Audio Engineering Society

Convention Paper 10178

Presented at the 146th Convention
2019 March 20 – 23, Dublin, Ireland

An automatic mixing system for multitrack spatialization for stereo based on unmasking and best panning practices

Ajin Tom[1], Joshua Reiss[2], and Philippe Depalle[1]

[1]*Sound Processing & Control Lab, Centre for Interdisciplinary Research in Music Media & Technology, McGill University, Montreal, Canada*
[2]*Centre for Digital Music, Queen Mary University of London, London, UK*

Correspondence should be addressed to Ajin Tom (ajin.tom@mail.mcgill.ca)

ABSTRACT

One of the most important tasks in audio production is to place sound sources across the stereo field so as to reduce masking and immerse the listener within the space. This process of panning sources of a multitrack recording to achieve spatialization and masking minimization is a challenging optimization problem, mainly because of the complexity of auditory perception. We propose a novel panning system that makes use of a common framework for spectral decomposition, masking detection, multitrack sub-grouping and frequency-based spreading. It creates a well spatialized mix with increased clarity while complying to the best panning practices. Both real-time and offline optimization-based approaches are designed and implemented. We investigate the reduction of inter-track auditory masking using the MPEG psychoacoustic model along with various other masking and spatialization metrics, extended for multitrack content. Subjective and objective tests compare the proposed work against mixes by professional sound engineers and existing auto-mix systems.

1 Introduction

Multitrack mixing is an iterative process in which various processing parameters such as loudness balance, EQing, compression are adjusted to achieve certain target output mix that complies to perceptual and objective criteria [1]. Research into automatic mixing systems has grown rapidly over the last ten years, with intelligent systems proposed for almost every aspect of audio production [2, 3, 4]. Intelligent tools that analyze the relationships between all channels in order to automate the mixing of multitrack audio content have been devised.

Automatic mixing could be classified as follows:
1) achieving an autonomous computerized mix by mimicking the iterative and adaptive approach of a sound engineer,
2) letting an intelligent system achieve a mix using complex processes that are not achievable by a sound engineer in finite time with tools available in a Digital Audio Workstation.

Both approaches can produce desirable mixes complying to certain rules and constraints. Panning has a larger effect in the overall improvement provided by automatic mixing than any of the other tools like autonomous faders and EQ for multitrack [5]. For example, in the case of positioning sources in a stereo field according to the first approach, mix engineers start panning spectrally similar sources by nudging the panning amounts of the individual tracks until a better sense of spatialization, clarity, instruments distinction and unmasking is achieved.

Previous work in the field of intelligent panning systems involves analyzing features from a multitrack recording to determine a panning amount for each track. The premise of [6] is that one of the primary goals of stereo panning is to 'fill out' the stereo field. This algorithm set target criteria as source balancing (equal numbering and symmetric positioning of sources on either side of the stereo field), spatial balancing (uniform distribution of levels) and spectral balancing (uniform distribution of

content within each frequency band). It further assumes that the higher the frequency content of a source, the more it will be panned, and that no hard panning will be applied.

A semi-blind stereo panning system is proposed in [7] in which tracks are given priority such that low priority tracks get placed at wider azimuthal angles. This was to comply to the general practice of placing bass heavy sounds and lead vocals in the centre of the mix.

Some concerns with the above panning techniques are that we may lose a stable centre image [8], harsh panning of an instrument on one side is often not preferred [9], and spectral centroid may not be an ideal descriptor of frequency content [6]. Since the above two techniques pan the track as a whole to either side, low frequency content may be panned, causing unwanted spectral imbalance.

The approach taken in [10] does not try to emulate traditional panning practices, and hence fits the second category of automatic mixing systems. Here, time frequency bins of each multitrack are assigned different spatial positions in the mix. Time-frequency based decomposition and modification techniques have also been used for spatial enhancement in [11].

Unwanted spectral masking is a commonly observed phenomenon that reduces audibility of sounds in multitrack mixing. When the output mix lacks clarity and instrument separation even after loudness balance, EQing and dynamic compression, we are left with no choice but to spatially separate the masked sources. The authors of [12] suggests that panning would remove most of the masking present in a mix. There is much more to explore on stereo positioning that can be carried out to objectively analyze multitrack content.

The panning system we propose involves subgrouping [13] similar tracks within a mix based on correlation of spectral content and each track within the subgroups are panned across the frequency domain using equivalent rectangular bandwidth (ERB)-scale sinusoidal shaped panning filters. The panning filters for the tracks are optimized in order to minimize masking cost function. A limitation of balancing using a panpot [8] is that the sources remain tied to one side and the width may be restricted [9]. These issues are addressed in our algorithm which generates a mix with a stable centre image while still giving a spatialized effect in a wider stereo field. These features are objectively measured using various spatialization metrics and computation-friendly masking models extended for multitrack. We discuss real-time and off-line optimization approaches, both of which make use of the same framework for spectral decomposition, multitrack grouping, masking detection and reduction. We focus on the headphone listening context, though the work may also be applicable to loudspeaker playback.

2 Background

2.1 Spatialization

One of the most important cues in spatial perception is source localization [14]. Spatialization refers to the positioning of sound objects in a virtual space and this is a key aspect in audio mixing, stereo reproduction in this context [15]. The perceptual quality of virtual sources in horizontal setups mainly depends on Interaural Level Difference (ILD) and Interaural Temporal Difference (ITD) cues [16, 17]. 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. The most frequently used technique to position sound sources in space is amplitude panning [18]. Modern audio production relies on amplitude panning techniques almost exclusively for the creation of azimuthal cues out of monophonic source signals.

