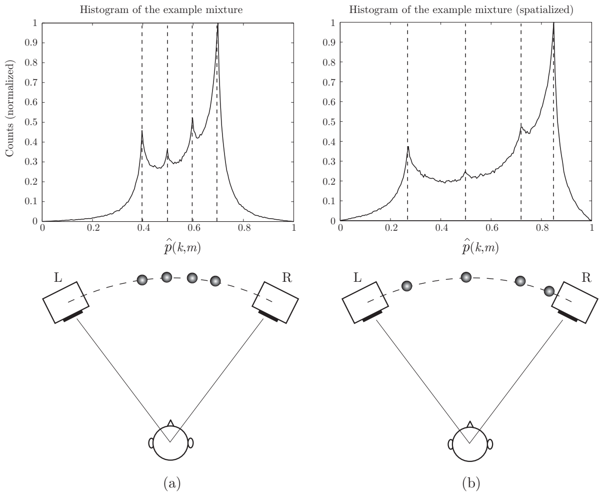


Fig. 5: Histograms of the example mixture before and after being processed together with the ideally perceived phantom sources. (a) Histogram before processing. (b) Histogram after spatialization

4.3. Stereo Spatialization

As previously explained, the values of $\hat{p}(k,m)$ are estimations of the virtual azimuth position of the dominant sound source at time-frequency point (k,m). The virtual azimuth position of a sound source is determined by the mixing process. Our aim is to modify the virtual source locations after the mix blindly, without any a priori knowledge relative to the mix. Moreover, we want to modify this according to a stereo enhancement objective, i.e. intending to spread the stereo field into the maximum degree arc in front of the listener (180 degrees for headphones or $2\theta_0$ for loudspeakers). Another criterion is that associated with the original mixing criterion of the sound engineer: the order of the sources in the stereo panorama should be preserved, i.e. the stereo scene must be widened, but the source arrangement must be consistent with the original recording. With this purpose, a non-linear warping function $\mathscr{T}(x,\rho)$ is introduced:

$$\mathscr{T}(x,\rho) = \operatorname{sign}(x)\frac{|x| + \frac{1}{\rho}|x|}{2|x| + \frac{1}{\rho} - 1}, \qquad (19)$$

where ρ is a user-defined parameter so that $0 > |\rho| \leq 1$. This parameter sets the degree of transformation (aper-

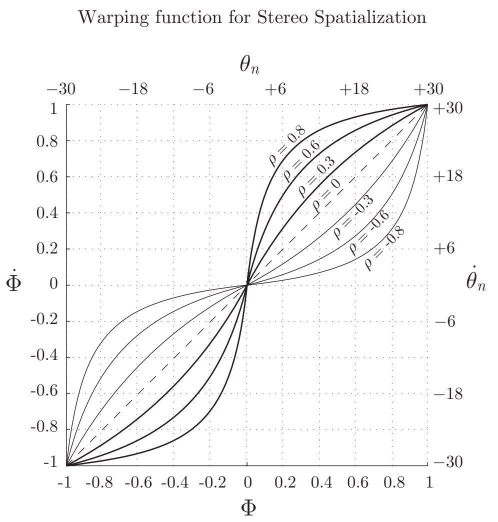


Fig. 6: Relation between original values of $p(k,m)$ and modified ones $\dot{p}(k,m)$ for different aperture parameter ρ. The correspondence with the original and modified perceived sound angles θ_n and $\dot{\theta}_n$ is also represented in the top and right axes.

ture) and the function sign(x) is used to denote the sign of a real number $x \in [-1,1]$. We apply this warping function to the estimated pan map values, which requires a previous mapping of $\hat{p}$ to the range $[-1,1]$:

$$p(k,m) = 2\hat{p}(k,m) - 1. \tag{20}$$

Then, the mapped values p can be transformed into the new ones by applying the transformation as follows

$$\dot{p}(k,m) = \mathscr{T}(p(k,m),\rho). \tag{21}$$

To carry out the re-spatialization of the sources, the modified values $\dot{p}(k,m)$ are mapped back to the range of the panning parameter p and the energy of the time-frequency points is selectively panned to the output similarly to Equations (7) and (8). The inverse mapping is easily performed as follows:

$$\dot{p}(k,m) = (1/2)(\dot{p}(k,m) + 1). \tag{22}$$

As shown by Figure 6, the transformation function is defined so that positive values of ρ broaden the aperture of the original stereo mix, while negative values narrow the spatial distribution of the input stereo sound.

