Automatic Mixing Systems for Multitrack Spatialization based on Unmasking Properties and Directivity Patterns

Ajin Tom

Department of Music Research
Schulich School of Music
McGill University
Montreal, Canada

November 2019

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of the Arts in Music Technology.

Abstract

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

This research investigates, develops and implements automated mixing strategies that optimize localization of sound sources in a multitrack mix using innovative approaches that rely on masking properties of perception and directivity of the musical source for coherent and flexible spatialization. The aim is to deliver sound for an immersive environment, in which sources can appear at any position with specific directivity patterns that quantify their directional dependent behaviour in the two or three-dimensional space around the listener. This thesis focuses particularly on spatialization aspects of multitrack mixing; one approach being frequency-based panning that relies on release from spectral masking using optimization techniques to obtain an unmasked and well-spatialized stereo mix. Another approach is aimed at multichannel systems beyond stereo for which the same optimization framework is used but with source directivity as constraints.

The proposed automix systems can be used in the mixing stage to place sources in the stereo/sound field, to produce a well spatialized mix with reduced auditory masking and improved perceived quality (clarity and intelligibility). The proposed algorithms for both techniques make use of a spectral panning linear system which generates optimized filters for each track, with constraints that comply to perception. The evaluation criteria involves both subjective as well as objective tests to obtain measures for unmasking amount and extent of spatialization. Audio samples generated by the proposed algorithms are available online. Both spatialization approaches proved to give a good sense of unmasking and spatialization. The proposed spatialization technique can be beneficial to design systems that create plausible 3D sound scapes. Using innovative audio effects/tools like source directivity coupled with optimization techniques can address how music can be meaningfully upmixed from the more common stereo to other playback formats like 5.1, 22.2 and Ambisonics.

Résumé

Le mixage multi-pistes est un processus itératif dans lequel divers paramètres de traitement, tels que l'équilibre de sonie, l'égalisation et la compression sont ajustés pour obtenir un signal de sortie cible conforme à des critères perceptifs et objectifs. La recherche sur les systèmes de mixage automatiques a connu une croissance rapide au cours des dix dernières années, proposant des systèmes intelligents pour presque tous les aspects de la production audio. Des outils intelligents analysant les liens entre tous les canaux audio afin d'en automatiser leur mélange ont été conçus.

Dans cette recherche nous étudions, développons et mettons en œuvre des stratégies de mixage automatisées en optimisant la localisation des sources sonores à l'aide d'approches innovatrices s'appuyant sur des propriétés de masquage perceptifs et/ou sur la directivité des sources musicales pour une spatialisation cohérente et flexible. Dans ma thèse, je me suis particulièrement intéressé aux aspects de spatialisation dans les mélanges multi-pistes selon deux approches: la première fondée sur un panoramique spectral qui repose sur une minimisation du masquage fréquentiel à l'aide de techniques d'optimisation afin d'obtenir un mixage stéréophonique non masqué et bien spatialisé; la seconde visant des systèmes multi-pistes au-delà de la stéréophonie pour lesquels le même cadre d'optimisation est utilisé mais contraint par la directivité des sources sonores.

L'objectif est de produire des sons pour un environnement immersif, où les sources peuvent apparaître à n'importe quelle position avec des motifs de directivité spécifiques qui caractérisent leur comportement anisotrope dans l'espace à deux ou trois dimensions entourant l'auditeur.

Acknowledgments

I would like to thank my supervisor Professor Philippe Depalle for his guidance and enthusiasm that inspired me throughout the course of my Master's degree and research. I am extremely grateful for the opportunity to study and carry out research in the Schulich School of Music at McGill University. In the diverse Music Technology program here, I got to use my skills and knowledge as a researcher, musician, sound designer and engineer. I am fortunate to have been a student of Gary Scavone, Philippe Depalle, and Marcelo Wanderley. In their enlightening seminars, I got to learn and work on several interesting topics related to Computational Modelling of Acoustical Systems, Audio Signal Processing, and Digital Music Instruments. I thank Darryl Cameron for his technical support and help during the course of my research.

I would also like to thank the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) for the outstanding facilities provided for Audio Research. I would like to thank Professor Joshua Reiss and my friends at the Centre for Digital Music (C4DM) for their guidance and support during my research internship at Queen Mary University London, UK, where I carried out the first phase of this research. I would like to thank my funding sources: MITACS, CIRMMT, Department of Music Research teaching assistantships and research assistantships under Professor Marcelo Wanderley in the Input Devices and Music Interaction Laboratory (IDMIL) and under Caroline Palmer in the Sequence Production Lab (SPL) and the Image Analysis Lab (IAL) funded by NSERC-CREATE.

Finally, I thank my friends and family for their love and encouragement. A special mention to Ankita Singh for being there for me throughout this journey. Their continuous support helped me pursue my passions and interests.

Contents

List of Figures

List of symbols

Notation	Description
θ	Azimuth angle
ϕ	Elevation angle
r	Physical distance (m)
s	Input sound signal
t	Time (s)
g	Gain
M_o	Number of output tracks / loudspeakers
M_i	Number of input tracks / sources
n	Time (samples)
S	Short-time Fourier transform of audio signal
k	Frequency bin
q	Short-time Fourier transform frame index
I	Hop size (samples)
N	Number of frequency points / FFT Size
f_s	Sampling rate (Hz)
H	Frequency response
y	Output sound signal
B	Bandwidth (Hz)
M_{track}	Masking amount
Ψ	Similarity measure
υ	Frequency of panning filter
δ	Phase offset of panning filter
ω	Angular frequency (rad/s)

Notation	Description
c	Speed of sound in air (c = 343m/s is assumed)
p	Pressure (Pa)
f	Frequency (Hz)
M_e	Number of elementary sources
D	Directivity function
ϑ	Scaling factor mapping directivity across frequency
η	Extent of directivity
λ	Wavelength (m)
v	Velocity (m/s)
L	Duration/length of signal (s)

Chapter 1

Introduction

Music production involves three major steps after the compositional aspects: recording (capturing sounds of sources), mixing (combining and processing sound sources/tracks) and mastering (processing the mix before final sound output). This dissertation deals with the mixing step, specifically, carrying out automatic placement of the sound sources' frequency content in the sound field.

Audio mixing is the process of combining several audio tracks / sources (referred to as multitrack) of a recording or live music session into a final mono, stereo or multichannel sound file or playback [3]. Multitrack mixing is a fairly complex process carried out by audio engineers. It involves a number of different steps each of which has several substeps depending on the sources, interaction with other sources and desired result. In the process of combining the individual tracks, their relative sound levels are adjusted and balanced after which they go through processes like equalization, compression, and several other frequency and time-based effects. For stereo, surround and other multichannel audio playback, placement of the tracks in the stereo, surround, or sound field is an important process in creating a widened mix with better sound clarity. This process is referred to as panning or spatialization [4].

The overall research goal of this dissertation is to investigate, develop and implement automated mixing strategies that optimize localization of the sound sources using innovative spectral panning approaches that rely on masking properties of perception based on best panning practices and directivity of the musical source for coherent and flexible spatialization.

1.1 Motivation

Democratization of audio technology has enabled music production on limited budgets, putting high-quality results within reach of anyone who has access to a laptop, a microphone and the abundance of free software on the web [5]. Similarly, musicians are able to share their own content at very little cost and effort, again due to high availability of affordable technology. Mixing is a crucial process in music production which requires a high level of expertise in order to deliver professional-standard material. Raw, recorded tracks almost always require a considerable amount of processing before being ready for distribution, such as level balancing, equalization, dynamic range compression and artificial reverberation [3]. An amateur music producer could run into sonic problems due to uninformed microphone placement, unsuitable recording environment, technical issues with the instrument or even poor performance by the musician. In the context of live music, the mixing task is quite demanding and sensitive due to problems related to acoustic feedback, room resonances and poor equipment. These observations indicate that systems taking care of the mixing stage of music production for live and studio-based recording situations, quickly and automatically, would be valuable [6].

With such automatic mixing systems in place, home recording becomes more affordable, smaller music venues can achieve better sound output for their loudspeakers and monitor systems when an expert operator is unavailable, and musicians can increase their productivity and focus on the creative aspects of music production. Meanwhile, professional mix engineers are often under pressure to produce high-quality mixes quickly and at minimal costs [7]. Computer-assisted music production tools that come out of automatic mixing research would be highly beneficial to these mix engineers since they can let these tools provide meaningful pre-mixes, in other words, prior parameters for the signal processors in the mixing chain. Indeed, research into such automatic mixing systems has grown rapidly over the last ten years, with intelligent systems proposed for almost every aspect of audio production, and many of these have seen commercial application; the potential for deeper understanding of auditory perception and mixing practices is huge [8].

1.2 Research strategy

This thesis focuses on sound spatialization systems that build on recent work that resulted in the emergence of a new class of cross-adaptive systems [9] aiming at automatic mixing. The idea is to determine the mixing parameters of each individual input in order to optimize objective (e.g., level balance) and perceptual (e.g., release from masking) characteristics of the mixed output signal. In the automatic mixing systems so far, spatialization has essentially relied on level-based or delay-based positioning of sources and their frequency content [10–12]. 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 [1, 13, 14].

In this research, automatic multitrack spatialization is carried out using two approaches both of which are based on the common framework of spectral unmasking using optimized frequency-based panning filters. Spectral masking in this context refers to spectral content of sources/tracks being undesirably masked by the rest of the multitrack [13]. The primary focus and interest in this research lies in studying spectral masking effects over spatial masking [15] since reverberation aspects are not considered.

The first approach, targeted for stereo, is a frequency-based spreading technique in which each track is assigned a panning filter across time depending on the spectral content of each track and the rest of the mix. The choice of these panning filters complies with spectral unmasking and best panning practices, eventually creating a well spatialized mix with increased clarity. Both a real-time and an offline optimization-based approach are designed and implemented. Reduction of inter-track auditory masking is investigated using the properties of the MPEG psychoacoustic model [2] along with various other masking and spatialization metrics [16], extended for multitrack content. Subjective and objective tests (led at the Centre for Digital Music, Queen Mary University, London, UK) indicate that the proposed auto-mixes sometimes outperformed or at least were at par with existing auto-mix works. The optimized auto-mix has consistent ratings that are comparable to professional sound engineer mixes. These results are published in a peer-reviewed paper which was presented at the 146th AES Convention, Dublin, Ireland [17].

The second approach addresses how spatialization systems could place sound sources in other formats beyond stereo, such as ambisonics or 5.1 surround for example. The core idea of this panning technique is to use directivity as a feature of the source [18].

Directivity patterns quantify directional dependent behaviour of sources thus guiding the distribution/spreading of spectral content on targeted mixing channels of the sound output format [19]. The system uses template directivity drawn from acoustics / geometry of the source and associated radiation, or directivity designed by the user for producing specific spatial effects. The use of directivity patterns will replace ad-hoc rules used in the first approach, to achieve multitrack mixing by automatically spreading the energy of the source over the considered set of tracks. Directivity patterns essentially determine frequency response for a sound source based on its directivity, position and orientation with respect to the listener. This frequency response is a more natural panning filter (since the radiation pattern of an acoustical instrument is a natural phenomenon) that can be applied to respective sources to carry out spectral panning. Directivity can also be used as an audio effect (instead of a plausible recreation of an actual acoustical directivity) to carry out innovative spatial effects.

The state-of-the-art technology for both stages relies on a population-based stochastic optimization technique called particle swarm optimization [20]: a frequency-based spreading of masked tracks constrained by best panning practices (approach one) and constrained by source directivity (approach two) to eventually generate the best possible spatialized mix. Eventually, users will be able to use either constraints or a mix of both. The evaluation criteria involve objectively calculating spectral unmasking [1, 17] across channels as well as fruitful extent of spatialization using a goniometer [10] and systematic analysis of directivity rendered around sources.

1.3 Thesis overview

This thesis has six chapters and two appendices: the first chapter serves as introduction to point out the general problem statement, research questions, objectives, research strategy and scope of the research. Since this research deals with automatic mixing and spatialization, Chapter 2 covers the background aspects of the art and science of sound recording and also about spatial audio reproduction. It aims to make a connection between the physics / acoustics, signal processing as well as sound recording aspects of mixing. Chapter 3 introduces the state-of-the-art behind automatic mixing and also presents the overall framework and implementation of the proposed automix system for spatialization and unmasking. The algorithm presented in Chapter 3 can make use of both or either of two kinds of spatializa-

tion approaches: Chapter 4 presents approach one - spatialization relying on ERB-based sinusoidal panning filters and Chapter 5 presents approach two - spatialization relying on directivity patterns, to carry out unmasking. Since the second approach is inspired from the acoustical nature of instruments having radiation patterns, several experiments are presented to build up the underlying concepts behind directivity. The sound examples for both the approaches are available online (https://www.ajintom.com/auto-spatial). Chapter 6 provides a brief summary, our findings and future work of our research. Appendix A presents the theory behind the optimization technique (Particle Swarm Optimization) used in the proposed algorithm. Appendix B presents derivation for the limit on the speed of temporal evolution of directivity. Lastly, a bibliography of sources cited in this work is presented.

1.4 Contributions and Scope of the Research

The goal of this research is to propose, implement and evaluate automatic spatialization systems. The aim is to deliver sound for an immersive environment, where sources can appear at any position with specific directivity patterns that quantify their directional dependent behaviour in the two or three-dimensional space around the listener. The findings of this research will also prove to be beneficial in the context of loudspeaker listening in an acoustic environment; directivity patterns can compensate loudspeaker displacements or dislocations by rebalancing the energy of each source on each loudspeaker, thus maintaining acoustic coherence. This research will also support the latest advent of VR/AR technologies which require better true-to-life 3D audio immersion. The proposed spatialization technique can be beneficial to decide how music can be meaningfully upmixed from the more common stereo to other playback formats like 5.1, 22.2, Ambisonics, etc. Moreover, the sound recording and mixing community can further develop and use such innovative audio effects like source directivity to carry out spectral panning and unmasking.

Chapter 2

The Art and Science of Sound Recording

This chapter aims to provide pre-requisite knowledge of sound recording concepts relevant to this dissertation, mainly focusing on multitrack mixing and spatial audio reproduction. The points discussed in the following sections are gathered from books on audio mixing and recording by Eargle [21], Izhaki [3] and Moylan [22].

Sound recording is the art of capturing sound and then reproducing it either immediately or from a storage medium through speakers or headphones. The main goal of sound recording is to have the most faithful representation of the actual sound scene that was produced in the recording space. In the context of this dissertation it is useful to explain the associated concepts from three perspectives: physics/acoustics, signal processing and sound recording. From a physics perspective, sound recording is a heavily undersampled capturing of the sound field by placing microphones in limited locations in the recording space. In a usual recording set up, microphone(s) are placed around the sound sources and in the room to capture **pressure** variations in **space** over **time**. It is important to note that microphones have limited/specific directionality, so they are able to capture only from limited portions of space. Since there is a distance between the microphone and source, there is a delay due to sound propagation. From a signal processing view point, these pressure variations over time are used to derive the sound signal as **amplitude** displacements over time, from which we can further derive **frequency**. The fusion of many amplitude and frequency components of a single sound form the **spectrum**, according to the Fourier

transform. Signal processing tools (effects) are used to carry out modifications (such as filtering) on the sound signals for further processing. The physical features of sound, namely **amplitude, frequency, time, spectrum** are linked to higher level perceptual features **loudness, pitch, duration, timbre (sound quality)** respectively. The translation process from the physical attributes to the perceived features is nonlinear, and differs between individuals [23].

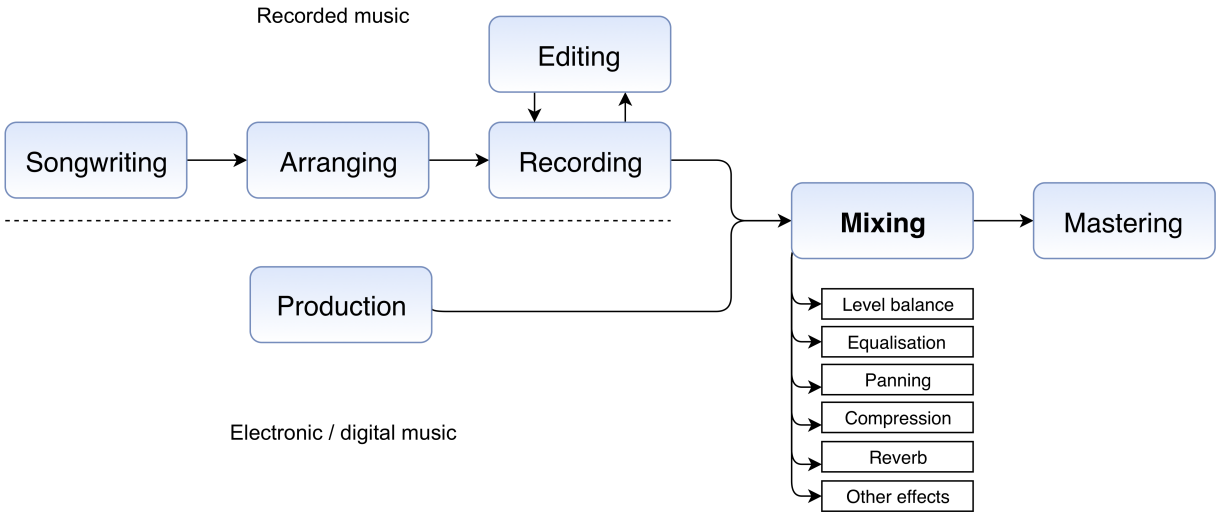

Fig. 2.1 Production chain

2.1 Overview of Music Production

Figure 2.1 illustrates the most common production chain for recorded as well as electronic/digital music. In the context of recorded music, once the compositional objectives related to songwriting and arranging (transforming individual compositional components into sets of voices and instruments) are done, the recording can take place effectively. Modern recordings and mixing are carried out on multitrack recorders and mixing consoles (Figure 2.2) which offer a multitude of inputs each linked to the signal acquisition chains of each of the sources being recorded. These input signals are run through respective signal processors and finally summed together. Sound engineers sometimes insert outboard gear like analog effect processors or amps in the signal path to extend processing possibilities. With the advent of faster computers, producers also rely on music production software,

known as Digital Audio Workstation (DAW, Figure 2.3). DAWs provide a full-fledged virtual workspace to choose samples, record and visualize audio content, edit (trim, crossfade, loop) audio material, apply signal processing with software plugins, etc.

Fig. 2.2 Typical music production set up: Recordist using a mixing desk, recording a saxophone player (left), Music producer using synthesizers and DAW, producing electronic/digital music (right) (Picture credits - UCLA School of Music)

Recording is usually carried out in a multitrack format; each source is recorded into a separate track such that each of them can be mixed, processed, edited, or otherwise altered at some future time, and without altering other sound sources. Before the introduction of multitrack recording, all sounds and effects that were to be part of a record were mixed at one time during a live performance. If the recorded mix wasn't satisfactory, or if one musician made a mistake, the excerpt had to be performed over until the desired balance and performance was obtained. The ability to record sounds into separate tracks meant that combining and treating these sounds could be postponed to the mixing stage. Electronic/digital music is different in nature to that of recorded music. In this case, the production stage is a mixture of songwriting, arranging and recording stages. Producers use virtual and/or analog synthesizers along with software plugins on their DAW to record samples or sounds based on how well they fit into the mix that they are building up (Figure 2.2).

Fig. 2.3 Example of Digital Audio Workstation (DAW) session

Once sound sources are combined, some of them might not fit perfectly well within the mix. For example, the kick drum and bass might sound exceptionally good when played in isolation, but combined they might mask one another. Filtering the bass might make its sound thinner, but will work better in the context of the entire mix. The mix as a whole is the final product; this does not mean that the sound of individual elements is not important, but the interaction between the content of the tracks and the overall mix takes priority. The final post production step is called mastering. This involves balancing sonic elements of a mix and optimizing playback across all target playback systems and media formats. In the following section various aspects of multitrack mixing are highlighted.

