A CROSS-ADAPTIVE DYNAMIC SPECTRAL PANNING TECHNIQUE

Pedro D. Pestana, *
CITAR - UCP,
R. Diogo Botelho, 1327, 4169-005 Oporto
(also ILID - ULL and CEAUL - UL)
ppestana@porto.ucp.pt

Joshua D. Reiss, †
C4DM,
Queen Mary University of London,
Mile End Road, London E1 4NS
josh.reiss@eecs.qmul.ac.uk

ABSTRACT

This work presents an algorithm that is able to achieve novel spatialization effects on multitrack audio signals. It relies on a cross-adaptive framework that dynamically maps the azimuth positions of each track's time-frequency bins with the goal of reducing masking between source signals by dynamically separating them across space. The outputs of this system are compared to traditional panning strategies in subjective evaluation, and it is seen that scores indicate it performs well as a novel effect that can be used in live sound applications and creative sound design or mixing.

1. INTRODUCTION

Recent work on adaptive digital audio effects has seen the emergence of a new class of cross-adaptive systems that aim to do automatic or computer assisted mixing (see [1] for a review). The architecture is one that allows for a mapping of all concurrent signals in a mix to determine the processing parameters on each of the individual inputs in order to optimize either a perceptual (e.g. loudness balance, see [2]) or objective (e.g. release from masking, see [3]) characteristic of the mixed output signal. In the present work we focus on panning as a tool to overcome masking problems, as done in [4] and others, but following a radically different approach that focuses in the creation of a novel type of tool.

Previous approaches have relied on grounded theory, trying to mimic the decisions that a human sound engineer would undertake. We propose that an interesting area of exploration for intelligent systems is to strive for processing techniques that are prohibitively difficult to achieve with traditional means and by human practitioners. Being non-standard, there is a good chance that results will be unconventional and careful analysis and subjective evaluation is paramount.

The approach we are suggesting follows close on [5], which looks at the possibilities of adaptive digital audio effects, but not specifically the problem of mixing. In that work the authors proposed a method for panning different frequency bins of a signal into different azimuthal positions, where the mapping decisions for panning placement may come from the analysis of a separate signal. We extend the idea so that the azimuthal mapping optimizes masking constraints arising from the need to sum multiple individual tracks in a mix, and that this can be extended to a time-varying system that will achieve a new approach to spatialization and spatial unmasking optimization. A similar approach is done in terms of frequency-unmasking in [6], and DirAC [7] touches on similar concepts, though it aims at 'transparent' reproduction, instead of acting as a cross-adaptive effect.

In Section 2 we elaborate on the theoretical reasons for why a spectral audio panner can be a viable tool, and the concepts on which spatial audio is built upon. Section 3 presents a detailed description of our cross-adaptive algorithm, while Section 4 presents an objective analysis of results in selected samples. In Section 5 we summarize a subjective evaluation performed on examples. Finally, Section 6 points to further directions and application, and provides an overview of our results.

2. THEORETICAL MOTIVATION

Spatialization depends on Interaural Level Difference (ILD) and Interaural Temporal Difference (ITD) cues, but the relative importance of each is still not fully understood. The former works mainly at high frequencies, where the head casts an acoustic shadow that is large enough to attenuate the level reaching the contralateral ear. On the other hand, the latter is predominantly an active mechanism for low-mid frequencies, where the phase difference can be resolved by the brain. At really low frequencies, sounds seem to have an enveloping place of origin. ITDs and ILDs are often distorted and conflicting, and the auditory mechanism uses the most consistent cue; the one that suggests the same direction over a broad spectral band [8].

According to Griesinger [9], with broadband sources cues are ambiguous, and human hearing appears to simply average over the various possible sound directions to determine the best-guess position for the source, while weighting-in past history. The brain seems to use mechanisms other than localization to stream sounds together and only afterward does it assign a common place of origin. Furthermore, Blauert [10] argues that in principle, within a critical band, all sound can only be perceived as a single source of wider or narrower character. To this extent, Pulkki [11] concludes that spatial realism is not needed in audio for the resulting image to have any verisimilitude.