Finally, the output channels are obtained using

$$Y_l(k,m) = \cos\left(\frac{\pi}{2}\dot{p}(k,m)\right) E(k,m) e^{j\angle X_l(k,m)}, \tag{23}$$

$$Y_r(k,m) = \sin\left(\frac{\pi}{2}\dot{p}(k,m)\right) E(k,m) e^{j\angle X_r(k,m)}, \tag{24}$$

where $E(k,m)$ stands for the energy of the stereo input at time frequency bin (k,m):

$$E(k,m) = \sqrt{|X_l(k,m)|^2 + |X_r(k,m)|^2}. \tag{25}$$

Note that the phases of the original input channels are not modified. Once the output signals have been calculated, they can easily transformed back to the time domain using the inverse STFT operator using an overlap-add scheme. The histogram of the output stereo signal corresponding to the example using an aperture factor of $\rho = 0.6$ and the representation of the corresponding virtual phantom sources are represented in Figure 5(b).

4.4. Spatialization Error

The previous subsections were centered on the description of the proposed spatialization technique. Although this technique is based on the WDO assumption, in real-world mixtures not every time-frequency point corresponds to a single source. In fact, the energy of a single point is usually shared between a dominant (or target) source s_T and an interference source s_I. Therefore, an error is committed when estimating the pan parameter of the target dominant source:

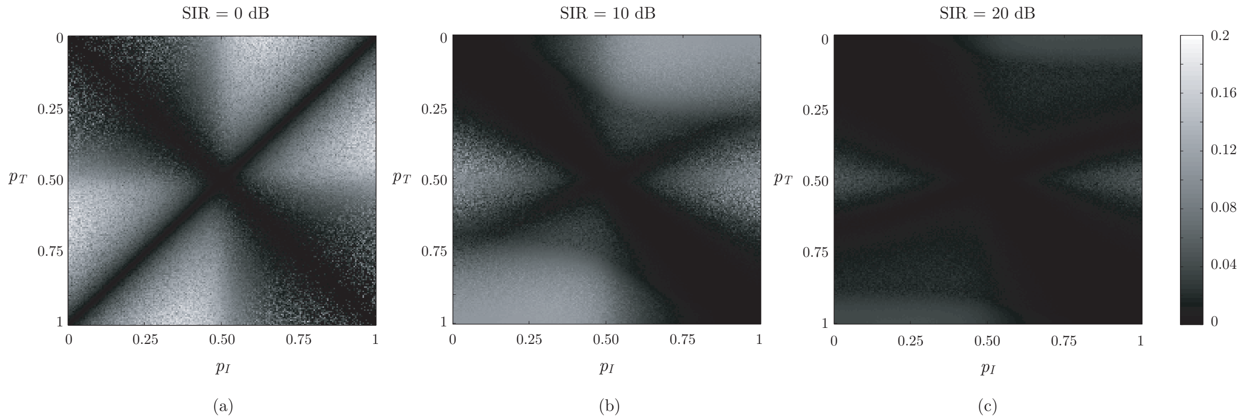
$$e(k,m) = \hat{p}(k,m) - p_T(k,m), \tag{26}$$

where $p_T(k,m)$ is the pan parameter of the target source. This error is dependent on the mixing parameters of the target and interference signals and their relative energy and phases. When the spatialization effect is applied, the transformation is based on the azimuth estimations given by $\hat{p}$, and the error e is forwarded to the output. This means that the transformed pan parameter $\dot{p}$ does not correspond exactly to the value of $\hat{p}$ that would result if the sources had originally been mixed accordingly to the transformed spatial locations:

$$\dot{e}(k,m) = \dot{p}(k,m) - \dot{p}_T(k,m), \tag{27}$$

where $\dot{p}_T(k,m)$ denotes the new pan parameter of the target source after spatialization.

Figure 7 shows the mean absolute error $|\dot{e}|$ for $\rho = 0.5$ as a function of the spatial configuration of the target



(a) (b) (c)

Fig. 7: Absolute error in $\dot{p}$ as a function of the spatial configuration of the mixed signals using $\rho = 0.5$. (a) SIR $= 0$ dB. (b) SIR $= 10$ dB. (c) SIR $= 20$ dB.

and interference signals (with panning parameters denoted as p_T and p_I, respectively). The graphs have been generated by considering the addition of random-phase phasors spatially distributed according to their simulated panning configuration, both representing the contribution of the target source and the interference source in a time-frequency point. The contributions of both signals are defined by means of the *Signal to Interference Ratio* (SIR), defined as SIR $= 20\log_{10}\left(\frac{|s_T|}{|s_I|}\right)$. Note that the error is higher when the spatial distance between the target and interference sources is big. Nevertheless, the mean error is quite small even for high SIR values which results in perfect quality reproduction when the transformation is not very severe $(\rho < 0.8)$.