2.2 Multitrack Mixing: A creative engineering process

Long before the advent of computer-based DAWs or analog mixing consoles, composers arranged sound sources in physical space, taking into account the properties of each instrument in an effort to create an overall balance. There are a lot of similarities in the

perspectives of this olden day "mixing" and the technology-driven mixing practices, which evolved from multi-tape recorders to modern day DAWs and multitrack mixers. The loudest instruments would most often be placed at the back of the stage or to a certain side, so that solo instruments such as violins and flutes cut through in the mix and similar sounding instruments were placed apart. In the context of modern music production, this problem of frequency overlap (which causes spectral masking, discussed in Section 3.6.1), is solved using EQing, panning and applying reverb. Later in the dissertation we will see how we can use or link olden day practices (mixing based on positioning of sources in the acoustic space) to innovative techniques like modelling orientation of acoustic instruments (on stage) using directivity patterns.

A basic definition of mixing is: a process in which multitrack material, whether recorded, sampled, or synthesized, is balanced, processed, and combined into a multitrack / multispeaker (sometimes referred to as multichannel) format such as stereo, surround, higher order ambisonics, etc. From an artistic perspective, the key objective with mixing is to best convey the performance and emotional features of a musical piece to the listener [3]. The mixing process involves both technical and creative tasks.

The technical tasks are usually associated to technical issues with raw recorded content and its interaction with the rest of the mix. Here are a few examples: many recordings require cleaning up of unwanted sounds such as background noise or buzz from guitar amplifiers, pre-singing coughs, etc. These sounds are often processed or simply trimmed out. Recorded material can suffer from various unwanted phase delays/mismatches (due to microphone placements) which result in dramatic effects; hence it is important to correct them at the beginning of the mixing process. Other technical aspects involve level balancing, EQing and panning tracks to make sure sounds blend well and still cut through the mix without being overly masked.

Creative tasks are linked to artistic choices involved in improving the perceived quality of the mix. A few examples include applying EQ on vocals to give them more presence, tweaking reverb to push back certain sources, using gates to shape timbres of percussive sounds or applying delay or distortion effects on guitar tracks based on artistic choices. These decisions that cater to the artistic vision of the mix are subjective, whereas technical aspects of mixing have objective criteria to some extent and they have evolved into rules or best practices.

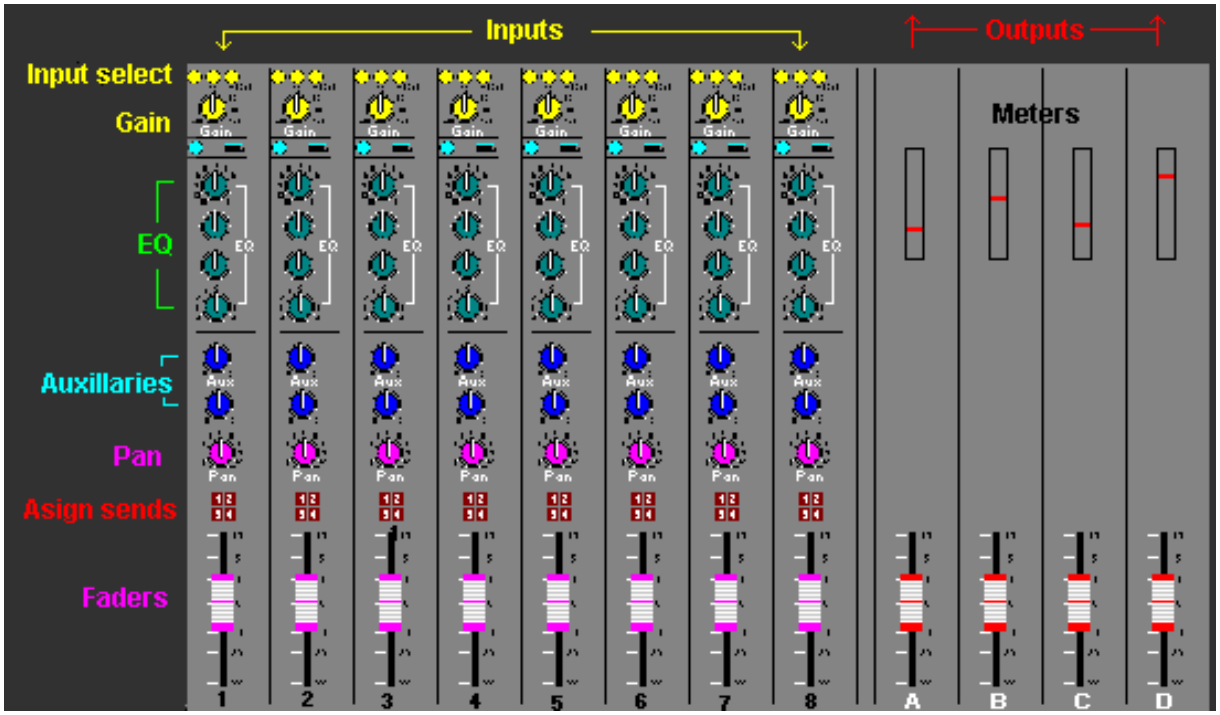

Fig. 2.4 Mixing console (Picture credits: Lenard Audio)

2.3 Signal processors in the mixing pipeline

Various signal processors are used to perform specific alterations to the sound signals. Mix engineers iteratively alter parameters linked to these signal processors to achieve a balanced mix while evaluating it according to its appropriateness to the musical idea and technical constraints. In the recording chain, signal processing is usually carried out on a DAW (Figure 2.3) using software plugins in a studio recording context, or on a mixing console/desk (Figure 2.4) which is usually the case in live scenarios. Effect processors can either be static or time-varying; DAWs have an additional time-varying processing capability in that mix engineers can draw parameter curves for each effect to vary over time (referred to as 'automation curves'). Though mixing consoles have built-in processors, mix engineers usually rely on tapping the signal chain to send them ('assign sends' in Figure 2.4) through outboard gear (hardware). Individual instruments or voices can be directed through any number of signal processors in its respective mixing pipeline/chain.

Similarly, groups of instruments can receive the same processing. The entire recording might also be processed, as is common in the mastering process.

The process of mixing can be divided into 4 main controls: **level**, **frequency**, **time** and **stereo**. Using these basic controls we can also play with higher level controls such as **spatialization**. Mix engineers usually aim to perceive the mix as if it exists on an imaginary sound stage, where instruments can be positioned left and right (stereo) or front and back (depth). In many cases, instruments end up masking (Section 3.6.1) each other and struggle to cut through in the mix. Carrying out unmasking using signal processing (spatial separation of the frequency content in this case) of the sources is the main goal of this dissertation. In this section, the most common signal processors are discussed:

2.3.1 Level control

Maintaining the relative levels / balance of the different sources in the multitrack is crucial in making sure each source cuts through / is heard with intended definition. In mixing consoles and DAWs, levels are usually adjusted using 'faders' (Figure 2.4). Another common effect is the compressor which falls in the category of dynamic processors. A compressor applies a negative gain to the signal whenever it exceeds a threshold. This negative gain is proportional to the signal level that exceeds the threshold, controlled by the *ratio* parameter. *Attack* and *release time* parameters control how fast the compressor reacts with its compensation gains. The *knee* parameter determines a smooth transition from the uncompressed to the compressed region. Other level-based effects are noise gates (reduces noise by attenuating signals below certain threshold levels), expansion (works opposite to compressors, it helps make up levels) and limiters (compressors with high ratio, help in avoiding clipping).

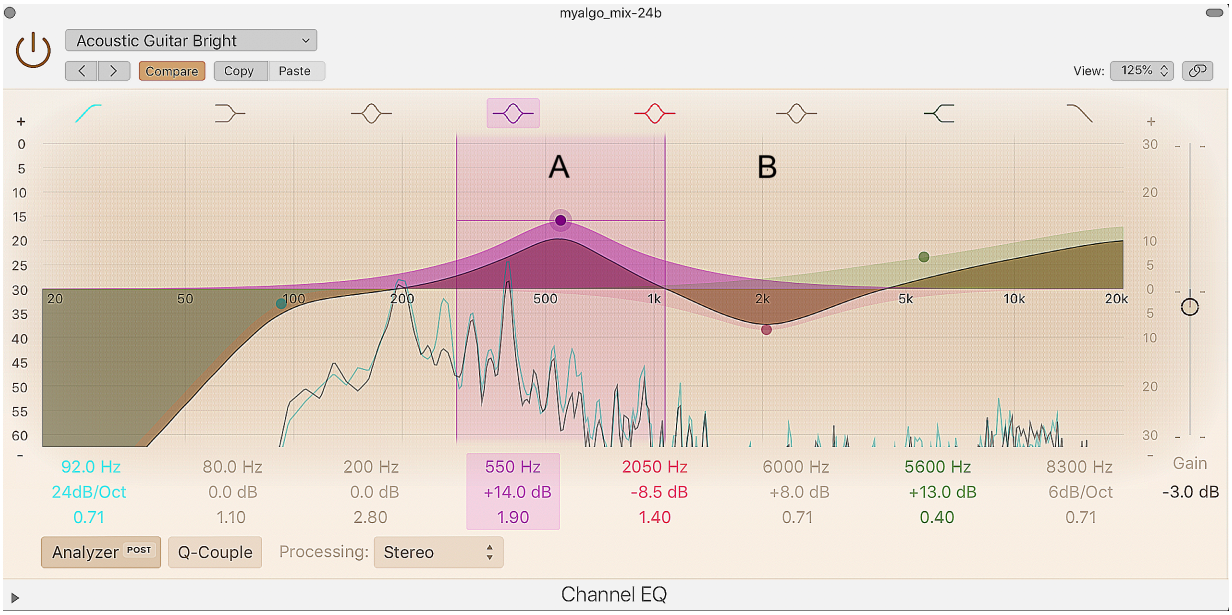

Fig. 2.5 Equalizer (frequency on X-axis and magnitude on Y-axis)

2.3.2 Frequency control

This control includes the filter, commonly called equalizer (EQ), which is one of the most essential processors in audio engineering. EQ allows mix engineers to emphasize and attenuate certain frequency components of the input signal.

Figure 2.5 illustrates some of the most common filters used in an EQ effect, from left to right: ***lo-cut/hi-pass filter*** which blocks/allows all frequency content below/above a certain cutoff point (in this case, 92Hz). ***Peaking filters*** emphasize or attenuate a frequency region around a centre frequency with a specified bandwidth. In the example illustrated in Figure 2.5, the peaking filter is centred around 550Hz (region A), with a quality factor* of 1.9, boosting frequencies between 200Hz and 1000Hz. The filter right next to it (region B), the peaking attenuates the respective frequency content. ***Shelving filters*** alter frequencies above a certain frequency by a fixed number of decibels. In this case, boosting frequency content above 5600Hz by 13dB). These filters help in balancing the spectral content of sources to bring out the desirable timbre out of the recorded signal and also to address spectral masking (Section 3.6.1).

*ratio of centre frequency and bandwidth

2.3.3 Temporal effects

Time-based (delay) effects are created by making single or multiple copies of a source sound, delaying the copies in time and then mixing the two together. These effects are implemented using delay lines. These effects are controlled using 4 parameters: *delay time* - time delay between original signal and delayed version, *feedback amount* - controls repeat amount, *modulation* - controls speed of playback as well as above two parameters to induce time-varying frequency effects. The primary effects in this domain include phasing, flanging, chorus, double tracking, slapback, echo, etc. These effects have specific delay amounts and functionalities.

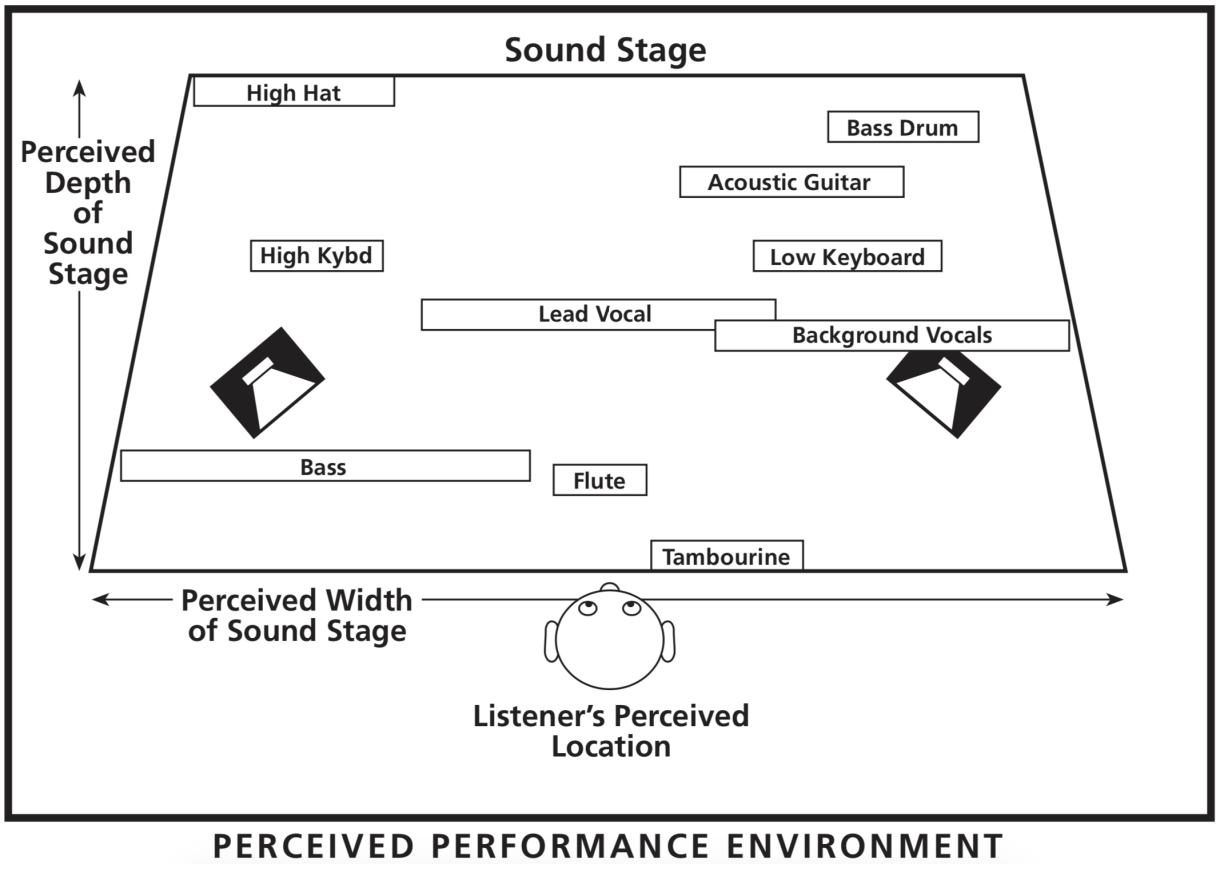

Fig. 2.6 Example of a sound stage with sources placed in certain lateral position and distance with various widths (from [22])

2.4 Spatialization

One of the most important cues in perception of space 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 [24]. The spatial properties of sound play important roles in complying to objective criteria based on best practices as well as in communicating the artistic message of recorded music [22]. Spatial properties may be used in supportive roles to enhance the character or effectiveness of musical ideas, to differentiate one sound source from another, to provide dramatic impact, or to recreate or reinforce reality by providing a performance space for the music. Sounds propagate from a source to the listener and are widely modified by the surrounding environment. Therefore, there are some spatial effects imposed by the physical and geometric characteristics of the environment on the sound signals arriving to the listener's ears. These spatial effects affect the timbre of the sound produced by the sources in the space/room. The most frequently used technique to position sound sources in space is amplitude panning [25]. Modern audio production relies on amplitude panning techniques almost exclusively for the creation of azimuthal cues out of monophonic source signals. **Depth** is simulated using reverberation effects. The aim of using these effects and techniques in the mixing pipeline is to create an aural image / illusion of the sound stage (Figure 2.6). The sound stage encompasses the area where all sound sources are perceived to be located. Each source on the sound stage is localized using lateral location and distance; this will eventually give the sound stage an overall perceived depth and width. Decisions with regard to instrument placement and creating the perceived performance environment in the mix are usually based on artistic choices.

In the following sections various aspects of spatial hearing, spatial audio reproduction systems and spatialization in the context of multitrack mixing are discussed.

2.4.1 Concepts of spatial hearing

The acoustical sound field around us is very complex. Direct sounds, refractions and reflections arrive at the listener's ears. The listener then analyses the incoming sounds and eventually develops the sense of space. Spatial hearing is an important part of the cognition of the surrounding world [26]. Humans associate spatial attributes, such as direction and distance, to auditory objects. We can localize sound sources and perceive some properties of the space they are in, using just hearing. We decode spatial information from different types

of cues: spectral content of ear canal signals, spectral or temporal differences between ear canal signals, and effect of head rotation to perceived binaural differences. We decode the differences of sound between the ear channels and use them to localize sound sources [27]. These differences are called binaural directional cues. Temporal difference is called the interaural time difference (ITD) and level difference is called the interaural level difference (ILD) [14]. Humans are sensitive to ILD at all frequencies, and to ITD mainly at frequencies lower than about 1.5 kHz. At very low frequency, below 100Hz, we do not perceive any stereo effect. ITD and ILD provide information on where a sound source is in the horizontal plane. This phenomenon is known as the duplex theory [28]. Due to the fact that the ears are located on different sides of the head, the arrival times of a sound signal vary with direction. Also, the head casts an acoustic shadow that causes the contralateral ear signal to be attenuated. The pinna and other parts of the ear may also change the sound signal. Head movements have a significant effect on binaural cues and these dynamic cues / information is used in source localization. For example, when a source is in front of the listener, and the listener rotates his head to the right, the left ear becomes closer the source, and the ITD and ILD cues change favouring the left ear.

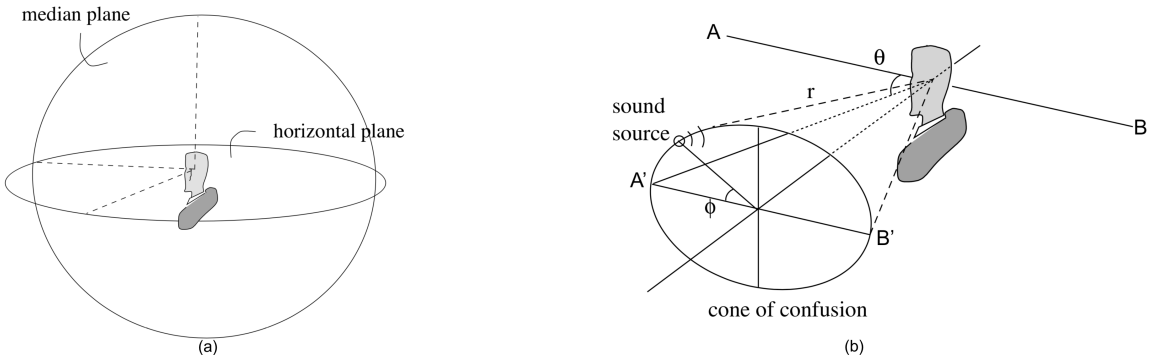

Fig. 2.7 Spherical coordinate system in spatial hearing (from [25]): a) Median and horizontal planes (left), b) Cone of confusion (right). Direction of the sound source is denoted as (θ,ϕ). In the figure, θ is the angle between the source and the listener's axis of symmetry, line segment AB, in the **horizontal plane**, ϕ is the angle between the source and the listener's axis of symmetry AB, (projected as A'B' for clearer illustration), in the **median plane**

2.4.2 Coordinate system in spatial hearing

Two planes are important in spatial hearing and in spatial reproduction, presented in Figure 2.7a [25]. The plane that divides symmetrically the left and right parts of the listener's space is the median / elevation plane. The horizontal / azimuthal plane divides the space into upper and lower parts. All points on the median plane are equidistant from both the ears of the listener, and all points in the horizontal plane share the same height with the ears. Spherical coordinates are often used to denote sound source directions in spatial hearing. Conventionally, they are denoted by distance r, azimuth θ and elevation ϕ. With respect to the listener, azimuth and elevation angles are used to refer to the sources in the horizontal and median planes, respectively, separated by a distance r.