Modern audio production relies on amplitude panning techniques almost exclusively for the creation of azimuthal cues out of monophonic source signals. In this work we shall ignore cues stemming from signal delay, though a translation of the technique could trivially be achieved. It is typical to distribute sound sources among the reproduced stage, as the spatial release from masking (SRM) that is achieved improves clarity and intelligibility. The relationship of ITD and ILD to SRM is not fully studied, but it seems level is panning is a sensible choice [12].

* The author P.D. Pestana was sponsored by national funds through the Fundação para a Ciência e a Tecnologia, Portugal, in projects: ''PEst-OE/EAT/UI0622/2014' and 'PEst-OE/MAT/UI2006/2014'.

† The author is supported by EPSRC grant EP/K007491/1, 'Multisource audio-visual production from user-generated content'

3. ALGORITHM AND CONSTRAINTS

Our algorithm is based on the typical phase vocoder implementation, and can be outlined as follows:

1. Perform a Short-Time Fourier Transform (STFT) on all the input tracks of an audio mix.
2. Pan each resulting time-frequency (t-f) bin on each track independently, placing the heaviest (magnitude-wise) bins of each track to non-colliding locations.
3. Perform the Inverse Fourier Transform to reconstruct the time-domain signals of the mix.

An audio mix is the result of a summation of an arbitrary number J of input tracks. Let us call each individual track x_j and assume out of simplicity that it is monophonic (single-channel) and that all tracks are equal in length. Let us define our notation and establish that the mix's time-frequency representation is then given by the STFT [13]:

$$X_j\left(n,k\right)=\sum_{m=-\infty}^{\infty}x_j\left(m\right)h\left(n-m\right)e^{-i2\pi km/N}, \quad (1)$$

with n indexing discrete time, k the discrete frequency bin and $h(n-m)$ a window function of length N. This results in a complex number at each element, which can then be decomposed into magnitude and phase angle. Consider:

$$Xmag_j\left(n,k\right)=\sqrt{\mathrm{Re}[X_j\left(n,k\right)]^2+\mathrm{Im}[X_j\left(n,k\right)]^2} \quad (2)$$

to be the 3-dimensional matrix of individual magnitudes for each t-f bin n-k of each track j. The phase angle will actually not be important for our application, but it is given by the inverse tangent of the ratio of imaginary to real part, for each element of the matrix. It is both prohibitive and irrelevant to perform the STFT at every point in discrete time, so one uses a fixed window length (N) overlapping with a hop size (I) which is a subdivision of N, and uses time frames that start at nI and are N samples long. A Hamming-windowed STFT with a hop size of $N/2$ will yield perfect reconstruction upon performing the inverse transform and adding all the individual frames. It is now clear that the matrix whose elements are described by equation 1 is $J\times G\times N$ elements long, where G is the signal length zero-padded to the next multiple of I and divided by I, and N, being the window length, is also the number of bins in the spectral domain.

Prior to reconstruction, our goal in the spectral domain is to perform a readjustment of each bin so that it yields two different results for a left (L) and a right (R) channel:

$$X_j\left(n,k\right)\longrightarrow\left\{\begin{array}{l}{Y_j}^{(L)}\left(r,k\right)\\{Y_j}^{(R)}\left(r,k\right)\end{array}\right. . \quad (3)$$

with r the outbound time-frame. For all practical purposes, we shall have $r=n$, since our analysis hop size is equal to the synthesis hop size. Reconstruction of channel Z (either L or R) can then be performed by overlap-adding the windows [13]:

$${y_j}^{(Z)}\left(n\right)=\sum_{r=-\infty}^{\infty}h\left(n-r\right)\frac{1}{N}\sum_{k=0}^{N-1}\left[e^{i2\pi rk/N}{Y_j}^{(Z)}\left(r,k\right)\right]e^{i2\pi nk/N} , \quad (4)$$

and scaling the output so that the overlap of a large number of hops does not increase the synthesized amplitude too much. Note that while the notation quickly becomes heavy, this is simply the strategy behind the well-known phase vocoder approach to spectral domain processing and equation 4 simply represents the summation of the Inverse Fourier Transform of all individual time-frames. The core part of our research is not the analysis-synthesis process, but how to implement the mapping in (3).