2.2 Panning practices, rules and constraints

Ideally, the various sources of a mix should have a defined position and spectral bandwidth in the stereo field. The placement of sound sources is achieved using various creative choices as well as technical constraints based on human perception of sound localization [18]. Since this work deals mainly with optimization based on panning constraints, we seek to embed the common practices used for placing sound sources:

a. Panning is an iterative process

Mix engineers usually begin to mix with all center-positioned monaural tracks. Panning positions are determined based on the track's content [19]. High priority tracks such as vocals are usually kept centre-panned [20]. Panning decisions are not made for individual channels, but rather the result of an interaction between the various channels in the mix. Therefore, both content of the channels and interaction with rest of the mix is considered while panning [21].

b. Low frequency - best kept centred

Having off-centre low frequency sources can provide uneven power distribution. Also, there is very little directional information below 200 Hz. The position of a low frequency source is often psychoacoustically imperceptible [22]. ILD is not a useful cue at low frequencies for loudspeaker playback, but it is crucial for headphone listening. Therefore, low frequency content should be fixed in the centre of the mix [20].

c. Mid frequency - minimise spectral masking

Separation and definition of each track in the low-mid frequency region is critical to achieve a clear and well

produced mix [22]; most instruments have their fundamentals in this region. Sources with similar spectral content cause spectral masking, hence they are best placed apart in the stereo field [20]. This is referred to as opposition panning: if an important monophonic track is panned to one side, then another track with a similar musical function is often panned to the other side.

d. Higher the frequency - higher the panning width

Analysis of mixing practice shows that sources with higher frequency are progressively panned further towards the left and right extremes. Moreover, high frequency sounds diffract less as they bend around the head and so the panning effect feels more evident when exaggerated for high frequency content [22].

2.2.6 Overall stereo picture - maintain balance

The panning process should take care of the changes in activity and loudness of tracks over time. The most important constraint while choosing panning locations is to maintain spectral and spatial balance between left and right channels. Spectral balance keeps the intensity of frequency content uniform across the various bands in the left and right channels. Typically a mix should make use of the whole stereo space without compromising the stable centre picture. As mentioned in [22], a panned sound makes the mix feel lopsided thus destabilizing the centre of the stereo picture. Hard panning is highly uncommon and best to be avoided. The use of opposition panning is essential to balance similar sources panned to either channels. The effectiveness of a panner lies in providing a sense of spatialization and stereo width without pulling the centre-stable stereo picture.

2.3 Masking

Masking is a perceptual property of the human ear that occurs whenever the presence of strong audio signal makes the spectral or temporal neighbourhood of weaker audio signals imperceptible [23]. Frequency masking occurs when two or more stimuli are simultaneously presented to the auditory system. The relative shapes of the masker's and maskee's magnitude spectra determine to what extent the presence of given spectral energy will mask the presence of other spectral energy [Fig. 1.a] [24]. Several experiments have been performed in order to estimate the shape of auditory filters in the basilar membrane [25]. Figures 1.b,c illustrates how adjacent frequency bands of a sound source overlap and mask each other, leading to a source masking itself.

In the context of multitrack mixing, a sound source may inevitably mask itself and other sources. When sources are combined, the perceived loudness of one source at a given frequency may be low with respect to the other sources in the mix. This partial masking results in a mix sounding underwhelming, poorly produced with lack of clarity [26]. From an automatic mixing perspective, this is an optimization problem which aims to minimize masking through adjustments of level balances, spatialization, spectral characteristics and so on. The optimal solution can be thought off as the final stereo mix of the multitrack audio, released from masking using various controls.

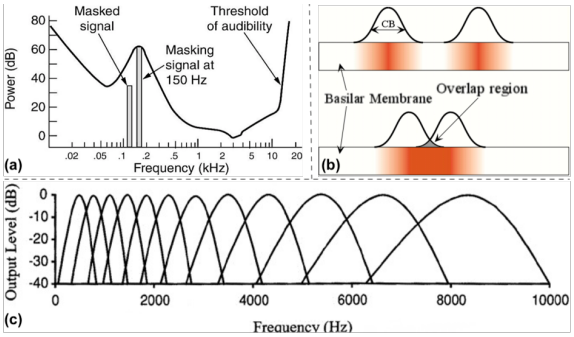

Fig. 1: (a) Frequency masking [12], (b)(c) Auditory filters in the basilar membrane [25].

For carrying out objective analysis of the multitrack unmasking improvement on the proposed auto-mixes, we use the MPEG Psychoacoustic Model, a well-established model used in audio coding/compression algorithms [27]. This model relies on a time-adaptive spectral pattern that emulates human auditory perception. The adaptation of the Masker-to-Signal ratio (MSR) from this model into a multitrack masking metric to be used for an optimization based automatic EQ is implemented in [12]. This model requires the computation of a multi-resolution Short-Time Fourier Transform (STFT), comprising of 6 parallel FFTs and each spectral frame filtered by a bank of level-dependent roex filters [28], which is costly. The choice of the masking metric will decide the algorithm's ability to work as an adaptive effect with real-time operability using side-chain processing [29] while complying to auditory perception. The proposed algorithm uses computation friendly strategies for spectral decomposition-reconstruction and choice of masking metrics and panning filters that comply to human hearing.

3 Methodology and implementation

In this section, we present the techniques and concepts used to carry out automatic spatialization and masking minimization. The framework of the proposed algorithm is as follows: monophonic audio tracks are fed in to an STFT framework (spectral decomposition), spectral similarity of each track with respect to every other track is computed (stored as a correlation matrix) to determine track pairs (subgrouping) that would undergo opposition panning [20], spectral masking of each track with

respect to the mix is computed, each track is assigned a panning filter based on the defined masking metric such that alternate spectral regions of each track are assigned particular positions and spectral bandwidth across the stereo field. The only difference between the real-time and the off-line approach lies in the assignment of the panning filters which determine the panning positions for each track. The former uses a particular ordering system to determine tracks in decreasing order of masking (from the correlation matrix) each of which would be assigned panning filters accordingly. In the off-line approach, the panning positions are determined by a particle swarm optimizer [30] that minimizes multitrack masking (cost function) while complying to the panning rules (constraints) discussed in section 2.2.