In spatial hearing an important concept is the cone of confusion [29]. The cone is defined as a set of points which satisfy the following condition: the difference of distances from both ears to any point on the cone is constant. A cone of confusion can be approximated by a cone having its axis of symmetry along a line passing through the listener's ears and having the origin in the centre point between the listener's ears, as in Figure 2.7b.

2.4.3 Spatial effects control using panning

Two-channel stereo has been the mainstay for hi-fidelity recording and playback systems since the first wave of stereo media was brought to the marketplace in the 1950s. Stereo systems are designed to create the illusion of a spatial sound scene with directional sound sources localized between two or more loudspeakers placed in front of the listener. Signals of mono sound sources are often mixed into 2-channel stereo programs creating multi-mono signals delivered through two loudspeakers. By duplicating a mono signal and routing it to both loudspeakers of a 2-channel stereo system, a phantom source (sound image/illusion) [30] appears between the loudspeakers. A phantom image may be perceived as a virtual point source, or be spread to exhibit some degree of width. This is carried out using **panning** which basically induces simple level differences between the loudspeakers to evoke auditory objects between the loudspeakers. This distribution of a sound signal (either monaural or stereophonic pairs) into a new stereo or multichannel sound field is determined by a pan control setting, called panpot (Figure 2.4).

In the context of mixing, mix engineers start panning spectrally similar sources by nudging the panning amounts of the individual tracks until a better sense of spatialization,

source distinction and clarity by unmasking is achieved. Some of the aspects linked to creating a stereo image using panning involve: **localization** (concerned with where the sound appears to come from on the left-right axis), **stereo width** (how much of the stereo image the sound occupies. A drum kit can appear narrow or wide, as can a snare reverb), **stereo focus** (a source can appear to be emanating from a very distinct point in the stereo image, or can be unfocused / smeared), **stereo spread** (spatial spread of the sources across the stereo image. For example, the individual drums on an overhead recording can appear to be coming mostly from the left and right, and less from the centre). There are several objective criteria to keep in mind while panning [4]: maintaining balance between left and right, stereo frequency balance, spreading masked sound sources, etc. These points are discussed in detail in Section 4.1, and these criteria form the basis for developing the automatic spatialization system for stereo. With the recent advancements in more affordable multichannel sound reproduction systems and efficient virtual acoustics synthesis algorithms, it is important to consider how sound sources will be spatialized beyond stereo. The later sections of this chapter cover background theory and technology behind spatial audio systems.

2.4.4 Depth control using room effects and reverberation

The spatial cues discussed above only consider direct sound coming from the source to the listener. However, there exist reflections and reverberation in real rooms and outdoor spaces. In a reverberant environment, sounds reach the ears through several paths. Although the direct sound is followed by multiple reflections, which would be audible in isolation, the first-arriving wavefront dominates many aspects of perception. This is known as the Precedence Effect [28] which is a suppression of early delayed versions of the direct sound in source direction perception. We can estimate the size of a room and even surface materials of objects and walls in the room by listening to sounds. Our perception relies on the density of the reflections and the length of the reverberation. Listeners use these cues to perceive distance from the source. When we hear less of the direct sound and more of the reverberant sound, we perceive the source to be far. This is quantified with the direct-to-reverberant ratio (DRR) of sound energies expressed in decibels.

In the context of mixing, depth control via effects like reverberation are usually artistic choices. Reverberation effects are either based on convolution in which measured impulse

responses in rooms are convolved with the dry signal recorded anechoically or synthesized using structural models like delay lines. Reverb enables us to create a sense of depth in the mix and further serves as a tool to position sources in the sound stage. A mix in which instruments appear close (higher DRR) is considered tight, and a mix with extended depth (more reverb resulting in lower DRR) is considered spacious. The objective criteria for reverberation is to make sure the depth assigned for the sources are coherent. For example, percussive elements like hi-hats are usually best represented in the mix when they are placed close or sounds tight in the mix, while orchestral elements like legato violin sections could use some effects like hall reverb so that they are less stereo focused and more stereo spread (terms discussed in Section 2.4.3) in the mix (like it would sound in a concert hall). The main objective is to create something natural or otherwise artificial but appealing [3]. In the following section we will see how spatial effects can also be utilized for artificial upmixing.

2.5 Spatial sound reproduction

Bringing a virtual three-dimensional sound field to a listening situation is one goal of the research in the field of audio reproduction [26]. The first recordings were monophonic; they created point-like sound fields. A big step was two-channel stereophonic reproduction, with which the sound field was enlarged to a line between two loudspeakers. Two-channel stereophony is still the most widely used reproduction method in domestic and professional scenarios. There has been various attempts to enlarge the reproduction sound field. In most systems, the loudspeakers are situated in a 2D horizontal plane to create pantophonic sound fields (horizontal-only). Some attempts to produce periphonic (full-sphere) sound fields with 3D loudspeaker placement and headphone listening exist, using Higher Order Ambisonics (HOA) [31], Wave Field Synthesis (WFS) [32], Head-Related-Transfer-Function (HRTF) modelling [33], etc.

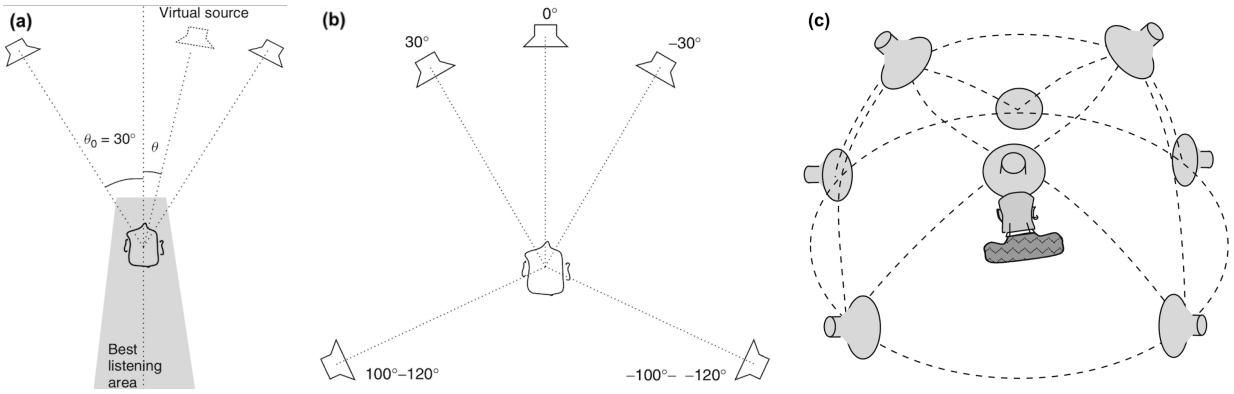

Fig. 2.8 Multispeaker systems layouts (from [27]), from left to right: a) Standard stereophonic listening configuration, b) 2D speaker layout, c) 3D speaker layout

2.5.1 Multichannel methods : Amplitude panning

Amplitude panning is the most frequently used virtual-source-positioning technique [27]. The aim is to make the listener perceive an illusion of a single auditory event (virtual / phantom sound source). Panning involves directing a sound signal, $s(t)$ to loudspeakers or output tracks with different amplitudes, formulated as follows (2.1):

$$s_i(t) = g_i s(t) \tag{2.1}$$

where, s_i (t) is the signal to be applied to loudspeaker i = 1,...,M_o, g_i is the gain factor of the corresponding output track, M_o is the number of output tracks / loudspeakers, and t is the time.

In **Stereophonic listening** there are two loudspeakers placed in front of a listener, as illustrated in Figure 2.8(a). If the listener is located at equal distances from the loudspeakers, the panning law estimates the perceived direction θ from the gain factors of loudspeakers. There are several published methods to estimate the perceived direction, one of them being the tangent law by Bennett et al. [34] formulated as (2.2) :

$$\frac{tan\theta}{tan\theta_o} = \frac{g_1 - g_2}{g_1 + g_2} \tag{2.2}$$

where θ is the perceived azimuth angle (panning angle) of the virtual source, θ_o is the loudspeaker base angle (Figure 2.8a), g_i is the respective panning gain for each output

track. The tangent law is based on a simple geometrical head model and is still the most popular panning law for pairwise panning [30]. The panning laws set only the ratio between the gain factors. To prevent undesired changes in loudness of the virtual source depending on panning direction, the sum-of-squares of the gain factors is normalized. In principle, the amplitude-panning method creates a comb-filter effect, as the same sound arrives from both loudspeakers to each ear at different times creating cross-talk [27].

The most commonly employed amplitude panning methods are Vector-Base Amplitude Panning (VBAP) [26], Multiple-Direction Amplitude Panning (MDAP) [30] and Distance-Based Amplitude Panning (DBAP) [35].

VBAP [26], is the generalization of the tangent law for amplitude panning in two-channel stereophony. VBAP is a method to calculate gain factors for pair-wise or triplet-wise panning, discussed below. In VBAP the number of loudspeakers can be arbitrary, and they can be positioned in an arbitrary 2-D or 3-D setups. VBAP produces virtual sources that are as sharp as possible with current loudspeaker configuration and amplitude panning methods, since it uses the minimum number of loudspeakers needed, one, two, or three at a time.

In **2-D loudspeaker setups** (Figure 2.8b) all loudspeakers are on the horizontal plane. Pair-wise amplitude panning [36] is the best method to position virtual sources with such setups, when there are more loudspeakers ($4 < M_o < 20$). In pair-wise panning the sound signal is applied only to two adjacent loudspeakers of the loudspeaker setup at one time. The pair between which the panning direction lies is selected. The number of active loudspeakers depends on the panning direction: two loudspeakers are active for directions between two loudspeakers and one is active for directions coinciding with a loudspeaker. **3-D loudspeaker setups** (Figure 2.8c) introduce height channels and use triplet-wise panning in which up to three speakers are divided into triangles (triangulation [37]) to pan a single virtual source.

A drawback of pair- and triplet-wise panning is that the spread of a virtual source depends on panning direction due to different numbers of loudspeakers producing the same signal. When there is a loudspeaker in the panning direction, the virtual source is sharp, but when panned between loudspeakers, some spreading occurs. This can be avoided by using **MDAP** [38]. In this technique, gain factors are calculated for multiple panning directions around the desired panning direction. The virtual signal is not applied to all loudspeakers, but a subset of them. The directionality therefore does not degrade as much

as it degrades with systems that direct a same sound signal to all loudspeakers. **DBAP** [35] offers an alternative matrix-based spatialization method where no assumptions are made concerning the layout of the speaker array (it takes the actual positions of the speakers in space as the point of departure) nor the position of the listener, thus making it useful for several real-world situations.

As stereo began to reach commercial viability in the late 1950s, work was already progressing on more ambitious multichannel audio formats. The most popular ones are surround systems like **5.1 and 7.1** which provide perceptual benefits over the other multichannel formats like stereo and quadraphonics [39]. The front centre channel anchored the stereo soundfield for off centre listeners and provides better tonal balance over a phantom centre [40].The rear channel positions provide a balance between the reconstruction of lateral energy and panning of sources to the rear of a listener and in addition the ability to envelope the listener using the rear channels, or to generate a new artificial room acoustic [41]. Further, Stuart states that the move from stereo to multichannel is significant, providing better sound source segregation with lower masking thresholds [42]. A 5.1 system (Figure 2.8b) includes a stereo pair and centre speakers on the front, stereo pair on the rear and a low-frequency effects (LFE) channel (subwoofer). 5.1 coarsely quantizes a soundfield and panning between the front and back speakers to the side of the listener causes sounds to jump between front and rear channels. 7.1 evolved from 5.1 by adding two extra speakers surround left and surround right, primarily designed for cinema applications with wide screens to fill in the gaps between the front and rear channels. There have been notions of adding height to surround systems for many years now. A surround system with height, which has been the subject of a large amount of research, is the **NHK 22.2 system** [43]. This system uses 10 channels at ear height, 9 channels above the listener and 3 frontal lower channels along with 2 subwoofers.

2.5.2 Wave Field methods : Ambisonics and Wave Field Synthesis

While amplitude panning techniques like VBAP and DBAP can achieve convincing spatial effects under certain conditions, both of these methods fundamentally encode audio output in relation to a specific arrangement of loudspeakers, and are thus classified as multispeaker/multichannel methods [44]. The term wave field methods refers to those spatial formats that seek to encode an entire sound field, regardless of the arrangement of

output transducers. This is done via Huygen's Principle, which states that each point on a progressing wavefront may instead be considered as a separate source [32]. Wavefront methods tend to be broadly separable into the following: 1) Ambisonics, which reproduces the incoming sound field around the listener, and 2) Wave field synthesis, which reproduces the outgoing sound field emitted by one or more acoustic sources.

Ambisonics was conceptualized by Gerzon in 1973 [45]. It is a complete method of recording transmission and reproduction of not only horizontal surround but also what is termed 'periphonics' or full sphere reproduction. Ambisonics is basically a microphoning technique. However, it can also be simulated to perform a synthesis of spatial audio [46]. In this case it is an amplitude panning method in which a sound signal is applied to all loudspeakers placed evenly around the listener with gain factors as follows (2.3):

$$g_i = \frac{1}{M_o}(1 + 2\cos\theta_i) \tag{2.3}$$

where, g_i is the gain of the i^{th} speaker, M_o is the number of loudspeakers, and θ_i is the angle between loudspeaker and panning direction. Second-order Ambisonics makes use of an additional term $2\cos 2\theta_i$. The sound is applied to all of the loudspeakers, but the gains have significantly lower absolute values on the opposite side of a panning direction. However, to get an optimal result, the loudspeakers should be in a symmetric layout, and increasing the number of them would not enhance the directional quality beyond a certain amount of loudspeakers. Ambisonics is utilized to reproduce a sound scene by either recording using a sound field microphone (Eigenmike [47]) or by synthesising using ambisonic encoding equations [31]. There are other aspects such as ambisonics decoding and normalization schemes considered in ambisonics systems [39].

Wave Field Synthesis (WFS) uses the same principle as ambisonics to achieve the opposite aim: given an infinite amount of microphones around an acoustic source and an infinite number of loudspeakers in the same arrangement, each producing the signal from its respective microphone, the wave fields in both cases should be identical [39]. It reconstructs the whole sound field within a listening room. Theoretically it is superior as a technique, but unfortunately it is impractical in most situations. The most restricting boundary condition is that the system produces the sound field accurately only if the loudspeakers are at a distance of maximally half a wavelength from each other. The centroids of loudspeakers should thus be a few centimetres from each other to be able to produce high frequencies

correctly; this cannot be achieved without a very large number of loudspeakers. Such systems have been constructed using roughly 100 loudspeakers. Accurate spatial reproduction is typically limited to about 1000 Hz [25]. The benefit of WFS is that the reproduction of the source is not only valid at one point in space but at any point within the whole area, delimited by the speaker array. One drawback of WFS is the fact that spatial aliasing occurs above a certain frequency caused by the physical distance between the transducers, thus correct frequency reproduction can only be guaranteed up to a frequency limit.

2.5.3 Binaural methods : Headphone listening

Considering the high cost of multichannel speaker systems, in terms of price, flexibility, space and other logistics, simulating the above methods on headphones is an efficient alternative. Most tracks produced for stereo can be reproduced reliably on consumer headphones since audio engineers mix and monitor on a stereo pair of headphones and/or monitor loudspeakers. The mapping of the stereo output tracks are directly carried out to the pair of small transducers on the headphones. However, implementing reliable 2D/3D audio reproduction over headphones (binaural methods) is challenging. Binaural techniques are loosely defined to be methods which aim to control directly the sound in the ear canals to match a recorded real case or with a simulated virtual case. This is done by careful binaural recordings, or by utilizing measured or modelled head-related transfer functions (HRTFs) [33] and acoustical modelling of the listening space, also known as auralization [48]. An HRTF is a response that characterizes how an ear receives a sound from a point in space. As sound strikes the listener, the size and shape of the head, ears, ear canal, density of the head, size and shape of pinna, nasal and oral cavities, all transform the sound and affect how it is perceived, boosting some frequencies and attenuating others. An interesting application for HRTF technologies with headphones is listening to existing multichannel audio material [49]. In such cases, each loudspeaker in the multichannel loudspeaker layout is simulated using an HRTF pair. A monophonic sound signal can be positioned virtually to any direction, if HRTFs for both ears are available, for a desired virtual source direction. For example, a signal meant to be applied to the loudspeaker in a 30^o direction is convolved with the HRTF pair measured from the same direction, and the convolved signals are directed to the headphones. The method simulates the ear canal signals that would have been produced if a sound source had existed in the desired direction. The head movements

of a listener should also be taken into account in processing, otherwise the sound stage will appear to be moving, causing inside-head localization.

2.5.4 Up-mixing and down-mixing

Audio mixing is performed for the purposes of down mixing or up mixing [50]. Down mixing is used to reduce the number of input tracks into a composite output mix with fewer output tracks, and upmixing is performed when the resulting mix has more output tracks than input tracks. An example of up mixing is the case of panning a monaural signal to achieve a false sense of stereophony. In most cases, each input consists of a single track recorded monaurally. In more complicated scenarios, for example recording an acoustic guitar or drum kit, there can be multiple microphones recording each source and they are then mapped to single or multiple output tracks. In practice, both mixing procedures are used together. For example, we may up-mix each of the monaural sources into a two channel source by panning (rebalancing the signal to different output channels), and then we may down-mix sets of two-channel signals into a single two-channel down-mix.

The majority of tracks produced are targeted for stereo, wherein each source is recorded with a single or a group of microphones and the energy balanced across the stereo width between two tracks, left and right, or stereo. These stereo tracks are played back on headphones, home loudspeakers, etc. When this has to be played back on higher channel systems like 5.1 surround or car audio systems, the stereo mix-down is up-mixed in an artificial manner using signal processing, for example : ambience extraction of a stereo track to produce audio content for the rear channels in a surround system [51]. In this thesis we address the situation of having access to the multitrack and the spatialization systems would place the audio content strategically on the different channels of the chosen output format.

Chapter 3

Automatic Mixing for Multitrack Spatialization based on Unmasking

In this chapter we give an overview of automatic mixing and dive into the state-of-the-art system for automatic spatialization of multitrack audio.

3.1 An automated systems approach to mixing multitrack audio

Mixing is one of the crucial aspects of music production, within which the sound signals from different sources are combined to form a coherent piece of music, called the 'mix'. Multitrack mixing is an iterative process in which various processing parameters such as loudness balance, EQ and compression are adjusted to achieve a certain target output mix that complies to perceptual and objective criteria [52]. In the previous chapter we presented the system used to mix these signals, which included a mixing console or a DAW, a chain of signal processing tools and spatial audio systems to playback the output mix. Advances in digital mixing technology have been significant in recent years, in part due to the rapid increase in computational power [53]. One effect of this is that music production has become far more accessible. Amateur music producers, particularly musicians, are now able to do all stages of the production process in an affordable manner without having to invest heavily in studio time and equipment. Another effect of these advances is an increase in the complexity of the mixing tools, which places a greater technical burden on amateur and professional mixing engineers alike. Advances in music production technology have led to a recent surge of interest in the field of research known as **automatic mixing**

to address the above challenges. The idea is to analyze the relationship and interaction between tracks to automate the mixing of multitrack audio content. These are 'intelligent' / expert systems [5] that perceive, reason, learn, and act intelligently. This implies that they must analyze the signals upon which they act, dynamically adapt to audio inputs and sound scene, automatically configure parameter settings, and exploit best practices in sound engineering to modify the signals appropriately. They derive the parameters in the editing of recordings or live audio based on analysis of the audio content, and based on objective and perceptual criteria.