A viable reconstruction of a spectrally processed signal depends upon a careful choice of windowing parameters. For our case, with a short window one would likely encounter amplitude modulation artifacts, while a long window would cause temporal aliasing and smearing. A similar balance results from the hop size; allowing for no overlap or little overlap results in clicking noises and too much overlap will cancel out the desired effect, restricting panning azimuth to a very narrow area. An heuristic approach convinced us that a window length of 2^{16} (working at a $44.1\ kHz$ sample rate; one update approximately every 1.49 seconds) with a hop size of one sixteenth the window length yields sonically acceptable results. There is temporal smearing that makes signals sound as if they had been put through a nonlinear reverb, but the amount is subtle enough so that it is still pleasing.

Each track's t-f bins can now be placed at position $p_j(n,k)\in[-1,1]$, where -1 represents full left and 1 full right. This position will then be converted to a specific azimuthal angle dependent on the position of the speakers (the standard stereophonic situation is defined by speakers at $\pm30^\circ$ from the median plane). The sine-cosine rule for amplitude panning states that we should have gains for the left and the right channel that follow:

$$g^{(L)}=\cos\left(\frac{\left(1+p_j(n,k)\right)\pi}{4}\right), \quad (5)$$

$$g^{(R)}=\sin\left(\frac{\left(1+p_j(n,k)\right)\pi}{4}\right). \quad (6)$$

We know from [10] that azimuthal discrimination is more accurate near the origin at the median plane than when we drift to the sides. It is thus sensible to envisage a position distribution scheme that mimics this perceptual phenomenon. We chose to allow for J discrete possible placements (as many as the track count) and balance them according to a variation on the roots of the first order Chebyshev polynomials, which gives us a well-behaved distribution. The root calculation is done through:

$$P_g=\mathrm{sign}\left[\cos\left(\frac{2g-1}{n}\pi\right)\right]-\cos\left(\frac{2g-1}{n}\pi\right), g=1,2...n, \quad (7)$$

Notice we are shifting the nodes so that they are center-heavy instead of tail heavy (the sign function is -1 for negative values, 1 for positive values and 0 for the value zero). We then transform $\{P_1,P_2,...,P_J\}\longrightarrow\{P_{1:n},P_{2:n},...,P_{J:n}\}$ by simply sorting the results, where $t_{k:n}$ denotes the k-th ascending order statistic. This gives us fixed position to which we shall map our J tracks differently for each t-f bin.

Figure 1 shows the constrained positions for four possible track counts. Notice that positions at the extreme left or right are never allowed, something that is considered a good practice by some sources [14]. Two other possible positional distributions were considered: equidistant spacing and the non-constrainment to discrete angles (relying instead on the energy distribution of each bin to achieve symmetrical balance). Informal testing showed that our

choice yields better perceptual envelopment than equidistant spacing and more stability than not having fixed discrete points (panning positions are less prone to large changes from frame to frame).

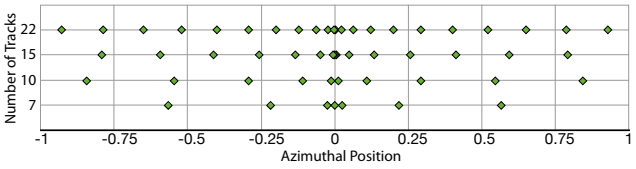

Figure 1: Allowable azimuthal positions P for four possible track counts J, according to our Chebyshev-roots based constraint.

The decision regarding which position to choose for each bin must be bound by some constraints, following [4, 14, 15]:

1. For each bin the sum of the individual track amplitudes contributing to the left and right signal must be balanced in terms of magnitude:

$$\sum_{j=1}^{J} p_j(n,k)\,|X_j\,(n,k)| = 0, \forall n,k. \tag{8}$$

2. The sum of the magnitudes of all weighted frequency bins for a fixed time-frame must be equal at all chosen azimuthal steps:

$$\sum_{k=0}^{N-1} p_j(n,k)\,|X_j\,(n,k)| = \alpha, \forall j,n, \tag{9}$$

where α is a time-frame-varying constant.