5. EXPERIMENTS

An informal evaluation campaign was carried out in the studio using a group of 7 expert listeners formed by 3 musicians, 1 mixing engineer and 3 people involved in audio research. A set of 20 fragments corresponding to different music styles (pop-rock, classical, jazz, hard rock, electronic) and stereo recording techniques (amplitude panning, XY stereo, MS stereo) were considered. A set of example tracks[1] before and after processing can be downloaded from http://dl.dropbox.com/u/2647896/examples.zip. Both headphones and loudspeakers were used as stereo reproduction systems in each song.

To simplify the interaction between the listeners and the spatialization software, the proposed approach was presented as an VST plug-in, which can be easily used in conjunction with popular sound recording and editing packages. A screen-shot of the plug-in window is shown in Figure 8. The basic enhancer effect is controlled by means of the ”threshold” fader, which changes smoothly the value of ρ in the spatial transformation function shown in the *Stereo Spatialization* panel. The effect can be switched on/off using the down-left button of the panel. To visualize in real-time the distribution of the panorama estimates before $(\hat{p})$ and after $(\dot{p})$ processing, two graphs are shown to the user in the *Panorama Analyzer* panel. No more functionalities were added at this time to reduce the learning time of the listeners when using it for the first time.

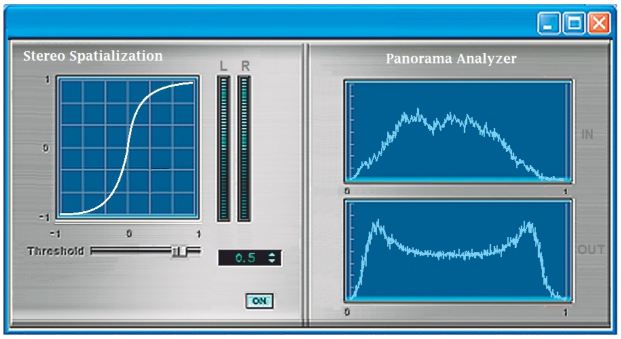

Threshold 0.5

Fig. 8: Screen-shot of the VST spatialization tool

The listeners were asked to experience themselves with

[1] Since the full tracks are copyrighted, only short fragments are provided

the plug-in as long as they wanted using the test sound library: turning the effect on/off, changing the degree of spatialization applied, and selecting between loudspeaker and headphone listening. After experimenting with each test track, they provided a verbal description of the effect and rated the overall quality of the output sound regarding spatial and timbral aspects. A 7-point grade scale was used with this purpose, ranging from -3 to $+3$, where negative/positive values indicate that the quality is worse/better than the original in the following intensity scale:

- -3: Substantially Worse.
- -2: Moderately Worse.
- -1: Slightly Worse.
- 0: Equal.
- +1: Slightly Better.
- +2: Moderately Better.
- +3: Substantially Better.

5.1. Results

In general, all the listeners reported an increased stereo aperture when using the effect that did not substantially affect the timbral quality of the mixture. They most agreed that, depending on the input music track, the test effect added a valuable spatial impression that could perceptually enhance the input sound. As expected, they all reported that the effect was more easily perceived with headphones than with loudspeakers. It is also worth to mention that many listeners reported that the stereo panorama center became emptier when the effect was applied with a very high threshold (ρ) value. The mean results for all the subjects and all test items are shown in Figure 9, together with their 95% confidence intervals. It can be clearly seen that the spatial properties are considerably enhanced using the proposed technique (Figure 9(a)). On the other hand, the overall timbral quality is not affected, which means that increasing the spatial impression does not produce perceivable artifacts or distortion.

6. ACKNOWLEDGEMENTS

The Spanish Ministry of Science and Innovation supported this work under the project TEC2009-14414-C03-01.

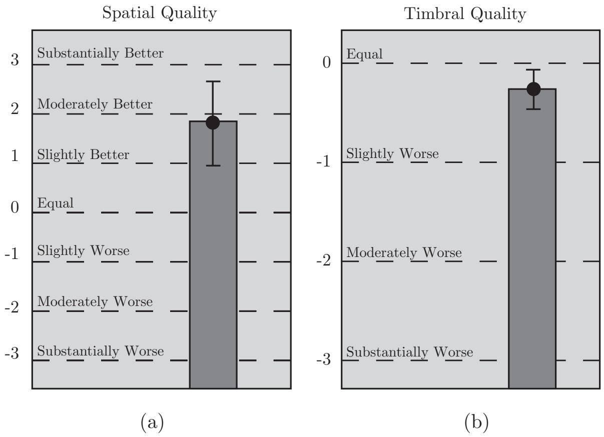


Fig. 9: Mean and 95% confidence intervals obtained from the scores given by the subjects.