For progress towards automix systems in these sound engineering domains, significant problems must be overcome that have not yet been tackled by the research community [5]. Yet multitrack signals are pervasive, and the interaction and dependency between output tracks plays a critical role in audio production quality [54]. Considering the advent of 3D audio technologies, it is important to develop automatic mixing tools using new, multi-input multi-output audio signal processing methods, which can analyze the content of sources in order to improve the quality of capturing, editing and combining multitrack audio for playback on spatial audio systems (section 2.5). Automatic mixing tools should enable non-experts to mix live music events, in which the practical issues are accounted for by the automatic system; and should help musicians to produce good quality mixes of their work, without the need for them to delve too deeply into the technical complexities of mixing. Furthermore, for professional applications, the automatic mixing tools should lighten some of the functional burden from the production process since technical tasks can interrupt the creative flow. Automatic mixing tools should be able to assist the engineer's decision making process by providing good prior parameter settings (based on technical/objective criteria). This will enable the mix engineer to concentrate on the more creative aspects of the mixing pipeline/process (to satisfy perceptual/subjective criteria). Generally, there is economic, technological and artistic merit in exploiting the immense computing power and flexibility that today's digital technology affords [8].

3.2 Automatic mixing approaches

A number of different approaches have been employed to address the research problem of automatic mixing. They are organized using these three labels as outlined in [55]:

3.2.1 Machine learning approach

In machine learning approaches, the system is trained on initial content from a database of sample mixes (for example, the Mix Evaluation Dataset [56]), to infer how to manipulate new content. The idea is to analyze the evolution of parameter settings of multitrack content in the dataset across several mixes (datapoints) and train the machine/deep learning model. An example of this is described in [57], in which a machine learning system used both individual tracks and final mixes of 48 songs as dataset. This was used as training data, and it was then able to apply time varying gains on each track for new content. However, this approach is limited due to the lack of well labelled datasets of multitrack content, which also sometimes face issues related to copyright.

3.2.2 Grounded theory approach

Grounded theory approach for automatic mixing and its methodology may be used to acquire basic knowledge which may subsequently be transferred to the automatic mixing system. In the context of audio production, this suggests psychoacoustic studies to define mix attributes, and perceptual audio evaluation to determine listener preference for mix approaches. An important downside of this approach is that it is very computationally intensive, considering that many of the psychoacoustic models are non-linear and complex. Though there has been some initial work in this area [11], it is too limited to constitute a sufficient knowledge base for the implementation of a robust automatic mixing system.

3.2.3 Knowledge engineering (KE) approach

A traditional approach to automatic design would exploit knowledge engineering which assumes that the rules are already known and they should just be integrated in the automatic mixing system. This involves integrating established knowledge and best practices into the rules and constraints under which the system operates. However, best practices in recording production are generally not known. The main resources to gather the best practices are from well established books, prior music production experiences and advice from peers and teachers in sound recording. This dissertation made use of all of the three resources to gather the best practices. The works by Senior [58], Izhaki [3], Moylan [22] and Eargle [21] provide a good amount of best practices and approaches that mix engineers can use in their production practice. Another challenge of using this approach is to under-

stand how to translate subjective descriptors like "the mix should not sound *muddy*", "the cymbals sound too *harsh*", etc. This information and preferences are not easily acquired or effectively transferred to the automatic system. Some work has been done on mapping high-level, subjective descriptors (such as 'bright', 'harsh', etc.) to lower level audio processing parameters [59,60], but putting this to use in an autonomous mixing system is less than obvious.

In this dissertation, we follow a knowledge-engineered approach, acquiring knowledge from practical literature. Though the knowledge-engineered approach is limited in that there are no well-defined rules available for a certain instrument or for a certain processor, it serves as a good starting point for building automatic mixing systems. Carrying out subjective evaluation of the auto-mixes against produced mixes, will help provide good evaluation criteria and further fine-tune such systems.

3.3 Automatic mixing : architecture and building blocks

From the point of view of signal flow, an audio mixer (mixing console and DAW) is composed of several chained audio processing effects. In the digital audio world, each audio effect has several recallable control parameters. Each individual audio effect takes in an unprocessed input signal and outputs a processed signal. Users control the signal processing parameters in order to produce the desired transformation of the input signal. Currently, multitrack audio editing tools demand manual intervention. Although audio editors are capable of saving a set of recallable static scenes [61] for later use, they lack the ability to take decisions based on the audio scene, such as adapting to different acoustic environments or different set of inputs. Rather than having sound engineers manually apply audio effects to all audio inputs and determine their appropriate parameter settings, automatic mixing systems can apply adaptive digital audio effects [9]. This will aid or replace the task normally performed by the user. Parameter settings of adaptive effects are determined by analysis of the audio content, where the analysis is achieved by a feature extraction component built into the effect. Thsese effects are then applied to the tracks either in a static or time-varying manner. Thus, intelligent audio effects [50] (of an automix system) may be used to set the appropriate equalization, automate the parameters on dynamics processors, and adjust panning amounts of each track to more effectively distinguish the sources, for example. It is also important to consider real-time operability of these au-

tomix systems, for live scenarios. Designing real-time systems involves a whole new set of challenges related to system stability, latency, etc. Off-line processing has the advantage of having access to all the time-frequency content of the multitrack which means it can utilize costlier optimization techniques and use more reliable perceptual models to conduct more informed feature extraction and processing. In this section, the building blocks of an automatic mixing pipeline, involving feature extraction and effects processing of the multitrack content are discussed.

3.3.1 Feature extraction

The feature extraction block is in charge of extracting a number of features per input track. The features measured are usually low-level ones which are further used to obtain source inter-dependency features. The ability to extract the features fast and accurately will determine real-time operability of the automix system. On the other hand, performance of the automix algorithm also depends on the reliability of the features. The closer the feature extraction model gets to perception, the better the algorithm's performance and better the automix system can mimic professional sound engineering practices. Automix system tools are classified as follows [50]: **accumulative** tools achieve data values and derive the features over time to converge to a static value for the processing parameters, and **dynamic** tools makes use of fast extractable features to derive processing parameters in real-time. In practice, we can compromise between dynamic and accumulative feature extraction by using relatively small accumulative time windows with dynamic effects processing.

3.3.2 Cross-adaptive digital audio effects

In multitrack mixing, effects processing is performed on a given signal source not just because of its own content but also because there is a simultaneous need to blend it with the content of other sources, so that a high quality mix is achieved. The cross-adaptive processing section of the automatic mixing tools is in charge of determining the interdependence of the input features in order to output the appropriate control data. The obtained control parameters are usually interpolated before being sent to the signal-processing portion of the automatic mixing tools [5]. As seen in Figure 3.1a, the features are extracted from all tracks are sent to the same feature processing block, where the control parameters for the effects are produced. The cross-adaptive feature processing is implemented by a set

of constrained rules that consider the interdependence between tracks. The output tracks are summed to produce the final mix. This system sometimes makes use of feedback loop from the final mix back to the features extraction and processing block to further fine-tune the control parameters [9]. This control approach to audio processing gives great design flexibility and adaptability to control of effects.

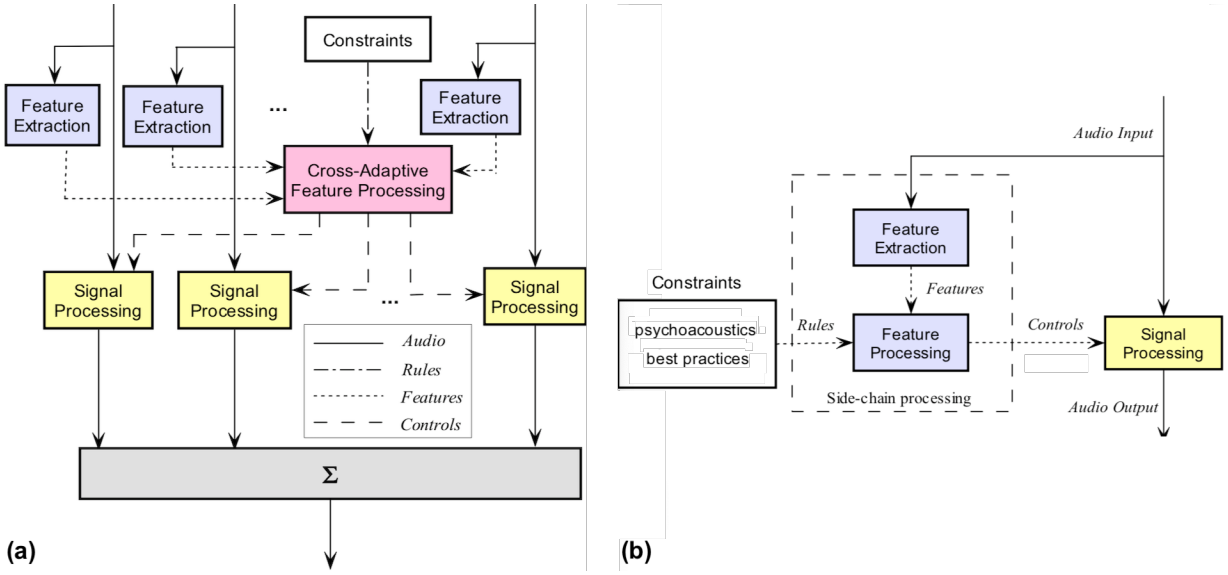

Fig. 3.1 Block diagrams : a) Cross-adaptive mixing system (left), b) Feature extraction and effects processing (right) [5]

3.3.3 Side-chain processing

The signal processing involved in the context of multitrack mixing is quite complex. Achieving the above discussed tasks in real-time is almost impossible. Real-time processing is a significant factor to be considered in music production, especially in live performance scenarios. For this reason, side-chain processing [5] is implemented and performed: the audio signal flow takes place in a normal manner in the signal processing device, while the required analysis, such as feature extraction and classification of the running signal is performed in a separate analysis instance. Once a decent amount of certainty is achieved on the analysis side (feature extraction), the control signals that are prepared meanwhile are sent to the signal processing side to trigger the desired parameter control change command. Figure 3.1b illustrates a side-processing chain for an audio effect, in which features are extracted by

analysis of the audio signal and are processed based on a set of constraints, which come from the three approaches discussed in section 3.2.

3.4 Automatic Spatialization based on Spectral Unmasking

Research into automatic mixing systems has grown rapidly over the last ten years, with intelligent systems proposed for almost every aspect of audio production [5,8,62]. Intelligent tools that analyze the relationships between all tracks in order to automate the mixing of multitrack audio content have been devised. De Man et al. [8] gives an overview and a list of research work that has dealt with various multitrack mixing processes like level, panning, EQ, compression, etc. Spatially separating sources (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 [63]. In this dissertation, we deal with spatialization aspects of multitrack mixing.

One of the most important tasks in audio production is to place sound sources across the stereo or sound 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. 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.

3.4.1 Previous work

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 Mansbridge et al. [10] 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.

Enrique et al. [11] proposed a semi-blind stereo panning system in which tracks could be given priority such that tracks of decreasing priority started to get alternately placed in 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 [46], harsh panning of an instrument on one side is often not preferred [4], and spectral centroid may not be an ideal descriptor of frequency content [10]. Since the above two techniques pan the track as a whole to either side, low frequency content may be panned, causing unwanted spectral imbalance.

Pestana et al. [12] took a different approach, in which different time-frequency bins of each multitrack are assigned different spatial positions in the mix. This approach is unique among other automix tools since it does not emulate traditional panning practices in which a source is panned as a whole to either stereo output tracks. This algorithm could be classified into the second category of automatic mixing systems, as mentioned below. Time-frequency based decomposition and modification techniques have also been used for spatial enhancement in [64]. The panning approach taken in this research uses the similar concept of frequency-based spreading using panning filters. Though the previous works addressed and aimed at reducing masking, there were no objective measures that optimized unmasking amount.

3.4.2 Spectral spatialization

Automatic mixing could be classified as follows:

- achieving an autonomous computerized mix by mimicking the iterative and adaptive approach of a sound engineer,
- 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. Many aspects of sound spatialization obeys standard rules: a stereo mix should be bal-

anced, hard panning should be avoided, etc. Sources with similar frequency content should be placed far apart in the listening field in order to improve the intelligibility of the audio content [58]. 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. Wakefield et al. [62] showed that this avoidance of spatial masking may be a far more effective way to address general masking issues in a mix than alternative approaches using equalizers, compressors and level balancing. Ronan et al. [1] 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.

In the case of positioning sources in a stereo field according to the first classification, 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. This mixing step is carried out using the panpot [46], discussed in Section 2.4.3. A limitation of balancing using a panpot is that the sources remain tied to one side and the width may be restricted [4]. 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 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 [12]. The relationship of ITD and ILD to SRM is not fully studied, but it seems amplitude panning is a sensible choice [65].

This dissertation introduces an automatic spatialization system based on the second classification of automatic mixing. It uses optimized frequency-based panning filters to carry out spatialization and spectral unmasking. The frequency content of the sources are spread across the output tracks in a coherent manner. The first approach is a knowledge-engineered one (Section 3.2.3) and is aimed for stereo. The automix system generates hand-crafted sinusoidal panning filters for each source; alternate frequency bands of each track are placed across the stereo field. The features of the panning filters comply to perception and the best practices in panning. The second approach is aimed at formats beyond stereo. This approach uses a more natural and acoustical approach to determine frequency spread, by using directivity as features of the source [18]. Directivity patterns

quantify directional dependent behaviour of sources thus guiding the distribution/spreading of spectral content on targeted mixing tracks of the sound output format [19]. The resulting frequency response from the directivity patterns forms a more natural panning filter that can be applied to the respective sources to carry out spectral panning, thus replacing ad-hoc rules to generate the panning filters in the first approach.

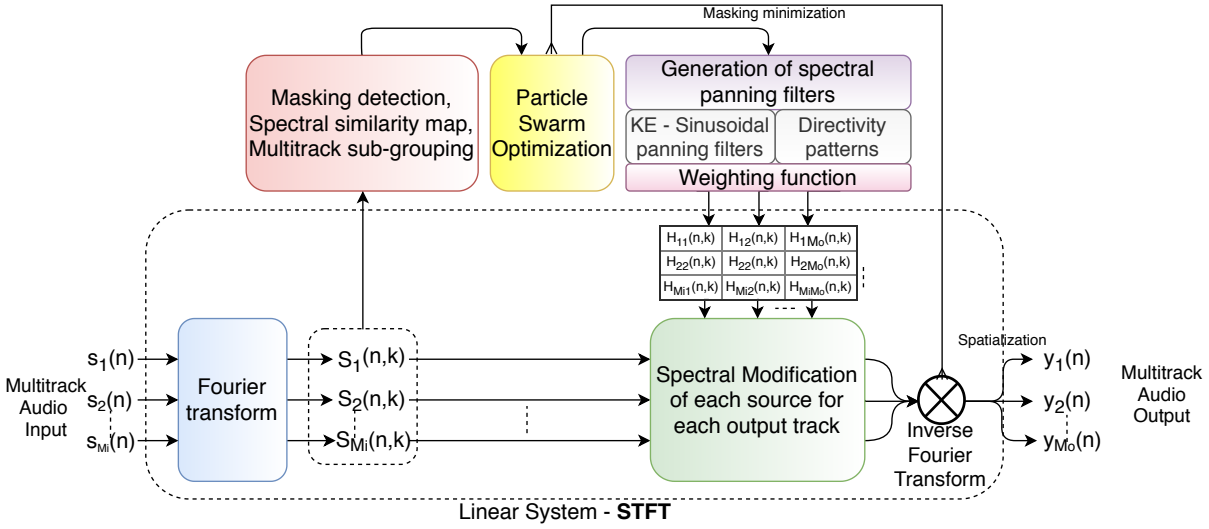

Fig. 3.2 Block diagram - Automatic Spatialization of Multitrack Audio

3.4.3 Framework, Methodology and Implementation

In this section, we present the techniques and concepts used to carry out automatic spatialization and masking minimization. The framework that carries out spectral spatialization is common to both the approaches. The input to the system is M_i monophonic tracks that need to be spatialized. The time frames of each signal, $s(n)$ are transformed to frequency domain, $S(n,\ k)$ using a Short-Time-Fourier-Transform (STFT) framework [66]. The three main blocks of this automatic spatialization system is the following: **masking detection** (feature extraction), generation of **spectral panning filters** for all the time frames using optimization for masking minimization across output tracks (cross-adaptive effects processing), **spatialization and masking minimization** by applying these optimized panning filters on the tracks across time (side-chain processing) to generate the final mix of M_o. The optimizer used in this automix system is particle swarm optimization [20]

(a population-based stochastic optimization technique, detailed explanation in Appendix A) to carry out frequency-based spreading of masked tracks constrained by best panning practices (approach 1) and constrained by source directivity (approach 2) to eventually generate the least masked, spatialized mix. Users will be able to use either constraints or a mix of both using the weighting function. Figure 3.2 illustrates the block diagram of the whole system. Following sections discuss each individual block in detail.

3.5 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. Those tracks that are meant to appear as a single/blended sound image/stream are grouped and summed together as a single monophonic track. 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 (STFT, Equation 3.1):

$$S(qI, k) = \sum_{m=-\infty}^{\infty} s(m)w(qI - m)e^{-i2\pi mk/N} \tag{3.1}$$

with STFT frame qI, discrete frequency bins k and a window function w of length N.

As far as STFT parameters are considered, a viable reconstruction of a spectrally processed signal depends upon a careful choice of windowing parameters [12] as well as proper choice of hop size. Having no/little overlap results in clicking noises or artifacts and too much overlap will cancel out the desired effect, resulting in narrow panned mixes [11]. This is explained in more detail in section 3.5.2. In this dissertation, the automix system uses STFT parameters from a past dynamic spectral panning work [12] which followed a heuristic approach to determine the following: long window size of 2^{15} with a hop size $I = N/16$ (for sample rate $f_s = 44100$ Hz).

3.5.1 Spectral modifications : time-varying panning filters

Control parameters of traditional multitrack effects do not vary over time. However, when it is really needed, the automation feature in DAWs allow mix engineers to provide automation

tracks, which is a time-line representation of the state of a parameter [50]. For example, if an electric guitar track sounds too bright during a piece while producing some high harmonics, the mix engineer might apply a notch filter at those high frequencies during those time instants before switching back to the default EQ setting for the guitar, using the automation feature. In the context of panning, when there is dense sound activity during a particular instant in the piece, the masked sources can then be spatially separated by panning / spreading them out in the stereo field. Automix systems (as per the second classification) have features that can address such effect-processing tasks using one or more objective criteria.

The integral part of the automix system is the spectral modifications carried out on each monophonic signal, $s(n)$ for the respective M_o output tracks. To carry out spatialization, the spectral modification block multiplies the optimized frequency response H (panning/spatialization linear system) with the spectrum S of signal s and inverse Fourier transformed to contribute to a signal y produced on an output track. In the next 2 chapters, both the approaches to generate these frequency responses are presented: 1) KE-based sinusoidal filters and 2) directivity patterns, both of which depend on the masking activity in the multitrack (discussed in later sections of this chapter). The frequency response is time-varying: each STFT frame has a corresponding frequency response $H_{m_i m_o}$ $(qI,\ k)$. The frequency responses $H_{m_i m_o}$ of the m_i tracks are different across the m_o output tracks. These level differences, or distribution of spectral content across the output tracks are what creates the effect of spatialization while accounting for unmasking. The inverse FFT of the overlapped-and-added STFT frames after spectral modifications of all the M_i sources for each output track m_o is as follows (Equation 3.2). The ratio $W(0)/I$ (where W is the Fourier transform of window w) is multiplied to satisfy the gain correction factor of OLA.

$$y_{m_o}(n) = \frac{I}{NW(0)} \sum_{m_i=1}^{M_i} \left[\sum_{k=0}^{N-1} S_{m_i}(qI,k) H_{m_i m_o}(qI,k) e^{i2\pi nk/N} \right] \tag{3.2}$$

3.5.2 System stability and limits

One major aspect to take care of while building automix systems is that the evolution of parameters over time should be realistic and within limits of the STFT framework to avoid aliasing. As discussed in Section 2.3, when it comes to source localization, mix engineers usually aim to perceive the mix as if it exists on an imaginary sound stage, where

instruments can be positioned left / right / front / back. Automix systems could end up moving around these sources over time in an unrealistic manner, just to satisfy objective criteria. For example, we certainly do not prefer hearing a guitar track on one side of the stereo field in one time instant and on the other side in the next instant, without any artistic intent. Therefore it is important to set up constraints on dynamic automix systems to make sure the effect's control parameters change over time in a meaningful manner.