3. Successive time frames cannot allow for a frequency-azimuth pair to shift more than a maximum azimuthal step-size. We have restricted the step size to two after some informal listening tests.

4. Bins representing frequencies below a cut-off point should be centrally panned, in keeping with most industry practices. We have chosen that frequency to lie in the vicinity of $150\ Hz$.

It is impossible to enforce constraints 1 and 2 when bounded by discrete positions and the need to fulfill constraint 3, so we use those as lax rules that provide a per-frame platonic ideal, that is then revised by the need to restrict sudden movement. We establish the positions for the first time-frame by applying a palindromic Siegel-Tukey type ordering [16] to each bin across tracks, according to ordered magnitude, and continue in the following time-frames with a greedy algorithm. The starting point choice and iterative proceeding rules are best explained by example:

Suppose we have a five track mix and that the first time-frame results in the following magnitude vectors:

$$\begin{cases} \mathbf{Xmag}_1\,(1,k) = \{0.2, 0.3, 0.25...\} \\ \mathbf{Xmag}_2\,(1,k) = \{0.3, 0.2, 0.15...\} \\ \mathbf{Xmag}_3\,(1,k) = \{0.4, 0.1, 0.35...\} \\ \mathbf{Xmag}_4\,(1,k) = \{0.6, 0.02, 0.6...\} \\ \mathbf{Xmag}_5\,(1,k) = \{0.8, 0.01, 0.5...\} \end{cases} \tag{10}$$

For the first bin (ignoring the fact that constraint 4 would force us not to use panning) track 5 has the highest magnitude value, followed by 4, 3, 2, 1. These would then be panned respectively to p_1, p_5, p_4, p_2, p_3. This ordering would place heavier track-bins towards the extremes, yet would tend to enforce constraint 1.

For the second bin, track 1 has the highest magnitude value, followed by 2, 3, 4, 5. They would be panned respectively to p_3, p_2, p_4, p_5 and p_1. This reverse Siegel-Tuckey ordering will help enforce constraint 2. We can move to bin 3 and simply shift the order so that it starts with the heaviest bin at p_5 and bin 4 will start with p_3 but move to p_4 and bin 5 will have equal positioning possibilities to bin 1. Thus, this part of the algorithm works as a 4-step mapping climbing up the frequency bins.

The second time-frame is first planned in a similar fashion, but we do not allow for individual bin shifts of more than two positions between consecutive time-frames. The target goal will sometimes become unattainable, and in such cases we try to minimize the least square errors between the intended position vector and the possible position vector. There is one special feature that we can use to optimize which is the symmetry of ordering: symmetrical azimuths about zero are equal terms of weight for our constraints 1 and 2. So if we see an intended shift of a specific bin from p_1 to p_4 in two successive time frames, constraint 3 would tell us to move no farther than p_3 (two steps), but given that p_4 is symmetrical to p_2, thus p_2 would present a better move. The programmatic implementation of the greedy algorithm is consequently messy and more prone to be described then notated. Since for each new frame we always calculate our pseudo-optimal positions, the actual overall placement never diverges too far from the intended one.

Finally, with the intended position for each t-f bin in mind, there are two possible approaches to performing the calculations: multiplying $X_j(n,k)$ (the complex spectrum as a whole, not the individual magnitudes) by the gains that were obtained ($g^{(L)}$ and $g^{(R)}$) or zeroing all t-f bins that are not going to be panned to a specific position for each track, doing the Inverse Fourier Transform of the result, and panning on the time domain. We chose to do the former for our examples, but the latter will be explored in the future. There are some audible examples of the algorithm's working at http://www.stereosonic.org/phd/specPanning/, alongside spectrograms that illustrate the change it brings about.

4. RESULT ANALYSIS

Figure 2 shows the intended positions for a segment of a five-track mix. The five discrete azimuthal positions are described in shades of gray from maximum-left (white) to maximum-right (black).