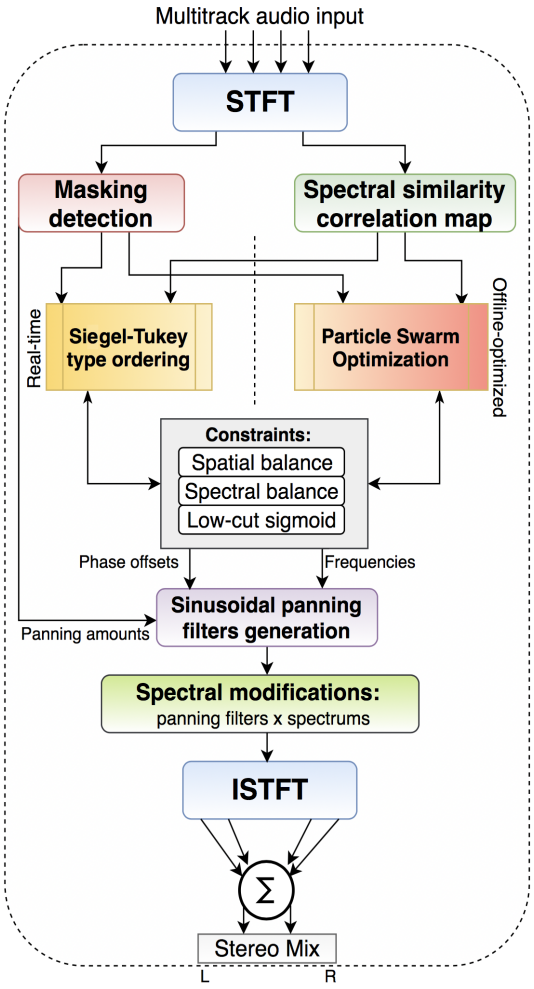

Fig. 2: Block diagram of the proposed algorithm.

3.1 Spectral decomposition and reconstruction framework

An audio mix is the result of a summation of an arbitrary number of input tracks. Since this work deals with spatialization; the input to the system is a monophonic multitrack. Since this panning technique involves time-frequency selective panning, we represent the tracks in the time-frequency domain by dividing the time domain signals into sequences of small overlapping frames to perform Short Time Fourier Transform (STFT):

$$X_{track}(n,k) = \sum_{m=-\infty}^{\infty} x_{track}(m)h(n-m)e^{-i2\pi km/N} \quad (1)$$

with discrete time n, discrete frequency bins k and a window function h of length N.

As far as STFT parameters are considered, a long window size of 2^{15} with a 1/16 hop length is chosen (for sample rate 44100 Hz), as per listening test results from [10]. Large window overlap caused temporal smearing which proved to cancel out desired effects, resulting in narrow panned mixes [31]. The plots and tables presented in Section 3 and 4 account for one such window frame during which all sources are active, as an example.

In Section 2.3 we discussed about previous work that uses multi-resolution STFT and roex filters to represent the displacement distribution and tuning characteristics across the human basilar membrane. However, in our proposed algorithm, since we aim to keep the computational cost low, we use STFT as defined in Equation 1. However, we carry out our spectral modifications in ERB scale (discussed in Section 3.3).

3.2 Spectral masking detection and multitrack subgrouping

Before applying the panning filters it is important to determine spectral masking in the multitrack. We discuss two features in this section:

1) Spectral masking for a given source [Fig. 3.a] can be measured by obtaining the amount of spectral overlap between the source and the rest of the mix. For a given track of interest, we define the spectral masking M of the track with respect to the rest of the mix, as follows:

$$M^2_{track}(n,k) = X^2_{track}(n,k) - X^2_{mix-track}(n,k) \quad (2)$$

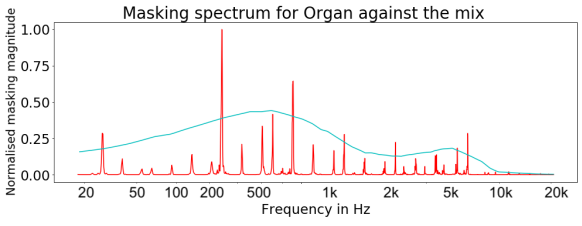

Fig. 3: a) Example of M_{track} in a multitrack (red), b) Smoothened average across frequency(blue)

2) To determine tracks that would undergo opposition panning, a correlation index matrix [Table.1] that measures similarity index [32] of each track with respect

	E.Guitar	Ac.Guitar	E.Piano	Organ	Sax
E.Guitar	-	**0.51**	**0.18**	**0.668**	**0.24**
Ac.Guitar	-	-	**0.158**	**0.423**	**0.261**
E.Piano	-	-	-	**0.298**	**0.401**
Organ	-	-	-	-	**0.33**
Sax	-	-	-	-	-

Table 1: Inter-track similarity - Correlation matrix

to every other track is computed. Spectral content of 2 given tracks X_i and X_j are compared using a similarity measure Ψ, computed as follows with forgetting factor $\lambda = 1$:

$$\phi_{ij}(n,k) = E\left\{X_i(n,k)X_j^*(n,k)\right\} \quad (3)$$

$$\phi_{ij}(n,k) = (1-\lambda)\phi_{ij}(n-1,k) + \lambda X_i(n,k)X_j^*(n,k) \quad (4)$$

$$\Psi_{ij}(n,k) = \phi_{ij}(n,k)|_{\lambda=1} \quad (5)$$

$$\Psi(n,k) = 2\frac{|\Psi_{12}(n,k)|}{[\Psi_{11}(n,k)+\Psi_{22}(n,k)]} \quad (6)$$

Intuitively, the above two metrics M and Ψ are coherent with perceptual masking since they determine frequency bins that are masked. The track-pairs that have highest spectral similarity are most vulnerable to spectral masking and hence each track within the pair is assigned time-frequency panning positions such that alternate spectral regions of the two tracks are placed on opposite ends of the stereo field. If a track has consistently low similarity index with all the tracks, the track is not processed further and remains unpanned in the mix; we observed this phenomenon for kick, bass and vocals most of the time and this is desirable, as we do not want to pan important tracks [7].