Another consideration to be kept in mind while implementing the STFT system is the selection of the rate at which $S(n,\, k)$ should be sampled in time n and frequency k to avoid aliasing [66]. The proposed automix system performs time-varying spatialization (panning filter computed for each STFT frame) before synthesizing the modified signal, it is important to make sure no aliasing occurs. The Fourier transform of $\mathrm{w}(n)$ in Equation 3.1 is a low-pass frequency response of bandwidth B in Hz. Therefore, the frequency bandwidth of $S(n,\, k)$ at a given channel k is the same than the one of the window, and thus according to the sampling theorem, $S(n,\, k)$ must be sampled at a rate of at least $2B$ samples per second to avoid aliasing. In this STFT framework, we use a Hamming window [67] of N samples, hence B (in Hz) would be calculated as follows:

$$B = \frac{2f_s}{N} \tag{3.3}$$

The synthesis method reconstructs the signal by overlap-adding (OLA) the individual time responses due to each analysis frame, with appropriate time shifts. The OLA technique requires analysis to be performed every $2B$ samples [66]. The total number of samples of $S(n,\, k)$ that is to be computed per second for N-point Hamming window $\mathrm{w}(n)$ is as follows:

$$(N)(2B) = (N)(\frac{4f_s}{N}) = 4f_s \tag{3.4}$$

The hop size is calculated as follows:

$$I = \frac{f_s}{2B} = \frac{f_s}{\frac{4f_s}{N}} = \frac{N}{4} \tag{3.5}$$

The hop size I for an N-point Hamming window is $N/4$, based on a 42-dB criterion on the log magnitude spectrum (bandwidth B is defined as the lowest frequency for which the log magnitude spectrum remains at least 42dB below the peak value). In this STFT framework using Hamming window, a properly sampled STFT requires at least 4 times more

information as it is required for the original signal $s(n)$ (see Equation 3.4). This redundancy provides a very flexible signal representation for which extensive modifications in both the time and frequency dimensions can be made [66]. Equation 3.5 provides the upper limit on the hop size for viable reconstruction of the signal after the STFT is performed.

3.6 Feature extraction : Masking

This section presents a brief background on masking and later presents the feature extraction and effects-processing block of the automix system which measures multitrack masking and minimizes it.

3.6.1 Background theory

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 [68]. Spectral/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 (Figure 3.3a) [69]. Several experiments have been performed in order to estimate the shape of auditory filters in the basilar membrane [13]. Figures 3.3b,c illustrate how adjacent frequency bands of a sound source overlap and mask each other, leading to source masking itself.

In the context of multitrack mixing, a sound source may inevitably 'mask itself', *i.e.* strong frequency content can mask the source's own neighbouring spectral regions as well as 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 [70]. 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 of as the final mulitchannel mix of the multitrack audio, released from masking using various controls (in this context, spectral panning controls).

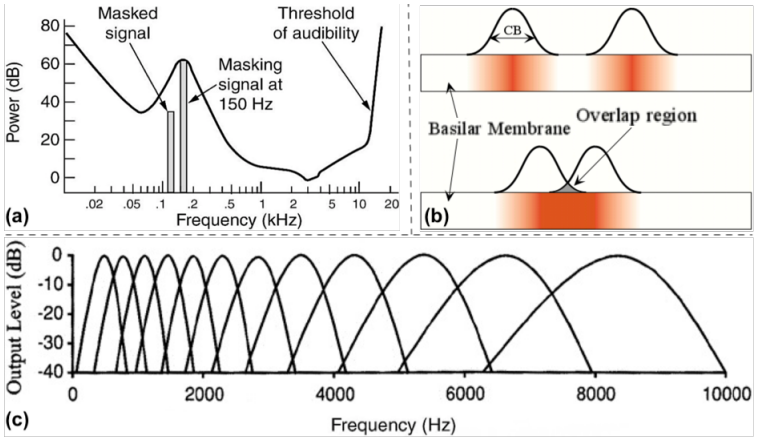

Fig. 3.3 a) Frequency masking [1], b,c) Auditory filters in the basilar membrane

3.6.2 Multitrack masking detection and subgrouping

Over the years, several perceptual models were formulated, most of which were developed for audio coding and compression domains. The underlying principle revolved around approximating the masked threshold of a signal to inform a bit-allocation algorithm or removing perceptually irrelevant time-frequency components [71–74]. However, since the proposed automix system considers real-time operability, it is important to develop computationally efficient algorithms. There has been quality research carried out on using simplified masking models which are lightweight, suited for real-time while still complying to perception to some extent [75–78]. The masking measures used in the proposed automix system is based on spectral similarity, which works well in the listening tests conducted for the approach one of this dissertation [17].

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 (Figure 3.4.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_{track} of the track with respect to the rest of the mix, as follows (Equation 3.6):

$$M^2_{track}(qI,k) = S^2_{track}(qI,k) - S^2_{mix-track}(qI,k) \tag{3.6}$$

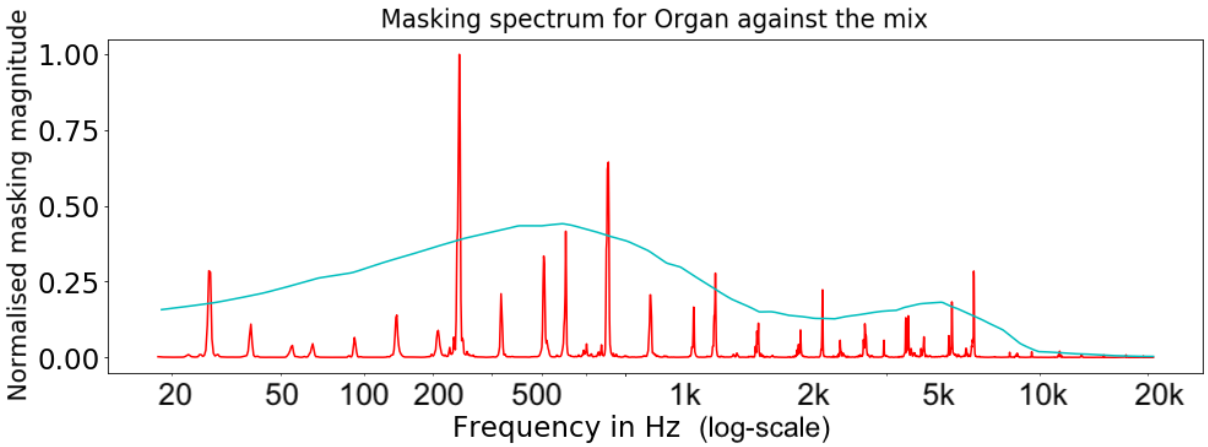

Fig. 3.4 a) Example of M_{track} in a multitrack (red), b) Smoothened average across frequency (blue)

2) To determine tracks that would undergo opposition panning (sources panned to opposite ends of the stereo field), a correlation index matrix (Table 3.1) that measures the similarity index [51] of each track with respect to every other track is computed. Spectral content of two given tracks S_i and S_j are compared using a similarity measure Ψ, computed as follows with forgetting factor $\lambda = 1$ (λ is used to determine the weighting given to past STFT frame):

$$\phi_{ij}(qI,k) = E\left\{S_i(qI,k)S_j^*(qI,k)\right\} \tag{3.7}$$

$$\phi_{ij}(qI,k) = (1-\lambda)\phi_{ij}((q-1)I,k) + \lambda S_i(qI,k)S_j^*(qI,k) \tag{3.8}$$

$$\Psi_{ij}(qI,k) = \phi_{ij}(qI,k)|_{\lambda=1} \tag{3.9}$$

$$\Psi(qI,k) = 2\frac{|\Psi_{12}(qI,k)|}{[\Psi_{11}(qI,k) + \Psi_{22}(qI,k)]} \tag{3.10}$$

Figure 3.5 illustrates a good example of masking based on spectral similarity; several frequency bins overlap for these two sources, eventually leading to masking. Informal listening tests suggested the similar observation, which means that simple FFT based spectral

	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 3.1 Inter-track similarity - Correlation matrix of input multitrack: electric guitar, acoustic guitar, electric piano, organ and saxophone

measures can provide decent masking measures without being too far from perception. Intuitively, the above two metrics M_{track} and Ψ are coherent with perceptual masking; they form a good approximation to 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 spatial locations that are farther apart. If a track has consistently low similarity index with all the tracks, the track is not processed further and remains unpanned in the mix; this phenomenon was observed for kick, bass and vocals most of the time and this is desirable, as we do not want to pan important tracks [11].

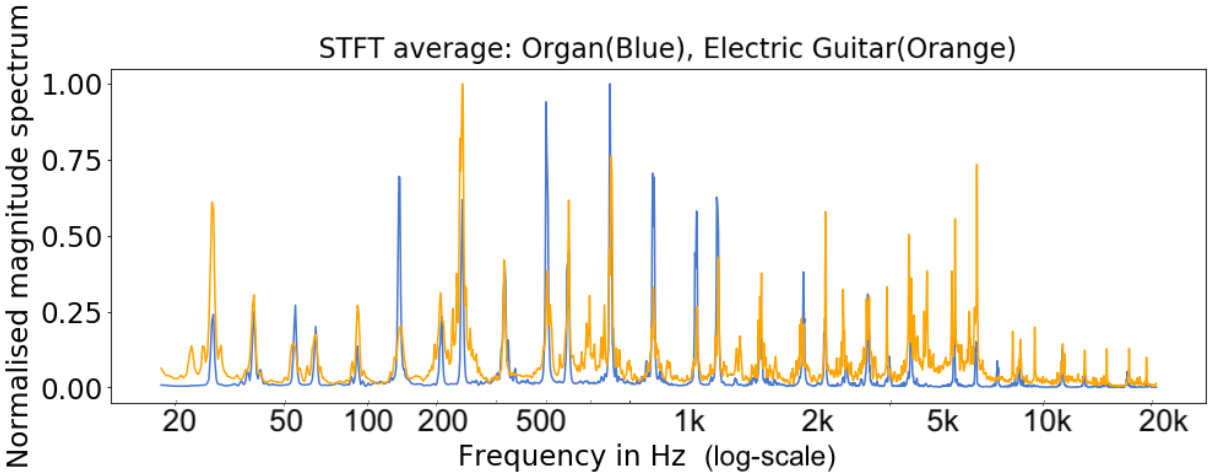

Fig. 3.5 Example of spectrums (averaged across STFT frames) of the most heavily masked sources in a multitrack

3.6.3 Masking metric based on MPEG Psychoacoustic Model

For carrying out objective analysis of the multitrack unmasking improvement on the proposed auto-mixes, we used the properties of the MPEG Psychoacoustic Model, a well-established model used in audio coding/compression algorithms [2]. 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 [1]. Figure 3.6 illustrates the flowchart of the calculation of MSR. This model requires the computation of a multi-resolution Short-Time Fourier Transform (STFT), comprising six parallel FFTs and each spectral frame filtered by a bank of level-dependent Roex filters [79], which is costly. This perceptual masking measure is used only as an objective performance metric to validate the unmasking carried out by the proposed automix algorithms. 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 [9] 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.

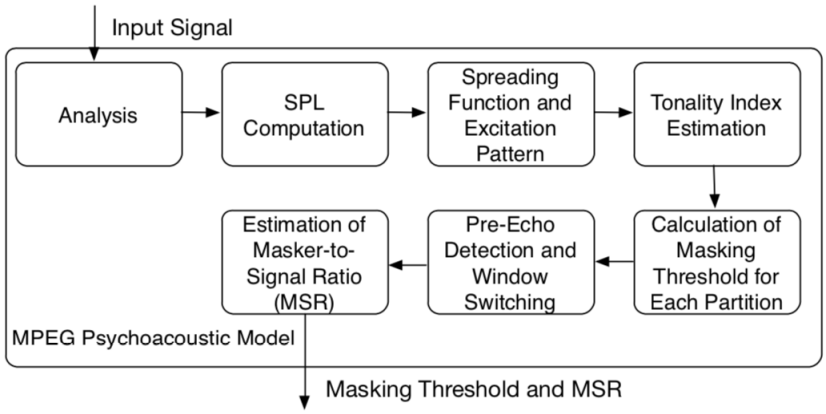

Fig. 3.6 Flowchart of the MPEG psychoacoustic model [2]

To evaluate the amount of multitrack unmasking achieved by the proposed algorithm, we carry out the following calculations to measure masking as per the properties of the MPEG Psychoacoustic Model 1 [2]. We specifically use the cross-adaptive Multitrack Masking

measure, M_n, for track n, as defined in [1]:

$$M_n = \sum_{sb \subset I_M} \frac{MSR_n(sb)}{T_{max}} \tag{3.11}$$

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:

$$MSR_n(sb) = 10log_{10}\frac{T_n(sb)}{E_n(sb)} \tag{3.12}$$

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 $E_n(sb)$.

3.7 Effects-processing: Unmasking using Particle Swarm Optimization

The tasks of a mix engineer mixing a multitrack can be viewed as a constrained, multi-objective optimization problem [53]. There are several control parameters to be adjusted to achieve artistic as well as technical objectives that contribute to a good mix. One of the biggest challenges, as mentioned before, is the fact that all subjective scores and metrics have to be brought down to objective measures for the automix algorithms to work on. In this dissertation, the main goal is unmasking using spatialization. Masking reduction in a mix involves a trial and error adjustment of the relative levels, spatial positioning, frequency and dynamic characteristics of each of the individual audio tracks. In practice, the masking reduction process embodies an iterative search process similar to that of numerical optimization theory [80]. Masking reduction therefore can be thought of as an optimization problem, which provides some insight to the methodology of automatic mixing in order to reduce masking. Given a certain set of controls for a multitrack, the final mix output can be thought of as the optimal solution to a system of equations that describe the masking relationship between the audio tracks in a multitrack recording [1].

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) [20] (this algorithm worked well in the past for automix works that aimed at unmasking [1, 17]) 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 features of the panning filters, which are objectively tuned to reduce the cost function M_m (multitrack masking) which is defined as the L_2 norm of $M_{track}(qI, k)$ in the spectral bands that need to be unmasked.

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 [1,81] as part of the global optimization process:

$$x_{min} = argmin_x(M_m(x)) + \underset{x,i,j,i\neq j}{\text{argmax}}(M_d(x, i, j)) \tag{3.13}$$

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

In the next two chapters, the generation of the spectral panning filters that carry out spatialization and unmasking are presented, along with specific details on the PSO constraints and evaluation metrics.

Chapter 4

Automatic Spatialization relying on Best Panning Practices

In this chapter, the techniques and concepts used to carry out automatic spatialization and masking minimization using the first approach is discussed. This spatialization effect is carried out using equivalent rectangular bandwidth (ERB)-scale sinusoidal shaped panning filters which are designed and further optimized based on rules / constraints from best panning practices (KE-approach to automatic mixing). The proposed panning system uses the masking metric and subgrouping discussed in the previous chapter to carry out frequency-based panning. Both real-time and off-line optimization approaches are implemented, both of which make use of the same framework for the feature extraction and effects-processing. We focus on the headphone listening context, though the work may also be applicable to loudspeaker playback. The results in this chapter are published as a 10-page peer-reviewed paper which was presented at the 146th AES Convention, Dublin, Ireland [17].

4.1 Panning practices, rules and constraints

Developing automatic mixing systems requires drawing inspiration from audio production methods. This spatialization approach follows a knowledge engineering approach (section 3.2.3). Though frequency panning is not a traditional mixing practice, the proposed spatialization approach has the same objective principles as the more common panpot-based panning. 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 [25]. Since this work deals mainly with optimization based on panning constraints, we seek to embed the common practices used for placing sound sources.

4.1.1 Panning - 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 [82]. High priority tracks such as vocals are usually kept centre-panned [83]. 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 [3].

4.1.2 Low frequency sources - 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 [3]. Interaural Level Difference (ILD - discussed in Section 2.4.1) 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 [83].

4.1.3 Mid frequency area - minimize spectral masking

Separation and definition of each track in the low-mid frequency region is critical to achieve a clear and well produced mix [3]; 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 [83]. 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.

4.1.4 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 [3].

4.1.5 Overall stereo picture - maintaining the 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 [3], 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.

4.2 Framework and implementation

The framework of the proposed algorithm is as follows: monophonic audio tracks are fed to an STFT framework (spectral decomposition), spectral similarity of each track with respect to every other track is computed (stored as a correlation matrix, Table 3.1 to determine track pairs that would undergo opposition panning [83]). Spectral masking of each track with respect to the mix is computed (Equation 3.6). The track-pairs that have highest spectral similarity are most vulnerable to spectral masking. Hence, 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 (frequencies and phase offsets of the panning filters) of the respective tracks are determined by a particle swarm optimizer [20] that minimizes multitrack masking (cost function) while complying to the

panning rules (constraints) discussed in previous section. Figure 4.1 illustrates the block diagram of this framework.

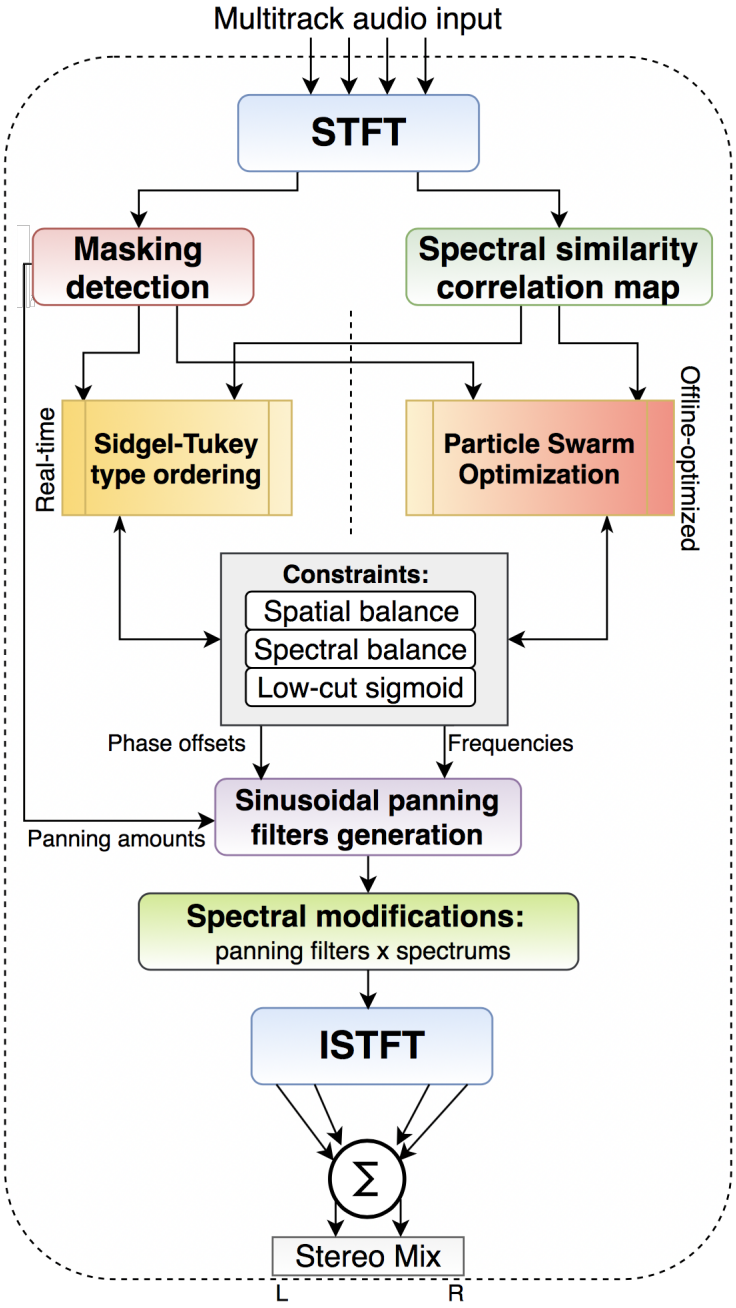

Fig. 4.1 Block diagram of automatic spatialization based on panning practices - approach one [17]

4.3 Opposition panning using panning filters

4.3.1 Motivation for spectral panning

In our perception, we want each instrument to have a defined position and size on the frequency spectrum. 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 the 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 [73,84]. Instead of removing masked frequency bins, the proposed algorithm places alternate frequency bands of the sources across the stereo field based on masking.