This is an illustrative example of something that is quite hard to visualize but serves to show 1) that pan position is different for each frequency bin of each track, when considering a fixed time-frame (if one looks at any column on any track), 2) that pan positions of each bin change over time (if one looks at any row on any track), 3) that there is some degree of inertia for each time frame, frequency bin (i.e, a row or column will not change position too drastically) and 4) Drums and Synth tend to be either full-left or full-right, the other tracks are much more inert.

An algorithm that proposes to work as a novel audio effect is quite hard to assess objectively. It is important to understand how closely the constraints 1 and 2 are met, in light of the fact that our rules are lax. We have looked at four multitrack songs in order to understand deviations from what is expected. We are dealing with multi-dimensional phenomena so there is little tangible feeling for how big a deviation can be and how to measure it. We use the following two metrics to determine how well we match the constraints. For constraint 1, we find the discrete frequency and

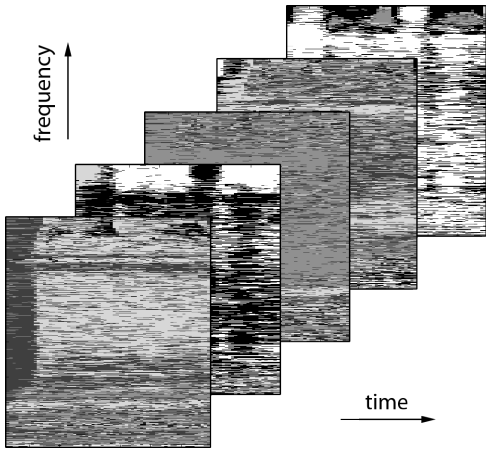

Figure 2: Intended pan positions for 5 tracks (front-to-back: bass, drums, guitar, vocals, synth. The time segments are always the same 5-seconds. White means left and black right.

discrete time mean of the absolute weighted azimuth deviations from zero:

$$\frac{\sum_n \sum_k \left| \sum_j p_j(n,k) \left| X_j(n,k) \right| \right|}{G \times N}, \tag{11}$$

where, as before, N represents the window length, and here is used as the number of bins, and G is the number of time windows used on the STFT.

For constraint 2, we find the mean over discrete time of the standard deviation between azimuthal magnitudes:

$$\frac{\sum_n s_{dev} \left(\sum_k |X_j(n,k)|_{|_{p_j(n,k)=P_g}} \right)}{G} \tag{12}$$

Table 1 shows the results for the four song segments ($\approx 20\ s$), for three different approaches. The benchmark approach is random panning of the t-f bins. We compare that with following our algorithm with a window length of 2^{15} and a hop size of $1/16$ and a window length of 2^{10} with a hop size of $1/2$.

Approach	Song #1	Song #2	Song #3	Song #4
Random panning	4.957	6.148	4.469	6.065
Window 2^{15}, hop 1/16	0.024	0.019	0.018	0.024
Window 2^{10} hop 1/2	0.077	0.101	0.992	0.140

Approach	Song #1	Song #2	Song #3	Song #4
Random panning	1.664	2.167	1.857	1.935
Window 2^{15}, hop 1/16	0.325	0.143	0.464	0.656
Window 2^{10} hop 1/2	1.431	0.542	1.161	1.548

Table 1: Approximation to constraints 1 (top) and 2 (bottom). Top table values are multiplied by 10^3, bottom table by 10^{16}.

The values show that constraint 2 will always be approached as a result of the law of large numbers, even for the case of random panning. However, both our approaches achieve a better result than a random strategy. A large window size will yield more frequency bins, which will likely smooth the variations at each position, so its relatively better score comes as no surprise. As for constraint 1, the algorithm shows a clear improvement against random panning. Here the larger window also seems to work consistently better, most likely because of the larger overlap, stemming from the hop size.