3.3 Panning filters: Frequency based spreading

The resulting mix of a heavily masked multitrack is underwhelming and confusing to listen to. In an ideal unmasked mix, all instruments are heard with relatively clear definition and there is an increasing spread of high frequency content across the stereo field. Audio engineers employ equalization and panning based on spectral masking to achieve this. In audio coding domain, masking models are widely used; the underlying principle is that the masked threshold of a signal is estimated to inform a bit-allocation algorithm or to remove perceptually irrelevant time-frequency components [33, 34]. Instead of removing masked frequency bins, the proposed algorithm places alternate frequency bands of the sources across the stereo field based on masking. The panning filters designed in our work draw inspiration from [10] in that no track is panned as a whole to either side, rather, time-frequency bins of each track are panned; however [10] did not account for masking reduction as an objective function, neither did its panning effect comply to auditory perception.

Multi-resolution STFT decomposition and reconstruction is computationally costly and not suitable for real-time effects, as discussed in Section 2.3. Instead we carry out spectral modifications accordingly (in logarithmic scale) to achieve perceptually relevant sonic results. We introduce sinusoidal shaped panning filters synthesized in ERB domain which accounts for decreasing frequency resolution of human hearing with increasing frequencies [35, 36]. The panning filter is projected from ERB domain to linear frequency domain. Figure 4 illustrates an exaggerated (extreme panning values) example of a monophonic track spread across the stereo field such that alternate frequency regions are placed on the Left and Right channel magnitude spectra. In Section 2.3, we also discussed about a source masking itself. Several experiments [25] have concluded that the auditory filters take the form of rounded complex exponential function like in [Fig. 1-b,c]. Our choice of sinusoidal-shaped panning filters solves this problem since a strong masking frequency component (masker) which would otherwise reduce the audibility of weaker components (maskee) in the same critical band would now be placed in the spatially opposite side of the maskee. The panning filter determines the spatial location of frequency bins of each track across the stereo field.

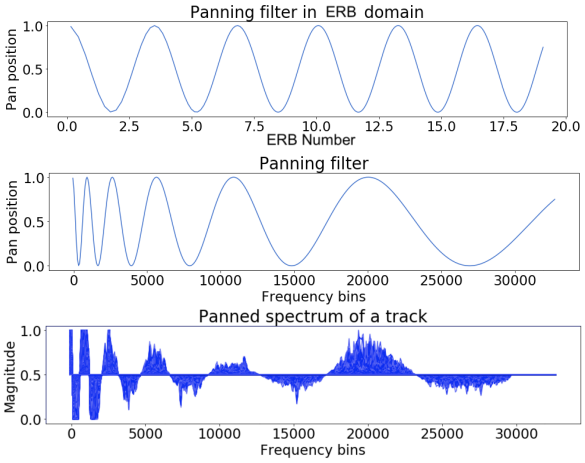

Fig. 4: a) Panning filter in ERB domain b) Panning filter in linear frequency domain c) Magnitude spectrum of a track's STFT frame after panning: 0 - hard left, 1 - hard right

The panning filter for each track is defined as follows:

$$P_j(n,k) = \rho_j(n,k) \cdot sin(21.4 \cdot \upsilon_j \cdot log(1+0.00437(b(k))) + \delta_j) \quad (7)$$

where $b(k)$ maps ERB number to frequency bin, υ_j and δ_j are frequency and phase offset of the panning filters for respective track (discussed in detail in Section 3.4), ρ_j is a spectral envelope filter whose role is as follows:

To comply to the panning rules discussed in Section 2.2, each panning filter is multiplied by a spectral envelope filter. The spectral envelope filter is computed as follows: the M_{track} function (Equation 2)[Fig. 3b] is multiplied with a low-cut sigmoid function with cut-off frequency at 200 Hz. Above 2 kHz, the M_{track} function is over-ridden and progressively set to 1.0 to avoid a sudden spike. The resultant is the panning filter envelope [Fig. 5] with 3 spectral regions whose roles are defined as follows:

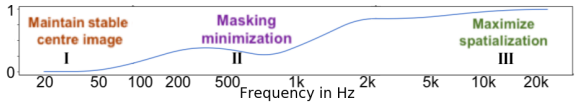

Fig. 5: Example of a spectral envelope filter ρ_j

The low-cut in Region I (< 200 Hz) ensures that low frequency content is not panned, thus contributing to a **stable centre image and spectral balance** [Section 2.2.b]. This feature becomes relevant especially in the headphone listening context in which spatial attributes/ILDs of low frequency content can be perceived [37].
Region II (500 Hz - 2 kHz) undergoes maximum masking in a multitrack [22], hence the panning amount in this band is determined by the M_{track} function. This function is smoothened across frequency by a frequency-varying averaging filter to avoid rapid variation of panning amounts [Fig. 3]. This region ensures **spectral masking minimization** [Section 2.2.c].
In Region III (> 2 kHz), the panning amounts are exaggerated by allowing the sinusoidal panning filters to spatialize alternate frequency bands above 2 kHz with maximum panning width [Section 2.2.d]. This region contributes to **spatialization** of the final stereo mix.

3.4 Multitrack masking minimization

We investigated and implemented both real-time and offline approaches to automatic spatialization under masking minimization. The difference between the two approaches lie in how the phase offsets δ_j and frequencies υ_j in Equation 7 are determined; these parameters are responsible in placing masked spectral content in different spatial locations across the stereo field.