Mix engineers use the panpot [46] to spatially position sources in the mix. A limitation of balancing using a panpot is that the sources remain tied to one side and the width may be restricted [4]. This issue is addressed in the algorithm which generates a mix with a stable centre image while still giving a spatialized effect in a wider stereo field. The panning filters designed in our work draw inspiration from the work by Pestana et al. [12] in that no track is panned as a whole to either side, rather, time-frequency bins of each track are panned; however [12] did not account for masking reduction as an objective function, neither did its panning effect comply to auditory perception. The proposed algorithm objectively minimizes masking using optimization; various spatialization and masking metrics extended for multitrack are used to validate the performance.

4.3.2 ERB-based sinusoidal panning filter

Multi-resolution STFT decomposition and reconstruction is computationally costly and not suitable for real-time effects, as discussed in Section 3.6.3. 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 that accounts for decreasing frequency resolution of human hearing with increasing frequen-

*The equivalent rectangular bandwidth or ERB is a measure used in psychoacoustics, which uses an approximation of the bandwidths of the filters in human hearing, relying on the unrealistic but convenient simplification of modelling filters as rectangular band-pass filters.

cies [23, 85]. The panning filter is converted from the ERB domain to linear frequency domain. Figure 4.2 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 3.6.1, we also discussed about a source masking some of its own components. Several experiments [13] have concluded that the auditory filters take the form of rounded complex exponential function like in Figure 3.3(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.

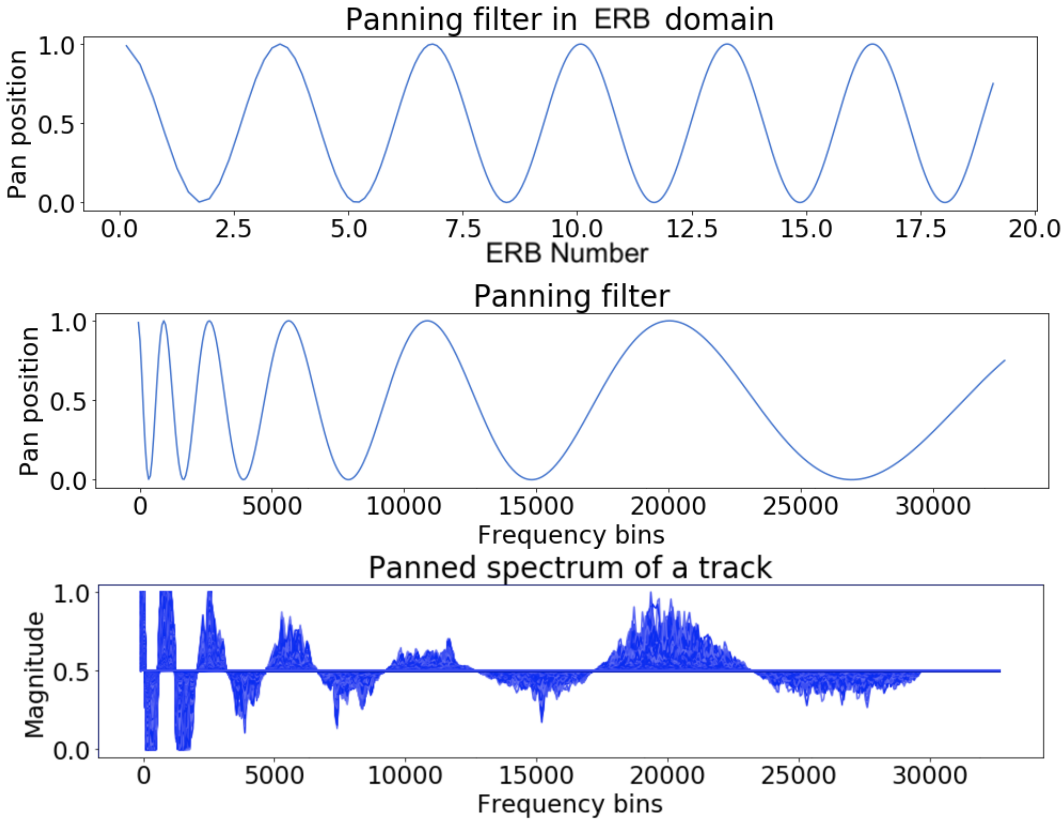

Fig. 4.2 a) Panning filter in ERB domain (top) b) Panning filter in linear frequency domain (middle) c) Magnitude spectrum of a track's STFT frame after panning (below the x-axis is pan amount to the left, above the x-axis is pan amount to the right) (bottom)

4.3.3 Panning filter design

The panning filter for each track is defined as follows:

$$P_j(qI, k) = \rho_j(qI, k) \cdot sin(21.4 \cdot \upsilon_j \cdot log(1 + 0.00437(b(k))) + \delta_j) \tag{4.1}$$

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 4.4, the role of spectral envelope filter ρ_j is as follows.

4.3.4 Spectral envelope filter

To comply to the panning rules discussed in section 4.1, each panning filter is multiplied by a spectral envelope filter. The spectral envelope filter is computed as follows: the M_{track} function (Figure 3.3b) 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 (Figure 4.3) with 3 spectral regions whose roles are defined as in the figure.

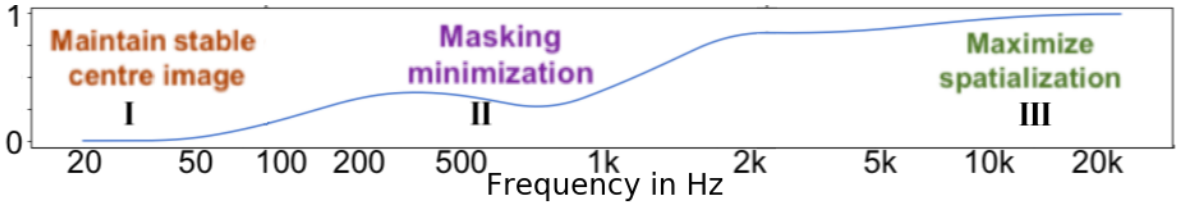

Fig. 4.3 Example of a spectral envelope filter ρ_j

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 4.1.2). This feature becomes relevant especially in the headphone listening context in which spatial attributes/ILDs of low frequency content can be perceived [86].

Region II (500 Hz - 2 kHz) undergoes maximum masking in a multitrack [3], 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 (Figure 3.4(b)). This region ensures **spectral masking minimization** (section 4.1.3).

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 4.1.4). This region contributes to **spatialization** of the final stereo mix.

4.4 Multitrack masking minimization

In this section the implementation of both real-time and offline approaches to automatic spatialization under masking minimization are discussed. The difference between the two approaches lie in how the phase offsets δ_j and frequencies υ_j in Equation 4.1 are determined; these parameters are responsible in placing masked spectral content in different spatial locations across the stereo field.

4.4.1 Real-time approach

In the real-time approach, we use a palindromic Siegel-Tukey type ordering [87] 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 given two groups of data tends to have more widely dispersed values than the other. In the example of Table 3.1 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, and 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. This panning method gives no perceivable effect; the tracks sound monaural.

4.4.2 Offline optimized approach

The PSO, as discussed in section 3.7, minimizes multitrack masking 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 $M_{track}(qI, k)$ in Region II.

The PSO is constrained by the panning rules described in section 4.1.5 and by bounds. It is important to maintain left-right balance across the entire frequency spectrum (spectral balance) and energy ratio of both channels (spatial balance) [11]:

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

$$Spec_{Band_i} = tan^{-1}(\sum_{k=B_i}^{B_{i+1}-1} |S_L[k]|^2 / \sum_{k=B_i}^{B_{i+1}-1} |S_R[k]|^2) \tag{4.2}$$

where S_L and S_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 centred 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 4.1.5. 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 4.4.1. 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 (Figure 4.4). 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.

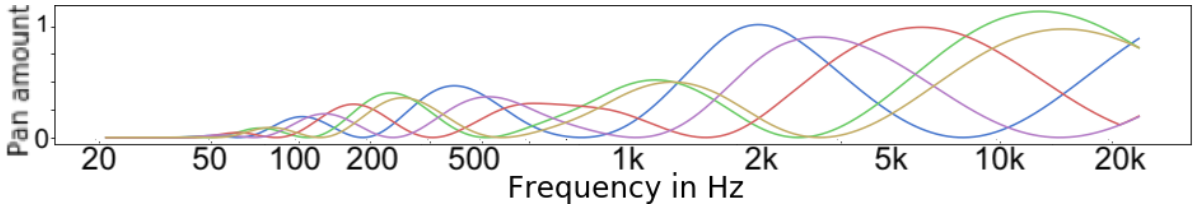

Fig. 4.4 Optimized panning filters (each color represents panning filter for respective track pair - Table 3.1)

4.5 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 [10–12] 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 [56].

4.5.1 Objective evaluation : unmasking and spatialization

Figure 4.5 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.

$\Delta Mask$		Folk	Country	Jazz	Funk	Pop	Rock
Real-time Mix	*[1]*	12.2	9.2	10	14.8	11.5	8.2
	ΔM_m	130	70	82	132	76	40
PSO Mix	*[1]*	19.5	14.6	15.1	26.7	17	12.6
	ΔM_m	142	75	91	159	80	55

Table 4.1 Change in masking (unmasking amount) : MPEG Multitrack Masking [1] and multitrack masking M_m

In Table 4.1 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 [88]. The perceptual masking metric (MPEG Psychoacoustic Model 1 [2]) in [1] 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.

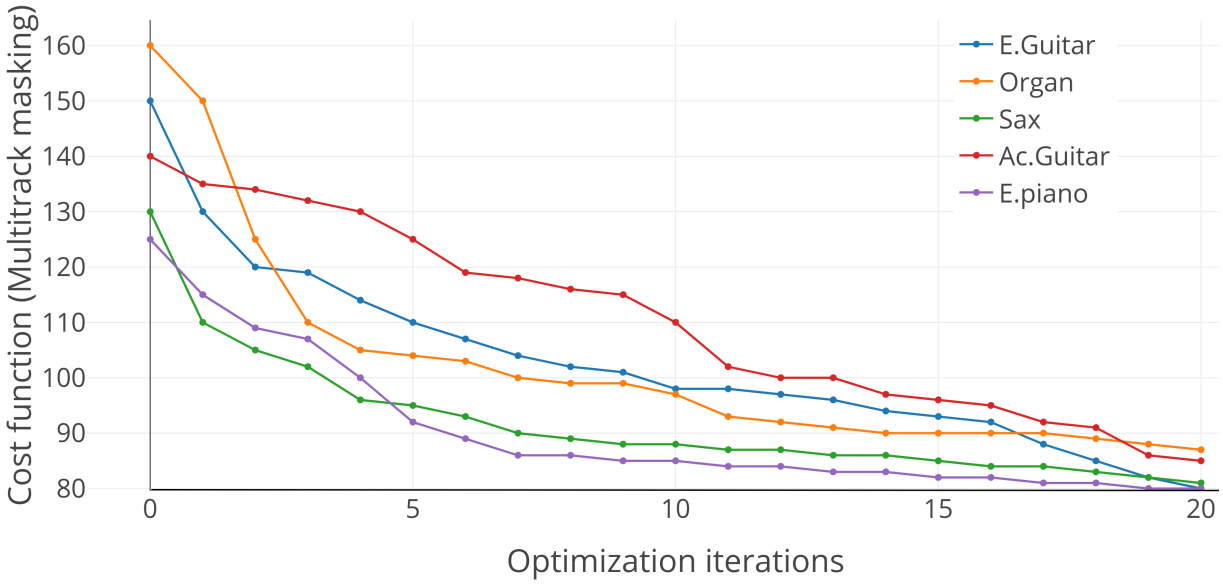

Fig. 4.5 PSO cost over iterations

The amount of spatialization was analyzed using the stereo panning spectrum (SPS) [51] and the panning RMS [16] 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 3.10) 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 Figure 4.6 for one STFT frame.

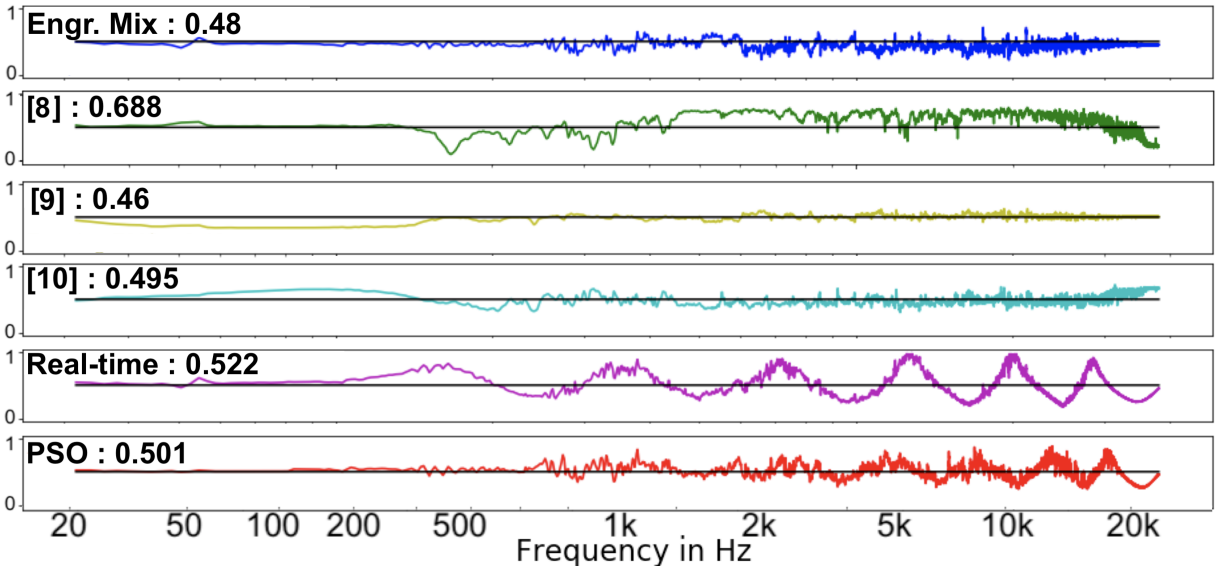

Fig. 4.6 Panning RMS value mentioned next to each mix, each color refers to stereo panning spectrum. Y-axis: Panning amount

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 Figure 4.7: spectral panning bandwidth (Equation 4.2) is calculated across audible frequency bands [10] for over 100 songs from the Open Multitrack dataset [88]. 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 3.1 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: [11] and [12] have unstable spectral balance in the low frequency end which results in an unstable stereo image, [10] has poor spatial balance since the energy is concentrated towards stereo right.

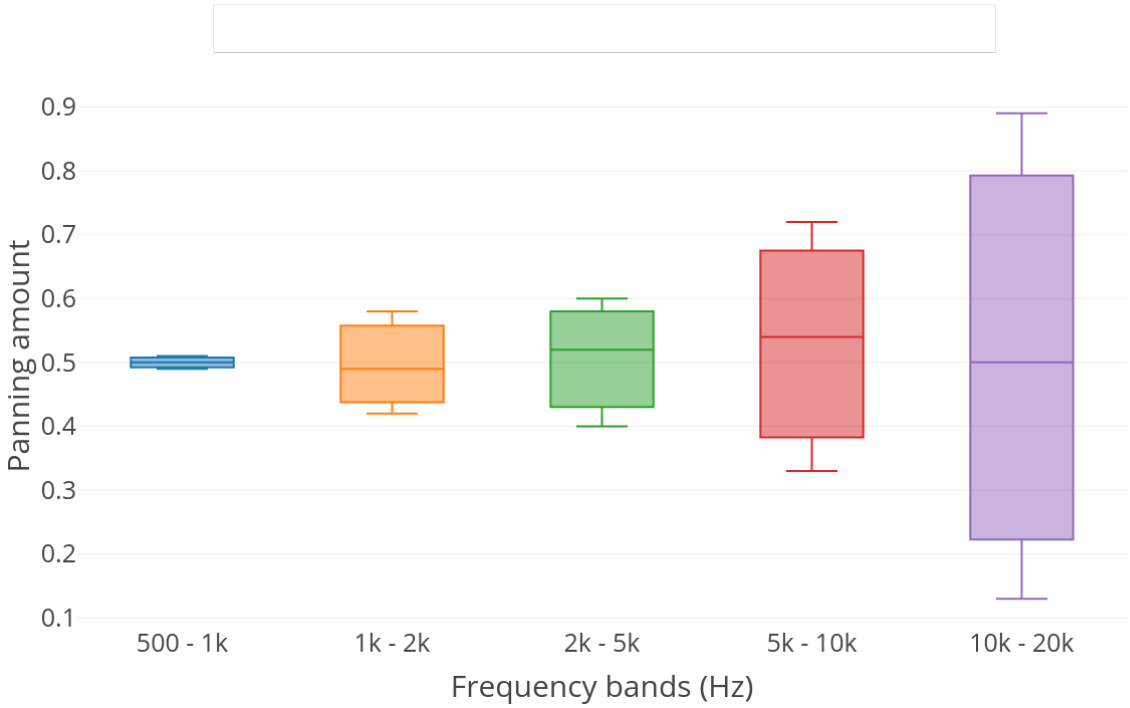

Fig. 4.7 PSO mix: Relative panning bandwidth across frequency bands, for 100 songs

To analyze the amount of spatialization and stereo activity, the mix outputs were run through a goniometer [10]. 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 (Figure 4.8) we observe the following: a) Human mix has stable centre image but has relatively narrow panning, b) [10] and c) [11] have lopsided spread thus having poor spatial balance, d) [12] and e) Proposed real-time mix has reasonable spatialization, f) PSO mix performs extremely consistent with stable centre image as well as maximum spatialization.