5. SUBJECTIVE EVALUATION

In order to understand whether the approach is worthwhile we performed a subjective multi-stimulus evaluation. Twenty subjects of moderate experience with audio engineering took place. The test signals were presented via headphones at a consistent listening level ($83\ dB$) through a steady signal chain. The use of headphones was a compromise, as the technique is better suited for a loudspeaker test; however, repeatability in different settings was important, so a compromise was chosen. Tests were double blind, using 4 different versions of 4 different songs (20 second segments). Versions of one such song can be heard online at http://www.stereosonic.org/phd/specPanning/. Procedures followed closely [17] and the different stimuli were:

1. A monophonic version that served as the base for the algorithm. Loudness balance done by a mixing engineer. This balance choice was kept throughout (M).
2. A traditional static stereophonic version, where panning decisions were done by a professional mixing engineer (S).
3. A spectrally panned version with our best time constants: window length of 2^{15} with a $1/16$th hop (Lg, for 'long window').
4. An anchor, a spectrally panned version with a time constant that some preliminary tests had shown problematic: window length of 2^{10} with a $1/2$ hop (Sh, for 'short window').

The order of presentation was randomized, a fact that was explained to the listeners in advance. Subjects were asked to rank the versions according to three parameters (one per listening run for a total of 3×4 runs with 4 versions per run): 'clarity', 'production value' and 'excitement'. The concept of clarity should be associated with the ability to segregate sources, thus with release from masking. Production value is inherently ambiguous and subjective, but because subjects were studying or had studied audio engineering, the meaning should be clear: technical quality of the mix. Excitement was expected to reveal a dimension that is not coupled to the evaluation of quality but of a rawer reaction to the mix.

Results are summarized in Figure 3, showing mean and confidence interval bounds for the three parameters. The professional version is superior in terms of production value perception, yet our algorithm competes in terms of clarity. This is an encouraging result, as we are comparing a trained professional approach to an experimental algorithm that goes dangerously against what would be considered accepted by traditional standards. Our algorithm also yields the best score for excitement, beyond the error bounds, which validates the hypothesis of using it on a mix for its 'special effect' character. The anchor version is perceived as unpleasing, yet it scores fairly well in terms of excitement. A Friedman test for evaluation consistency among users only finds evidence for randomness in terms of production value, which might be explainable by the difficulty in having consensus on what "Production Value" means. Further investigation reveals that subjects are evaluating songs differently in that case. Separating by song, our algorithm is perceived as excelling in production value for one of the four examples (exactly the one that can be found online).

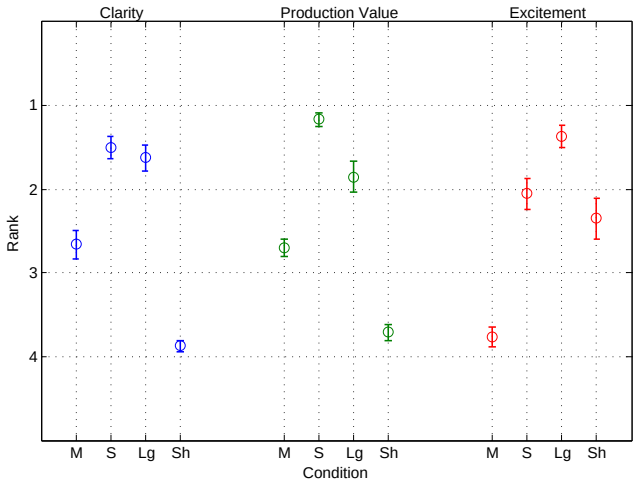

Figure 3: Evaluation results for Clarity (left), Production Value (middle) and Excitement (right), averaging both in terms of songs and subjects. See main text for conditions.

6. CONCLUSIONS AND FURTHER WORK

A technique of spectral unmasking through cross-adaptive dynamic panning is explored. It is to be understood as an experimental effect and not a method that can be used in traditional mixing. An absolute constraint equilibrium cannot be met constantly, and slow variations through time are shown to produce better results. Subjective evaluation confirms the algorithm can be used as a novel effect, resulting in an appreciated level of excitement. It is shown to compete with a professional mix in terms of clarity, so the purpose of unmasking is to some degree accomplished (particularly if one remembers the evaluators will always understand the algorithm as "strange"). Such a radical panning strategy could very well be understood as nonsensical, but results show otherwise, at least when considering the longer time-constant version. It is safe to assume that a listener will not understand that there is an explosion of panned events, and is still very much able to identify each sound source as being one.