3.4.1 Real-time approach

In the real-time approach, we use a palindromic Siegel-Tukey type ordering [38] to list the tracks in decreasing order of masking according to the correlation matrix, which determines inter-track spectral similarity. A Siegel-Tukey test determines if one of two groups of data tends to have more widely dispersed values than the other. In the example of [Table.1] in which the decreasing order of track pair similarity are illustrated, this type of ordering would give: Electric Guitar, Organ, Acoustic Guitar, Electric Piano, Saxophone. Phase offsets are computed such that both tracks of a track-pair from the correlation matrix are panned to opposite ends with $\delta_j = 180°$. The phases for each track-pairs' panning filters are offset with respect to the first panning filter's initial phase. The offsets are computed such that the maxima of each panning filter lies between the maximas of the first panning filter. This technique ensures that spectral-masked frequency bands of the tracks are spatially well separated in the stereo field thus making the masker and the maskee audible.

Considering the example of the first track-pair, in the left channel's spectrum, spectral content would alternate between the frequency bands of Organ and Electric Guitar, vice-versa for the right channel's spectrum. υ_j determines the number of oscillations of each panning filter. From preliminary listening tests, the effect of the panning envelope was most perceivable at normalized frequency $\upsilon_j = 0.01$, to obtain 6 frequency band splits. Extreme values of υ_j gave undesirable results: at low values, the panning filter divides each track into 2 broad spectral bands, each of which are panned to left and right channels, thus giving heavy spectral imbalance. High values of υ_j result in every alternate frequency bin of each track being panned in opposite direction, which prove to give no perceivable panning effect; the tracks sound monaural.

3.4.2 Particle Swarm Optimization approach

Mix engineers iteratively keep adjusting panning and EQ amounts of individual tracks until they achieve a well spatialized clear mix. Similarly, the proposed algorithm relies on Particle Swarm Optimization (PSO) [30] with the same objective: to minimize multitrack masking and to create a well spatialized mix with high perceived quality. The particles in this context are the phase offsets δ_j and the frequencies υ_j of the panning filters, which are objectively tuned to reduce the cost function M_m (multitrack masking in mid-frequency band) which is defined as the L_2 norm of $M_{track}(n,k)$ in Region II.

Due to the complexity and the nonlinearity of this iterative process, the optimization process tends to have multitrack influences, in that unmasking of one track leads to increased masking of other tracks. To balance the masking across all tracks, a second objective function with a min-max framework is used [12, 39] as part of the global optimization process:

$$x_{min} = min_x M_m(x) + \underset{x,i,j,i\neq j}{\operatorname{argmax}}(M_d(x,i,j)) \qquad (8)$$

where $M_d(x,i,j) = ||M_i(x) - M_j(x)||_2$
for $i, j = 1 \rightarrow$ no. of panning filters, $x = [\upsilon_j, \delta_j]$.

PSO is constrained by the panning rules described in Section 2.2.e and by bounds. It is important to maintain left/right balance across entire frequency spectrum (spectral balance) and energy ratio of both channels (spatial balance) [6]:

Balance angle per band for each STFT frame of the multitrack mix is calculated as follows:

$$Spec_{Band_i} = tan^{-1}\Big(\sum_{k=B_i}^{B_{i+1}-1} |X_L[k]|^2 \Big/ \sum_{k=B_i}^{B_{i+1}-1} |X_R[k]|^2\Big) \quad (9)$$

where X_L and X_R are the spectra of the left and right channels of the multitrack mix, $Band_i$'s are 5 bands that cover the audible frequency spectrum centered at 750 Hz, 1500 Hz, 2500 Hz, 7.5 kHz and 15 kHz with starting frequency index B_i for each band.

Spatial balance is calculated as the inverse tangent of the ratio of RMS energy of the left and right channels of the multitrack mix.The aim here is to make all the active sources converge to the centre such that the overall stereo balance is maintained between left and right channels at 0.5 as discussed in Section 2.2.e. The tolerance for the above metrics are bounded between 0.45 and 0.55. The bound of the overall frequency for each panning filter is $0.008 < \upsilon_j < 0.012$ for reasons discussed in Section 3.4.1. The frequency of the panning filter is linked to the spectral bandwidth of each track. The PSO thus minimizes multitrack masking by optimizing the panning filters within the constraints to comply to the panning rules[Fig. 6]. Each track's spectrum is multiplied by respective optimized panning filters, converted back to time domain by inverse-STFT and summed to obtain the final multitrack stereo mix.

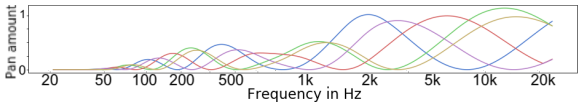

Fig. 6: Optimized panning filters (each color represents panning filter for respective track pair - Table 1).

4 Results

In this section we present the results of the proposed algorithms based on quantitative scores of several spatialization and masking metrics followed by subjective evaluation. The comparison involves the monophonic sum of multitrack, professional sound engineer mix, existing auto-mix works [6, 7, 10] and the 2 proposed auto-mix algorithms for various multitracks from different music genres. The sound engineer mix chosen for all the comparisons was the mix with the best mix rating in terms of spatialization, spatial balance and clarity from the Mix Evaluation dataset [40].

$\Delta Mask$		Folk	Country	Jazz	Funk	Pop	Rock
Real-time Mix	[12]	12.2	9.2	10	14.8	11.5	8.2
	ΔM_m	130	70	82	132	76	40
PSO Mix	[12]	19.5	14.6	15.1	26.7	17	12.6
	ΔM_m	142	75	91	159	80	55

Table 2: Change in masking : MPEG Multitrack Masking [12] and multitrack masking M_m.

4.1 Objective evaluation

Figure 7 illustrates the results of the optimization process in which the masking measure M_m (used as cost), reduces over the 20 iterations of a multitrack recording.