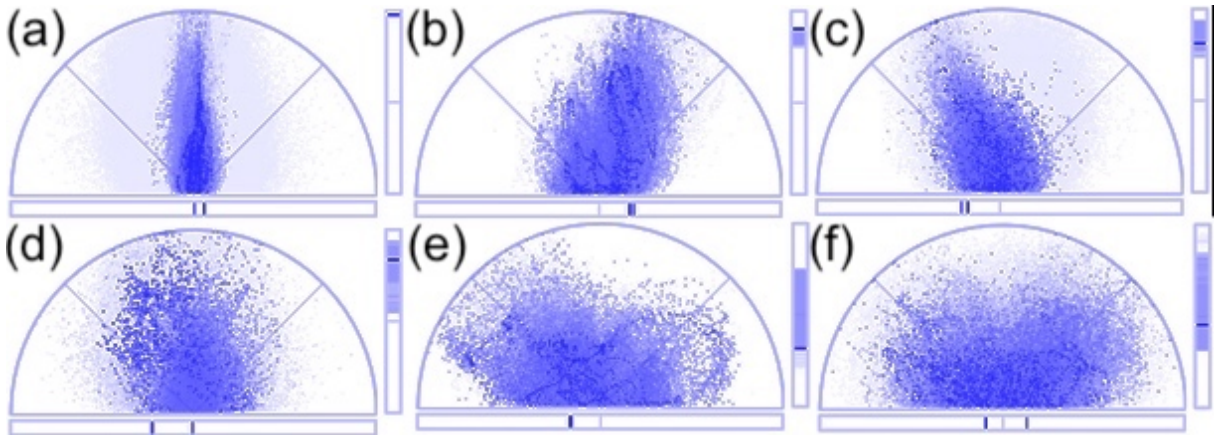

Fig. 4.8 Goniometer output: a) Sound engineer mix, b) [10], c) [11] d) [12] , e) Real-time mix, f) PSO mix

4.5.2 Subjective evaluation : listening test

All listening tests were conducted at the Centre for Digital Music, Queen Mary University, London, UK. 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 [89] 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 (Figure 4.10) 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), "wierd 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 subjective APE test (Figure 4.9) is available online (*headphones recommended*): http://webaudio.gutech.edu.om/test.html?url=tests/pantest1.xml All the audio samples used in the test are available at the end of the following webpage: https://www.ajintom.com/auto-spatial

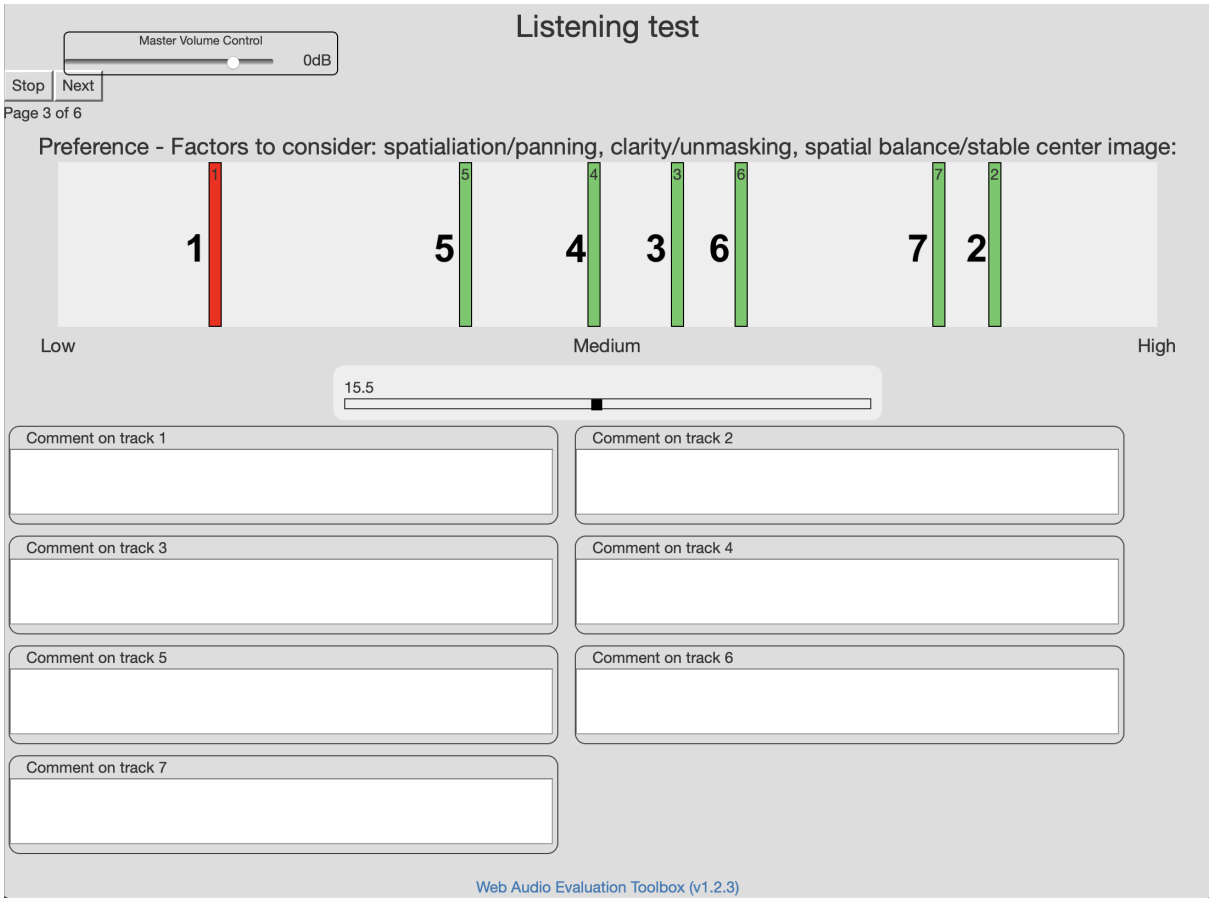

Fig. 4.9 APE listening test interface (page 3 of 6). *Track numbers across the green markers - 1: Monosum, 2: Sound Engineer Mix, 3: [10], 4: [11], 5: [12], 6: Real-time mix, 7: PSO mix. There were six such pages each for the following genres (across page numbers) - 1: Folk, 2: Country, 3: Jazz, 4: Funk, 5: Pop, 6: Rockballad.* On each page the subjects were required to click on the green markers to listen to the respective audio samples (the marker turns red on clicking) and drag them across the preference scale from Low to High and later comment on each of the 7 audio tracks. The audio samples and genres were randomized when the listening test was conducted

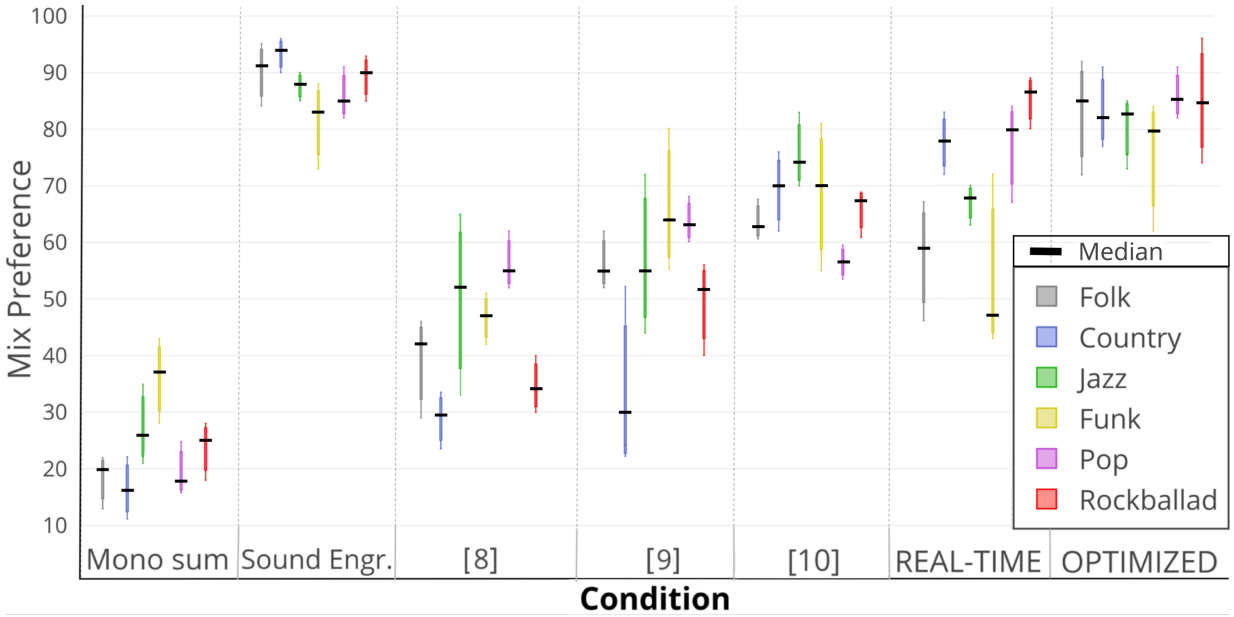

Fig. 4.10 Listening test results : mix quality rating for multitrack monosum, professional sound engineer mix, existing auto-mix works [10–12], proposed real-time mix, PSO mix; represented as a box plot with first value, median, last value

4.6 Discussion and conclusion

This chapter describes a frequency-based multitrack panning automation algorithm based on a knowledge engineering approach. 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 concept of frequency panning might intuitively sound odd, considering that a sound image could appear blurred throughout the stereo field with frequency components arriving from several directions. However, it is important to note that in the proposed algorithm, the

lobes of the panning filter are tuned such that, the majority of the mid-frequency content will take up one direction in the stereo field, the low-frequency content remains unpanned. Only the high-frequency bands are widely panned alternately across the stereo field, and from the listening test we noted that, the heavy panning of high frequency content did not affect our perception of spatial location of the source; the main timbre characteristics and contributions of a source usually appears from the mid-frequency bands, which in this case are tied to one side in the stereo field.

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. The 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.

Chapter 5

Automatic Spatialization relying on Source Directivity

This chapter presents the main idea and the implementation for the proposed spatialization system relying on source directivity. In the earlier section chapter (approach one), we used hand-crafted features and rules based on best panning practices to carry out frequency-based panning that relies on release from spectral masking using optimization. We now introduce a different spatialization approach, for which we use the same optimization framework but with source directivity as a constraint. This chapter will address how spatialization systems could place sound sources in other formats beyond stereo, such as ambisonics or 5.1 surround for example.

5.1 Motivation for 3D spatialization and directivity patterns

With the advent of VR/AR technologies, audio content creators, hosts, consumer electronics manufacturers and broadcasters require technological frameworks to capture and render true-to-life immersive 3D audio experiences. Latest formats like the MPEG-H which support scene-based audio technology are designed to overcome key limitations of traditional audio formats [90]. In the film industry, ambisonics microphones are used to create a 3D sound image. It is also possible to synthesize this image using amplitude panning techniques based on the scene. However, in the context of music listening, we are not that used to listening to 3D mixes. The ones that are produced are again recorded using special microphones or complex sound recording arrangements, which also require spatial audio

set-ups like 5.1, 22.2 or reliable binauralization technologies for reproduction/mixing. This calls for new techniques to determine how such works can be mixed; the main challenge being the sound mapping between sources and output tracks. In this dissertation we use directivity patterns of sources to determine this sound mapping. This idea dates back to the age-old art of symphony orchestra performances, specifically the context of arranging the instruments on stage with specific orientations to manage frequency overlap and separation.

The spatial signature of a source is characterized by its directivity / directionality. Directivity is a measure of the directional characteristic of a sound source, which determines how the energy is spread around the source [19] across the angles for each frequency. This energy spread determines the frequency response from the source to listening point (in the sound field) which is at a certain angle and distance with respect to the source (discussed in detail in later sections). The use of directivity patterns will replace ad-hoc rules to achieve multitrack mixing by automatically spreading the energy of the source over the considered set of tracks. The evaluation criteria involve objectively calculating spectral unmasking across output tracks as well as the fruitful extent of spatialization using a goniometer and systematic analysis of directivity rendered around sources. The aim is to deliver sound for an immersive environment, where sources can appear at any position with specific directivity patterns that quantify their direction-dependent behaviour in the two or three-dimensional space around the listener. The spatial properties of sound fields are important for human sound source localization in daily life and greatly affect the perceived sound quality and intelligibility, which has been explored in audio applications since the early days of the two-channel stereophonic reproduction, remaining an active field of research and development [91]. In an aesthetical context, spatial control of sound has been widely used in contemporary electroacoustic music too [92].

In the following sections we discuss about the directional characteristics of sound sources as well as the design and computation of the sound field around these sound sources which will eventually be given directivity features. The underlying idea is to not only render acoustical / synthesized directivity but to play/interact with directivity parameters in a mixing system.

5.2 Sound field around a source

A sound source can be characterized by the temporal and spatial properties of the radiation/sound field that it produces, under free-field conditions. Free-field condition refers to the absence of any reflected waves due to room or obstruction effects. In other words, there is a mechano-acoustic coupling between the radiating body and the sound field it produces. Strictly speaking, this coupling depends on the characteristics of the room the source is placed in, so that the room affects the acoustical properties of the sound source. However, since most of the sound sources placed in ordinary rooms present a mechanical impedance much higher than the sound field, the room has just a minor effect on the source dynamics, so that it can be neglected [93]. The spectral and temporal structures of the audio signals reaching the eardrums play a major role in the human perception, which is categorized in musical and psychoacoustical attributes such as pitch, duration, dynamics, timbre, loudness and localization [14, 23]. These attributes/features of a sound source can be captured and analyzed by techniques based on Fourier analysis [66].

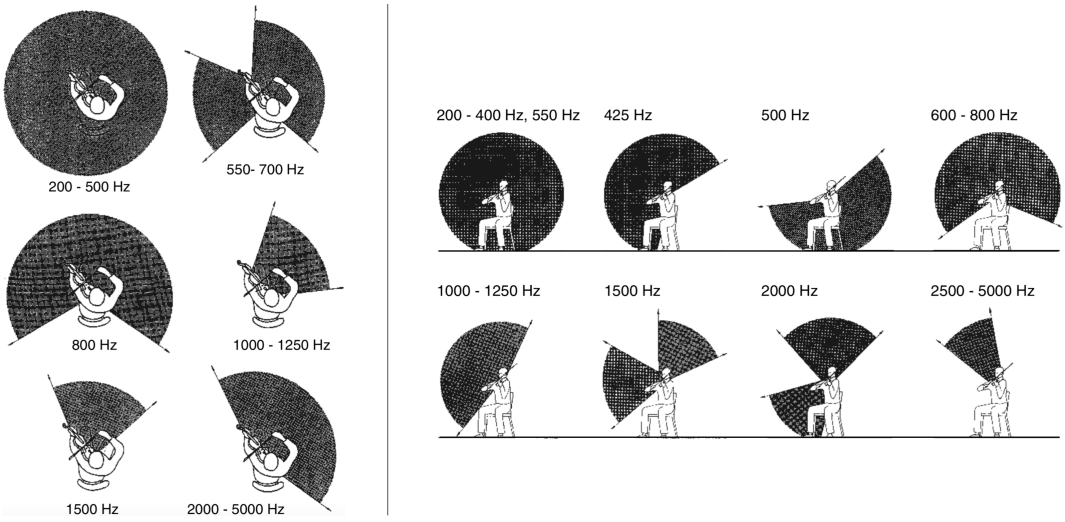

Fig. 5.1 Principal radiation directions of a violin in its horizontal (left) and frontal (right) plane (*Acoustics and the Performance of Music: Manual for Acousticians, Audio Engineers, Musicians, Architects and Musical Instrument Makers by Meyer* [91]). The shaded areas represent the directions in which the sound pressure level is within 3 dB of its maximum value averaged over a given frequency range in the respective planes

Musical instruments radiate sound power in all directions. The sound field around these sources is characterized in terms of sound pressure whose attributes are directivity, acoustic intensity, and sound power. Our focus is the dependence of the pressure amplitude with respect to distance, orientation and frequency. We limit ourselves to the physical description of the radiation of instruments. We do not consider room acoustics and obstruction due to physical objects in the room/space. Figure 5.1 depicts an example of the directivity pattern of a violin in the horizontal and frontal plane according to [91]. The directivity of acoustic instruments producing sounds in their lower registers / notes is usually omnidirectional; as they produce higher frequencies the directivity starts becoming more prominent / complex / varying across angles. For a given azimuth θ, elevation ϕ and distance r from a source, a listener will experience a particular frequency response $H(r,\theta,\phi,\omega)$.

The directivity pattern of a sound source is relevant for human perception. This feature of a source can be objectively measured by rotating a given sound source around its axis. Directivity is obtained through measurements in an anechoic chamber. However, the experimental procedures are complex, time-consuming and require expensive facilities which are not readily available in many acoustic laboratories. The most comprehensive and referenced works in the context of presenting overall directional characteristics of several musical instruments are the works by Meyer [91] and Blauert [14]. Another approach is to develop analytical methods to model directivity as features of sources and to use directivity more as an audio effect than an acoustical feature. Our spatialization system develops directivity modelling using the following techniques: template directivity drawn from acoustics / geometry of the source (and associated radiation) as well as directivity designed by the user for producing specific spatial effects using template radiation patterns.

5.2.1 Building directivity using elementary sources

One of the key considerations in musical instrument design is sound power radiation: this is a necessary condition for listeners to perceive sound. The sound that reaches the listener (or the microphones) mainly depends on the properties of the sources (the musical instruments). Musical instruments are complex systems in which numerous acoustical and vibratory phenomena are intrinsically combined together. Before diving into developing complex directivity patterns, we introduce the basic concepts of radiation applicable to musical instruments.

Elementary sources are limiting cases that are often very useful, as a first approximation, in order to describe complex sources such as musical instruments. Building up directivity using elementary sources follows from the principle of superposition which is valid in linear acoustics; any extended source can be represented as a distribution of elementary sources. This result forms the basis of the calculation of the acoustic field radiated by a musical instrument with arbitrary geometry as illustrated in [94]. Studying directivity using this physical approach also serves as a guide for the selection of construction parameters relating to geometry. In this context, we limit ourselves to the physical description of the radiation of the instruments, leaving aside the questions linked to the acoustics of the room and on the psychoacoustical aspects of the sound perceived by the listeners.

A given sound source can be viewed as composed of of infinitely many elementary point sources. Each of these point sources radiate in all directions. However, putting together these point sources in a certain manner to physically make up the geometry of an instrument will provide an overall directivity to this sound source that occupies volume in a room or space. This directivity is formed due to the constructive and destructive interferences of the radiating point sources. From a physics perspective, these individual point sources can be viewed as 'pulsating spheres' to illustrate the concept [94]. The pulsating sphere is a typical example of perfectly omnidirectional source. In this case, the amplitude of the sound pressure only depends on the distance from the source, and not on the angle of observation, as expected due to symmetry principle in physics. A monopole, or point source, can be viewed as the theoretical limit of a pulsating sphere when its radius tends to zero. It is an idealized system which is not feasible in practice. However, one can build reasonable approximate monopole sources under the condition that their dimensions are kept small compared to the wavelength and/or the distance to the observer (listener).

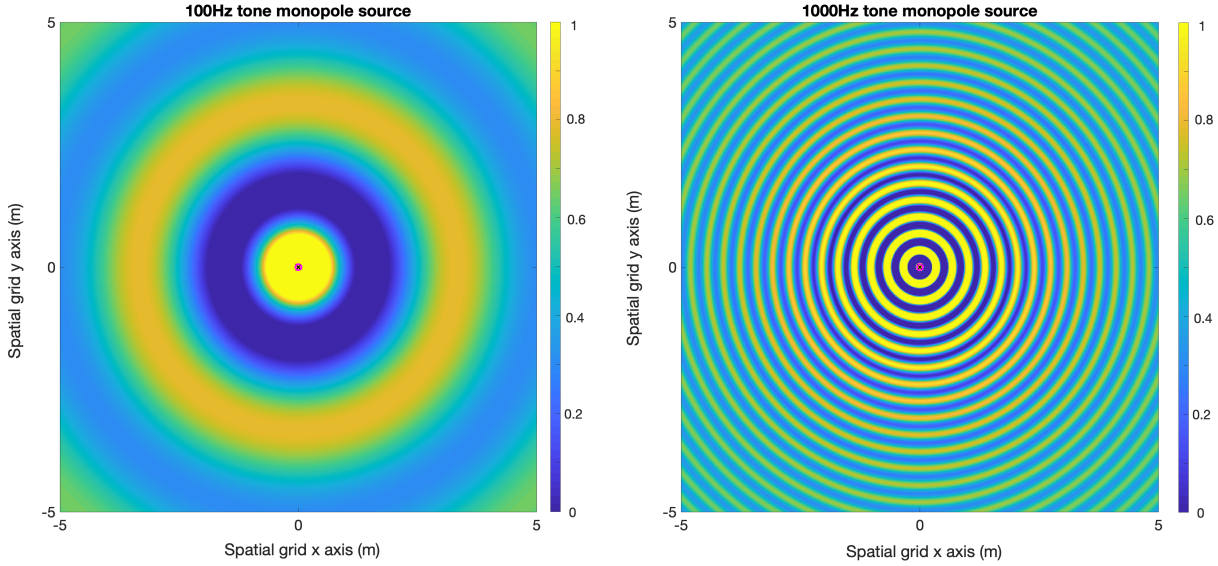

Fig. 5.2 Sound field around a monopole source

Using the wave equation expressed in spherical coordinates, we get the expression of the radiated pressure field at a distance r from the centre of the sphere. Figure 5.2 illustrates the sound field produced by a monopole source generating a tone. From this general solution, with speed of sound propagation c, pressure p at a distance r from a monopole producing angular frequency $\omega = 2\pi f$ can be expressed as factors of time and position dependence:

$$p(r,t) = \frac{1}{r}T(t)P(r) \tag{5.1}$$

$$= \frac{1}{r}e^{jwt}e^{-j\frac{w}{c}r} \tag{5.2}$$

$$= \frac{1}{r}e^{jw(t-\frac{r}{c})} \tag{5.3}$$

Note that for all examples hereafter we consider the global origin of the sound field / sound stage to be the centre of the 2D spatial grid. While representing pressure or frequency response at a given point in the sound field, the distance r considered is the one between the centroid of the source and the listening point (local origin, which might vary over time). In the case of monopole arrays, we consider the distances between each elementary source and the listener, $\underline{r} = (r_1, r_2, \ldots, r_{M_e})$ to compute the pressure p. Each

source's orientation is considered with respect to the horizontal axis and the listener is at a distance r, azimuth θ and elevation ϕ with respect to the source.