7. REFERENCES

[1] J. Blauert, *Spatial Hearing. The Psychophysics of Human Sound Localization*. MIT Press, 1996.

[2] G. Gatzsche and F. Melchior, “Spatial audio authoring and rendering: Forward research through exchange,” in *Proceedings of the International Computer Music Conference (ICMC2008)*, Belfast, UK, August 2008.

[3] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools*. Oxford, UK: Focal Press, 2008.

[4] A. D. Blumlein, “Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems,” British Patent 394,325, 1933.

[5] V. Pulkki, “Spatial sound generation and perception by amplitude panning techniques,” Helsinki University of Technology, Helsinki, Finland, Tech. Rep., 2001.

[6] G. Theile, “On the naturalness of two-channel stereo sound,” *Journal of the Audio Engineering Society*, vol. 39, pp. 761–767, October 1991.

[7] W. B. Snow, “Basic principles of stereophonic sound,” *Journal of the Society of Motion Picture and Television Engineers*, vol. 61, no. 11, pp. 567–589, 1953.

[8] J. M. Eargle, Ed., *AES Anthology: Stereophonic Techniques*. New York: Publications of the Audio Engineering Society, 1986.

[9] B. Bartlett, *Stereo Microphone Techniques*. Focal Press, 1991.

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

[11] R. Rabenstein and S. Spors, *Springer Handbook of Speech Processing*. Springer, 2008, ch. Sound Field Reproduction, pp. 1095–1113.

[12] R. C. Maher, "Old and new techniques for artificial stereophonic image enhancement," in *Proceedings of the 101st Convention of the Audio Engineering Society*, Los Angeles, CA, USA, 1996.

[13] A. Klayman, "Stereo enhancement system," U.S. Patent 4,748,669, 1988.

[14] G. J. Barton, "Signal enhancement method for stereo system," U.S. Patent 4,910,778, 1990.

[15] S. W. Desper, "Automatic stereophonic manipulation system and apparatus for image enhancement," U.S. Patent 5,412,731, 1995.

[16] M. Dressler, "Dolby Surround Pro Logic II decoder principles of operation," Dolby Laboratories Information, 2000.

[17] C. Avendano and J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *Proceedings of the AES 22nd Conference on Virtual, Synthetic and Entertainment Audio*, 2002, pp. 121–130.

[18] F. Baumgarte and C. Faller, "Binaural cue coding - part i: Psychoacoustic fundamentals and design principles," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 509–519, 2003.

[19] C. Faller and F. Baumgarte, "Binaural cue coding - part ii: Schemes and applications," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 520–531, 2003.

[20] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, June 2007.

[21] ——, "Directional audio coding in spatial sound reproduction and stereo upmixing," in *Proceedings of the AES 28th International Conference*, Pitea, Sweden, July 2006.

[22] L. Cohen, *Time-Frequency Analysis*. Prentice Hall, 1995.

[23] J. J. Burred, "From sparse models to timbre learning: New methods for musical source separation," Ph.D. dissertation, Technical University of Berlin, 2008.

[24] M. D. Plumbey, S. A. Abdallah, T. Blumensath, M. G. Jafari, A. Nesbit, E. Vincent, and B. Wang, "Musical audio analysis using sparse representations," in *Proceedings of 17th COMPSTAT 2006*, Rome, Italy, August-September 2006.

[25] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 91–98, 2006.

[26] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Performance evaluation of sparse source separation and DOA estimation with observation vector clustering in reverberant environments," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Paris, France, 2006.

[27] S. Rickard and F. Dietrich, "DOA estimation of many w-disjoint orthogonal sources from two mixtures using DUET," in *Proceedings of the 10th IEEE Workshop on Statistical Signal and Array Processing (SSAP2000)*, Pocono Manor, PA, August 2000, pp. 311–314.

[28] S. Rickard and O. Yilmaz, "On the w-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics*, Speech, and Signal Processing, pages 529-532, Orlando, Florida, May 2002.

[29] J. J. Burred and T. Sikora, "On the use of auditory representations for sparsity-based sound source separation," in *Proceedings of the 5th International Conference on Information, Communications and Signal Processing (ICICS 2005)*, Bangkok, Thailand, December 2005.

[30] M. Cobos and J. J. Lopez, “Stereo audio source separation based on time-frequency masking and multilevel thresholding,” *Digital Signal Processing*, vol. 18, no. 6, pp. 960–976, 2008.