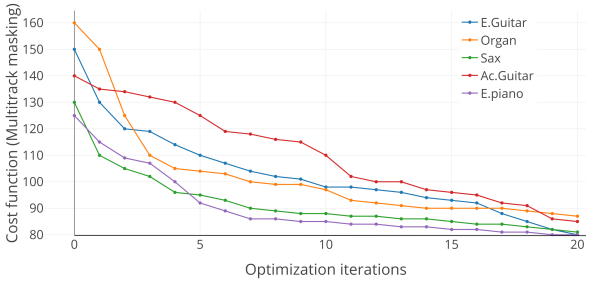

Fig. 7: PSO cost over iterations.

To evaluate the amount of multitrack unmasking achieved by the proposed algorithm, we use the state-of-the-art MPEG Psychoacoustic Model 1 [27]. We specifically use the cross-adaptive Multitrack Masking measure, M_n, for track n, as defined in [12]:

$$M_n = \sum_{sb \subset I_M} \frac{MSR_n(sb)}{T_{max}} \quad (10)$$

with I_M being the set of masked bands and $MSR_n(sb)$ being the masker to signal ratio for track n at band sb:

$$MSR_n(sb) = 10log_{10}\frac{T_n(sb)}{E_n(sb)} \quad (11)$$

where, $T_n(sb)$ is the masking threshold caused by rest of the mix, $E_n(sb)$ is the energy in band sb of track n and T_{max} is the predefined maximum amount of masking distance between $T_n(sb)$ and $E_n(sb)$.

In Table 2, we present the change in masking that occurred as a result of the proposed real-time and PSO (20 iterations) auto-mix for 6 songs of various genres chosen from the Open Multitrack dataset [41]. The perceptual masking metric in [12] is compared with the proposed multitrack masking measure M_m. Both metrics follow the same trend. The PSO mix gave higher masking reduction than the real-time non-optimized mix. This shows that the optimization step is beneficial to obtain the optimal panning filters for masking reduction. Though we do not have a clear understanding of

the masking release trend across genres, it appears that genres containing wideband sounds (like rock, with distorted guitar) do not release from masking as much as Funk in which the rendition of sounds are sparse and more percussive. The number of tracks also affect the unmasking amount.

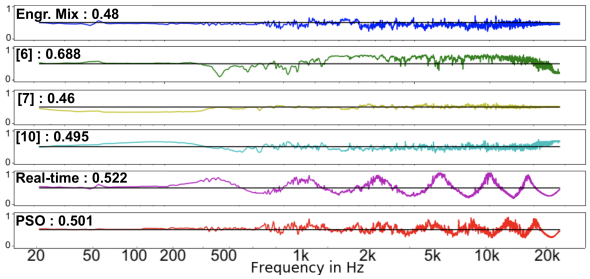

Fig. 8: Stereo panning spectrum and Panning RMS, (Y-axis: Panning amount).

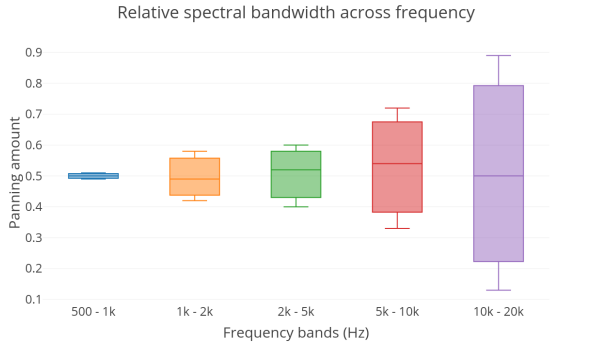

Fig. 9: PSO mix: Relative panning bandwidth across frequency bands, for 100 songs.

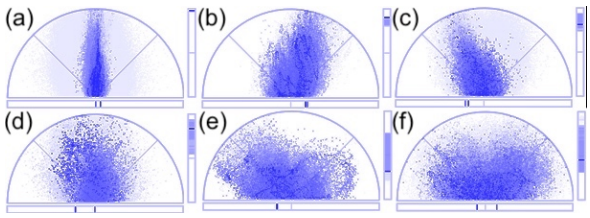

Fig. 10: Goniometer output: a) Sound Engr.mix, b) [6], c) [7], d) [10], e) Real-time mix, f) PSO mix.

We investigated the amount of spatialization achieved by analyzing the stereo panning spectrum (SPS) [32] and also summarized the result by comparing the Panning RMS [42] for all the mixes. These measures are used to determine the amount and distribution of panning in different frequency bands as well as its dynamic evolution over time. SPS is a panning index across frequency obtained by shifting and scaling the similarity function (Equation 6) and it is a measure of overall panning in a stereo signal. The basic idea behind the SPS is to compare the left and right signals in the time–frequency plane to identify the different panning gains associated with each time–frequency bin. Panning RMS is the root-mean-square of SPS. The results are illustrated in [Fig. 8] for one STFT frame.

The proposed algorithms, just like professional sound engineer mix, achieve desirable spatial balance (panning RMS closer to 0.5), stable center image at low frequencies and increasing panning width with increasing frequency. This is illustrated in [Fig. 9]: spectral panning bandwidth (Equation 9) is calculated across audible frequency bands [6] for over 100 songs from the Open Multitrack dataset [41]. Spectral balance is maintained at around 0.5 for all frequency bands. The result complies to the best panning practices discussed in Section 2.2 and the performance of the proposed algorithms remain consistent throughout the mixes from the dataset. The cyclic dependence of SPS (more dominant in the real-time mix) is prominent due to the sinusoidal-shaped panning filters which are consequently placed at constant offsets such that masked track pairs from the correlation matrix [Table 1] are on spatially opposite sides. This variance did not seem to give undesirable audible output. Rather, a bias in the SPS results in poor spatial and spectral balance: [7] and [10] have unstable spectral balance in the low frequency end which results in an unstable stereo image, [6] has poor spatial balance since the energy is concentrated towards stereo right.