A technique such as this can be useful in several scenarios:

1. Whenever amplitude panning of whole sources is frowned upon, such as large outdoor live shows, where elements are kept in the middle. The algorithm would allow audience members on the center to feel a sense of spaciousness while allowing everyone to still have the perception of a full mix.
2. In song sections, where the producer wishes to add a subtle other-worldly feeling to a part.
3. In excessively dense mixes, where the mixing engineer is struggling for clarity.

Several extensions of the idea can be researched in the future if one considers that the azimuthal positional choices on this work were mainly heuristic. There are many explorations on the nature of the constraints, and the reasons for their choice, that can result in clearer and more effective results. The research for a real-time strategy is in the works. A comparison with techniques involving filter-bank methods either in the analysis or synthesis is also an important step.

7. REFERENCES

[1] Joshua D. Reiss, "Intelligent Systems for Mixing Multichannel Audio," in *17th Intl Conf on Digital Signal Processing (DSP)*, 2011.

[2] Enrique Perez Gonzalez and Joshua D. Reiss, "Automatic Gain and Fader Control For Live Mixing," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009.

[3] Enrique Perez Gonzalez and Joshua D. Reiss, "Improved Control for Selective Minimization of Masking Using Inter-Channel Dependancy Effects," in *Proc of the 11th Intl Conf on DAFx*, 2008.

[4] Stuart Mansbridge, Saoirse Finn, and Joshua D. Reiss, "An Autonomous System for Multi-track Stereo Pan Positioning," in *Proc of the 133rd AES Convention*, 2012.

[5] Vincent Verfaille and Udo Zölzer, "Adaptive Digital Audio Effects (A-DAFx): A New Class of Sound Transformations.," *IEEE Transactions on Audio, Speech, and Language.*, vol. 14, no. 5, pp. 1817 – 1831, 2006.

[6] Piotr Kleczkowski and Adam Kleczkowski, "Advanced Methods for Shaping Time-Frequency Areas for the Selective Mixing of Sounds," in *Proc of the 120th AES Convention*, 2006.

[7] Ville Pulkki, "Spatial Sound Reproduction with Directional Audio Coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[8] Ville Pulkki, "Localization of Amplitude-panned Virtual Sources I: Stereophonic Panning," *Journal of the Audio Engineering Society*, vol. 49, no. 9, pp. 739–752, 2001.

[9] David Griesinger, "Stereo and Surround Panning in Practice," in *Proc of the 112th AES Convention*, 2002.

[10] Jens Blauert, *Spatial Hearing: the Psychophysics of Human Sound Localization*, MIT Press, Cambridge, 1997.

[11] Ville Pulkki, T. Lokki, and Davide Rocchesso, "Spatial Effects," in *DAFx*, Udo Zölzer, Ed., chapter 5, pp. 139–180. John Wiley & Sons, Chichester, second edition, 2011.

[12] J. M. Dillon H. Cameron S. Glyde, H. Buchholz and L. Hickson, "The importance of interaural time differences and level differences in spatial release from masking," *Journal of the Acoustical Society of America*, vol. 2, no. 134, pp. 147–152, 2013.

[13] R.E. Crochiere, "A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis," *IEEE Trans. on Speech and Aud. Proc.*, vol. 281, no. 1, pp. 99–102, 1980.

[14] Pedro D. Pestana, *Automatic Mixing Systems Using Adaptive Digital Audio Effects*, Phd, Universidade Católica Portuguesa, 2013.

[15] Enrique Perez Gonzalez and Joshua D. Reiss, "A Real-Time Semiautonomous Audio Panning System for Music Mixing," *EURASIP Journal on Advances in Signal Processing*, 2010.

[16] S. Siegel and John W. Tukey, "A nonparametric Statistics for the Behavorial Sciences," *Journal of the American Statistical Association*, vol. 55, pp. 429–445, 1960.

[17] S. Bech and N. Zacharov, *Perceptual Audio Evaluation — Theory, Method and Application*, John Wiley & Sons, Chichester, 2006.