To analyze the amount of spatialization and stereo activity, the mix outputs were run through a goniometer [6]. The proposed algorithm proved to give highest stereo activity and centre-image stability consistently throughout the length of each song for all the songs. From the goniometer snapshots of all mixes [Fig. 10] we observe the following: a) Human mix has stable centre image but has relatively narrow panning, b) [6] and c) [7] have lopsided spread thus having poor spatial balance, d) [10] and e) Proposed real-time mix has reasonable spatialization but loses spectral balance at certain frequencies, f) PSO mix performs extremely consistent with stable centre image as well as maximum spatialization.

4.2 Subjective evaluation

25 participants with more than 10 years of formal experience in Music Production were asked to rate all mixes in an audio perceptual evaluation (APE) preference test [43] in terms of panning quality, instrument separation and clarity on a single scale ('Low' to 'High'). All tests were conducted in an isolated listening room, with identical headphones, and same listening level. It was a reference free test with all conditions presented in a randomized manner. The results [Fig. 11] indicate that the proposed algorithms outperform existing auto-mix works and the PSO mix has consistent ratings comparable to professional

sound engineer mixes. Comments by the participants include "very clean centre image, well balanced, can hear instruments distinctly" (PSO mix), "weird panning but there is a nice sense of space which gives a live feeling but maybe a little too wide" (real-time). Past auto-mix works got comments like "good instrument separation but harsh panning", "off-centre bass". The mixes used for the APE test are available online at http://webaudio.gutech.edu.om/test.html?url=tests/pantest1.xml

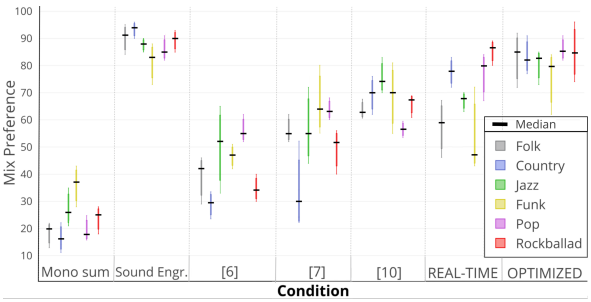

Fig. 11: Listening test results: mix quality rating for multitrack monosum, professional sound engineer mix, existing auto-mix works [6,7,10], proposed real-time mix, PSO mix; represented as a box plot with first value, median, last value.

5 Conclusion

This paper describes a frequency-based multitrack panning automation method. It achieves an increased sense of spatialization and masking reduction while complying with the well known panning practices. Both real-time and optimized off-line approaches are presented, implemented and evaluated. They rely on the same framework for spectral decomposition, multitrack subgrouping, masking detection and reduction. The proposed framework is computationally low enough to be implemented as a real-time plug-in, yet the sonic output of the proposed algorithms comply to human perception since we use ERB-scale sinusoidal shaped panning filters. The sinusoidal nature of the panning filters also addresses the problem of a source masking itself.

The proposed algorithm computes inter-track similarity to determine tracks that would undergo opposition panning, thus giving improved clarity and intelligibility to the final stereo mix. The proposed auto-mixes are compared against existing automatic panning works as well as professional mixes by sound engineers. They were evaluated for unmasking using the MPEG Psychoacoustic Model and various panning measurements using goniometer and panning RMS. The panning filters complied to some well known panning practices: maintain stable centre balance at low frequencies, unmask the mid-frequency area and high panning width at high frequencies. The subjective results of the proposed auto-mixes show consistently higher ratings than existing auto-mix works, and are comparable to professional sound engineer mixes. Our proposed intelligent system can be used in the mixing stage to place sources across the stereo field, to produce a well spatialized mix with reduced auditory masking and improved perceived quality. It is the intent of the authors to extend this work to automatic placement of sound sources in a multispeaker environment and also consider binaural effects of masking.

References

[1] E. Perez-Gonzalez, J. D. Reiss, "Automatic gain and fader control for live mixing," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, USA*, pp. 1–4 (2009).

[2] J. D. Reiss, "Intelligent systems for mixing multichannel audio," *17th Intl. Conference on Digital Signal Processing, IEEE, Corfu, Greece*, pp. 1–6 (6-8 July, 2011).

[3] B. De Man, J. D. Reiss, R. Stables, "Ten years of automatic mixing," *Proc. of the 3rd Workshop on Intelligent Music Production (Salford, UK)* (2017).

[4] J. Wakefield, C. Dewey, "An investigation into the efficacy of methods commonly employed by mix engineers to reduce frequency masking in the mixing of multitrack musical recordings," *138th Audio Engineering Society Convention, Warsaw, Poland* (7-10 May, 2015).

[5] E. C. Matz, Daniel, J. Abeßer, "New Sonorities for Early Jazz Recordings Using Sound Source Separation and Automatic Mixing Tools," *Proc. of the 16th Intl. Society for Music Information Retrieval Conference (ISMIR), Malaga, Spain*, pp. 749–755 (26-30 Oct, 2015).

[6] S. Mansbridge, S. Finn, J. D. Reiss, "An autonomous system for multitrack stereo pan positioning," *133rd Audio Engineering Society Convention, San Fransisco, USA* (26-29 Oct, 2012).

[7] E. Perez-Gonzalez, J. D. Reiss, "A real-time semiautonomous audio panning system for music mixing," *EURASIP J. Adv. Signal Process. https://doi.org/10.1155/2010/436895*, vol. 2010, pp. 1–10 (2010).

[8] M. A. Gerzon, "Panpot laws for multispeaker stereo," *92nd Audio Engineering Society Convention, Vienna, Austria* (March 24-27, 1992).

[9] R. Izhaki, *"Panning" in Mixing Audio: Concepts, Practices and Tools, chapter 13, pp. 184–203* (Focal Press/Elsevier, Burlington, USA, 1st edition) (2007).

[10] P. D. Pestana, J. D. Reiss, "A Cross-Adaptive Dynamic Spectral Panning Technique," *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx), Erlangen, Germany*, pp. 303–307 (September 1-5, 2014).

[11] J. Jot, C. Avendano, "Spatial enhancement of audio recordings," *23rd Audio Engineering Society Conference: Signal Processing in Audio Recording and Reproduction, Helsingør, Denmark* (May 23-25, 2003).

[12] D. Ronan, Z. Ma, P. M. Namara, H. Gunes, J. D. Reiss, "Automatic Minimisation of Masking in Multitrack Audio using Subgroups," *ArXiv e-prints* (March, 2018).

[13] D. M. Ronan, H. Gunes, J. D. Reiss, "Analysis of the subgrouping practices of professional mix engineers," *142nd Audio Engineering Society Convention, Berlin, Germany* (May 20-23, 2017).

[14] J. Blauert, *"Spatial hearing: the psychophysics of human sound localization", Rev. edn. Cambridge, Mass.* (MIT press) (1997).

[15] G. Gatzsche, F. Melchior, "Spatial audio Authoring and Rendering: Forward Research through Exchange," *Intl. Computer Music Conference (ICMC), Belfast, Ireland* (August, 2008).

[16] V. Pulkki, M. Karjalainen, "Localization of amplitude-panned virtual sources I: stereophonic panning," *Journal of the Audio Engineering Society*, vol. 49, no. 9, pp. 739–752 (2001).

[17] V. Pulkki, "Localization of amplitude-panned virtual sources II: Two-and three-dimensional panning," *Journal of the Audio Engineering Society*, vol. 49, no. 9, pp. 753–767 (2001).

[18] V. Pulkki, "Spatial sound generation and perception by amplitude panning techniques," *Ph.D. Thesis, Helsinki University of Technology* (2001).

[19] D. Self, *"Recording consoles," in Audio Engineering: Know It All* (vol. 1, chapter 27, pp. 761–807, Newnes/Elsevier, Oxford, UK, 1st edition) (2009).

[20] E. Benjamin, "An experimental verification of localization in two-channel stereo," *121st Audio Engineering Society Convention, San Fransisco, USA* (Oct 5-8, 2006).

[21] R. Neiman, *"Panning for gold: tutorials," Electronic Musician Magazine* (http://emusician.com/) (2002).

[22] R. Izhaki, *"Mixing audio : concepts, practices, and tools", Third edn. London: Taylor and Francis* (Focal Press) (2013).

[23] B. R. Glasberg, B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138 (1990).

[24] A. J. Oxenham, B. C. Moore, "Modeling the additivity of nonsimultaneous masking," *Hearing research*, vol. 80, no. 1, pp. 105–118 (1994).

[25] B. C. Moore, "Masking in the human auditory system," *Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction* (May 1, 1996).

[26] Z. Ma, J. D. Reiss, D. A. Black, "Partial loudness in multitrack mixing," *53rd Audio Engineering Society Conference: Semantic Audio, London, UK* (Jan 26-29, 2014).

[27] K. Brandenburg, G. Stoll, "ISO/MPEG-1 audio: A generic standard for coding of high-quality digital audio," *Journal of the Audio Engineering Society*, vol. 42, no. 10, pp. 780–792 (1994).

[28] T. Irino, R. D. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 412–419 (1997).

[29] V. Verfaille, U. Zolzer, D. Arfib, "Adaptive digital audio effects (A-DAFx): A new class of sound transformations," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1817–1831 (2006).

[30] J. Kennedy, "Particle swarm optimization," *In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, USA*, pp. 760–766 (2011).

[31] E. P. Gonzalez, J. D. Reiss, "Improved control for selective minimization of masking using interchannel dependancy effects," *Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx), Espoo, Finland*, p. 12 (September 1-4, 2008).

[32] C. Avendano, J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," *22nd Audio Engineering Society Conference: Virtual, Synthetic, and Entertainment Audio, Espoo, Finland* (June 15-17, 2002).

[33] A. Gersho, "Advances in speech and audio compression," *Proc. of the IEEE*, vol. 82, no. 6, pp. 900–918 (1994).

[34] P. Balazs, B. Laback, G. Eckel, W. A. Deutsch, "Time–frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 1, pp. 34–49 (2010).

[35] E. Zwicker, H. Fastl, *"Psychoacoustics: Facts and models", 3rd edn. Berlin: Springer*, vol. 22 (Springer Science & Business Media) (2013).

[36] S. S. Stevens, H. Davis, *Hearing: Its psychology and physiology* (American Institute of Physics for the Acoustical Society of America New York) (1938).

[37] J. C. Middlebrooks, D. M. Green, "Sound localization by human listeners," *Annual review of psychology*, vol. 42, no. 1, pp. 135–159 (1991).

[38] S. Siegal, *"Nonparametric statistics for the behavioral sciences",* (McGraw-hill, New York), 2nd ed. (1956).

[39] P. McNamara, S. McLoone, "Hierarchical demand response for peak minimization using Dantzig–Wolfe decomposition," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2807–2815 (2015).

[40] B. De Man, J. D. Reiss, "The mix evaluation dataset," *Proc. of the 20th Int. Conference on Digital Audio Effects (DAFx), Edinburgh, UK* (September 5–9, 2017).

[41] B. De Man, M. Mora-Mcginity, G. Fazekas, J. D. Reiss, "The open multitrack testbed," *137th Audio Engineering Society Convention, Los Angeles, USA* (Oct 9-12, 2014).

[42] G. Tzanetakis, R. Jones, K. McNally, "Stereo Panning Features for Classifying Recording Production Style." *Proc. of the 8th Intl. Society for Music Information Retrieval Conference (ISMIR), Vienna, Austria*, pp. 441–444 (23-30 September, 2007).

[43] B. De Man, J. D. Reiss, "APE: Audio perceptual evaluation toolbox for MATLAB," *136th Audio Engineering Society Convention, Berlin, Germany* (April 26-29, 2014).