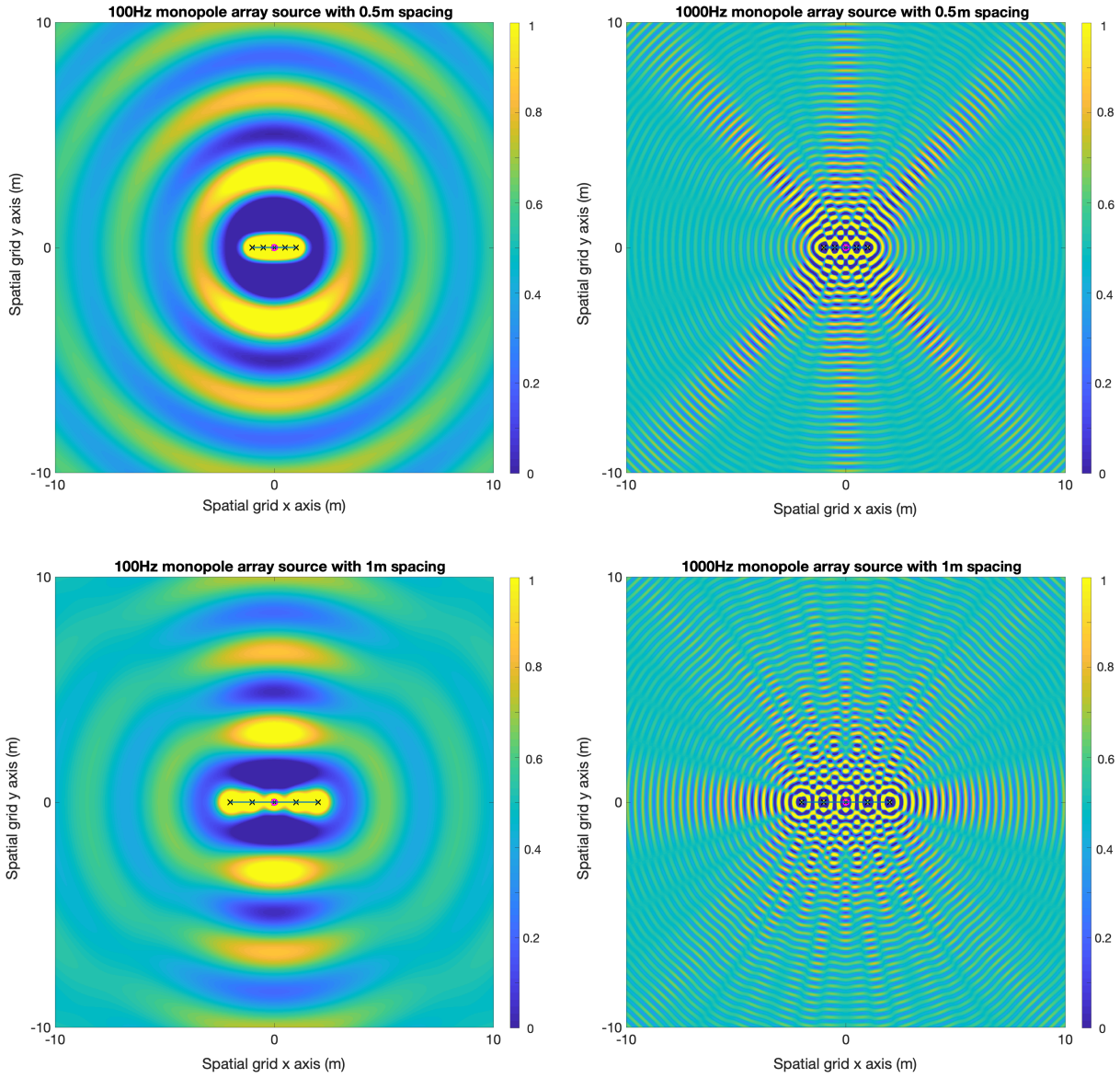

Fig. 5.3 Sound field around a monopole array of sources (marked in black, centre of the source marked in red) arranged linearly with certain spacing producing a tone

This result can be extended to build a sound source using a discrete set of monopoles distributed in space (Figure 5.3). One can derive the resulting sound field / pressure

at a given point (position of observer/listener) by summing the contributions of the M_e elementary sources from a vector of distances $\underline{r}$ (Equation 5.4):

$$p(\underline{r},t) = e^{jwt}\left[\frac{1}{r_1}e^{-jw\frac{r_1}{c}} + \frac{1}{r_2}e^{-jw\frac{r_2}{c}} + \frac{1}{r_3}e^{-jw\frac{r_3}{c}} + ... + \frac{1}{r_m}e^{-jw\frac{r_{M_e}}{c}}\right] \tag{5.4}$$

$$= e^{jwt}\sum_{m=1}^{M_e}\frac{1}{r_m}e^{-jw\frac{r_m}{c}} \tag{5.5}$$

One main consequence of monopoles aligned as linear array of sources is that the directivity increases compared to the case of a single monopole. Musical instruments do not escape this rule. For example, a slender structure like a marimba/xylophone bar vibrating on a high mode can be approximated as a linear array of elementary sources. In the above case each elementary source is assumed to produce a single frequency tone of same amplitude, located at a certain distance from the observer thus causing a phase shift. The pressure varies across the sound field not only due to distance attenuation, but also due to phase interference between the sound fields radiated by each elementary source; directivity pattern is created by this phase interference in the sound field by each elementary source. In order to determine the pressure resulting from the association of elementary sources generating non-monochromatic wave, we integrate the pressure across frequency (Equation 5.6). Figure 5.4 illustrates the sound field around a linear array of elementary sources producing a complex tone.

$$p(\underline{r},t) = \frac{1}{2\pi}\int_{\omega} p(\underline{r},\omega)e^{j\omega t}d\omega \tag{5.6}$$

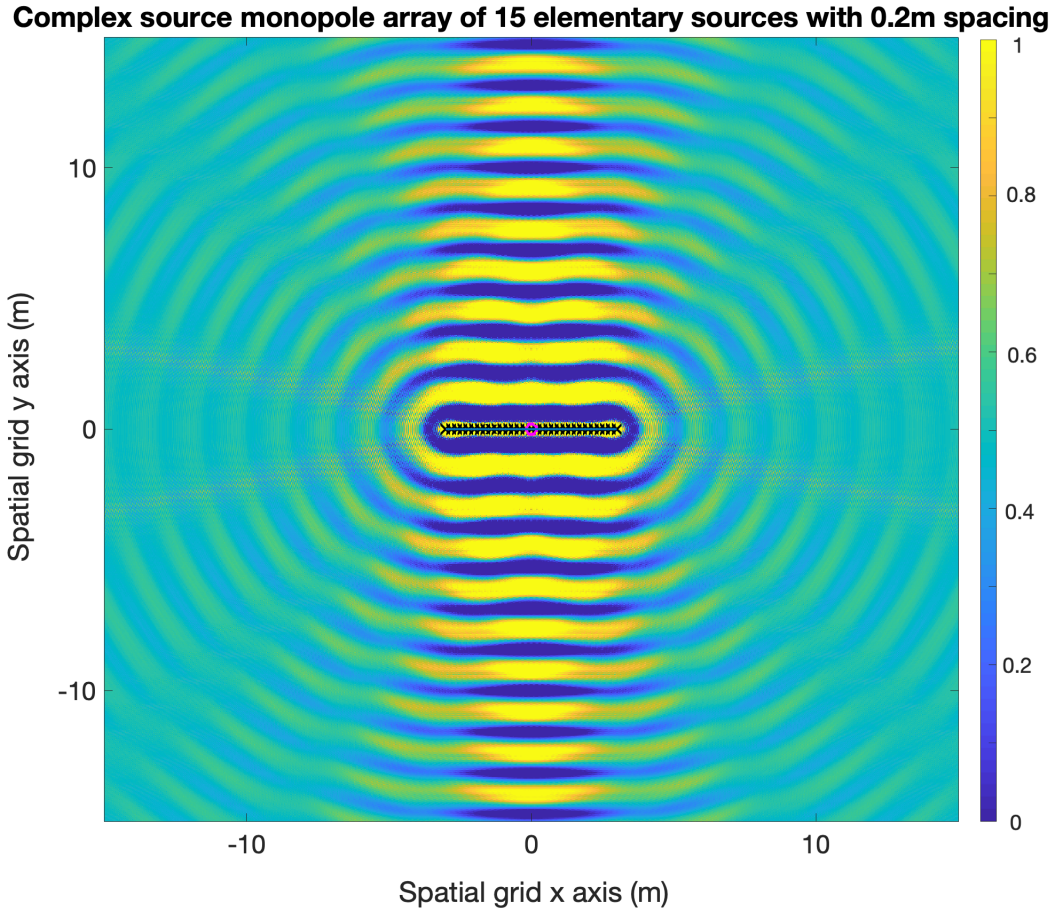

Fig. 5.4 Sound field around a linear array of elementary sources playing a marimba tone (note B3 : 247 Hz)

As discussed in Chapter 3 (Figure 3.2), the proposed spatialization algorithm is designed as a time-varying linear system (based on an STFT framework) to which each input signal is fed in and we get the output signal after a processing step (frequency response H applied to each STFT frame of the input signal). In the context of the discussion in this section, we have input pressure $p_i(t) = e^{j\omega t}$ which is processed by a propagation system (made of the path between the sources and the listening position) to generate the resultant output pressure $p_o(\underline{r},\, t) = e^{j\omega t}\ H(\underline{r},\omega)$ at the listener position. The frequency response H is the ratio of p_o and p_i.

For a monochromatic wave of frequency ω we consider the following formulation for

pressure at the listening point by elementary source s_m ($m \in [1, M_e]$) located at (r_m,θ_m, ϕ_m) from the listener (origin):

$$p_m(r_m, \theta_m, \phi_m, \omega) = e^{j\omega t} \cdot F(r_m, \theta_m, \phi_m, \omega) \tag{5.7}$$

where F is a generalization of P as in Equation 5.1. Considering the more general case of contribution from all M_e elementary sources, we have the resultant pressure at the listener position as follows:

$$p_o(\underline{r}, \underline{\theta}, \underline{\phi}, \omega) = \sum_{m=1}^{M_e} p_m(r_m, \theta_m, \phi_m, \omega) = p_i(t) \cdot \sum_{m=1}^{M_e} F(r_m, \theta_m, \phi_m, \omega) \tag{5.8}$$

Therefore, the frequency response of the propagation system providing the directivity across space and frequency at a given time instant (STFT frame) is as follows:

$$H(\underline{r}, \underline{\theta}, \underline{\phi}, \omega) = \frac{p_o}{p_i} = \sum_{m=1}^{M_e} F(r_m, \theta_m, \phi_m, \omega) \tag{5.9}$$

The propagation system, characterized by F($\underline{r}$, $\underline{\theta}$, $\underline{\phi}$, ω) accounts for the resultant pressure p_o (at the listening point) due to phase interference caused by pressures p_m generated by each elementary source s_m ($m \in [1, M_e]$).

5.2.2 Template directivity using user-defined radiation

The control strategy adopted in the context of this dissertation is to provide sources with pre-programmed basic directivites using template directivity equations. For the sake of simplicity, the method will be discussed using a two-dimensional model; this can be readily extended to three dimensions. The radiation pattern defines the variation of magnitude radiated by a source as a function of angle θ and frequency ω. Since we consider only the far-field scenario [95], the distance attenuation is governed by inverse distance law (though distance attenuation is considered in the final implementation, in this section we focus on just the frequency response caused by directivity due to orientation of the source with respect to the listener). The directivity function used in this dissertation is adapted from antenna directivity theory [96] (specifically concepts relating to the formulation of *array*

factor in chapter 6) to study this variation*. The total field of an M_e-element array (forming a secondary source) is equal to the field of a single element positioned at the origin (centroid of the secondary source) multiplied by a factor which is widely referred to as an *array factor*. This factor is characterized by a dirichlet kernel [97], or periodic sinc function) which is a good approximation of directivity since it has a low pass structure, thus complying with natural radiation patterns observed in acoustic instruments. We designed Equation 5.10 empirically studying and modifying the array factor formula to have desirable limits and flexibility required for our use-case (building template directivity for a musical source). We discuss the discretized formulation hereafter with sampling in time domain with $f_s = 44100$ Hz, frequency domain with STFT parameters ($N = 2^{15}$, $I_{max} = N/4$, $k = 0, 1, 2, 3, \ldots, N\text{-}1$) discussed in Section 5.3.1.

$$D_{k,\vartheta}(\theta, \eta) = \left| \frac{sin(\eta\pi\vartheta(k)sin(\theta))}{\eta sin(\pi\vartheta(k)sin(\theta))} \right| \tag{5.10}$$

Since we want to bound the outer *sin* of the dirichlet, we replace the argument variable of the dirichlet to $sin(\theta)$. D is symmetric about $\theta = 0$ and attains maximum value at that angle. In this context, a symmetric function would suffice as we are interested mainly in the directivity observed in the frontal plane of a source. Following are scaling factors incorporated into the original dirichlet kernel: η determines the extent of directivity and ϑ is a linear mapping of [0, N-1] to (0, 0.78). The value 0.78 was found empirically and we use $\pi/4$ henceforth as it is a good approximation. For values of $\vartheta > \pi/4$ the sidelobes overshoot the mainlobe, which is undesirable since we want to maintain maximum power in the horizontal direction ($\theta = 0^o$) of the source. We use $sin(\theta)$ instead of $cos(\theta)$; *cos* would direct maximum energy in the perpendicular direction ($\theta = 90^o$). Extent of directivity η maps the sharpness of the directivity across frequency.

Basically, as frequency increases, the directivity D gets sharper towards the direction of maximum power. Considering the spectrum, η technically rescales the dirichlet kernel; it is homogeneous to temporal bandwidth. As η increases, the number of sidelobes increases and the sidelobe levels (SLL) decrease. A brief explanation of the original dirichlet kernel

*Concepts from antenna design draws several parallels with concepts discussed in the previous section, for building directivity or radiation patterns. Several applications in the communication field require antennas to have highly directional characteristics. This is accomplished by constructing a geometrical configuration of arrays (linear, circular, spherical, etc.) with specific relative displacements of the elements. The total field of the array is determined by combining the contribution of radiation fields by the individual elements as discussed in the previous section.

and its limit cases are presented in [98]. The working of our directivity function is presented in https://www.ajintom.com/dir-graphs

We now discuss the trends of how D evolves across parameters k, θ and η:

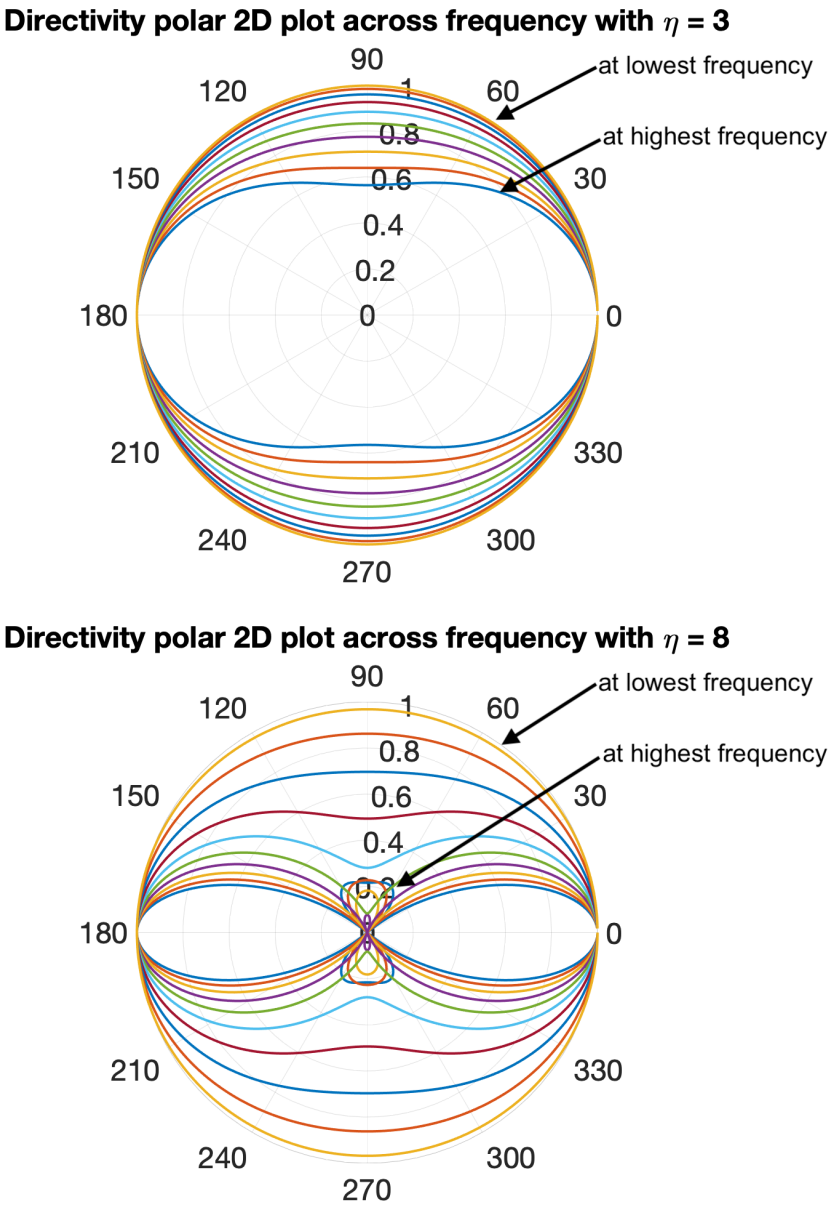

Fig. 5.5 2D polarplot of directivity varying across frequency for two values of η

Figure 5.5 illustrates how the directivity patterns evolve across frequencies and angles for different values of η. In Figure 5.5, the different colors represent the isodirectivity of D for 10 frequencies $k \in [0, N-1]$. The outermost plot is for the lowest frequency ($\vartheta(0) = 0$); we can see that the magnitude is roughly constant across all angles θ. The innermost plot, ($\vartheta(N-1) = \pi/4$), corresponding to the highest frequency, has varying directivity magnitude across the angles. For the highest frequency, at $\theta = 0^o$, the isodirectivity magnitude is maximum, at $\theta = 45^o$, it is minimum, and at $\theta = 90^o$ the magnitude is close to 0.6 for $\eta = 3$. At $\theta = 90^o$ the magnitude is close to 0.2 for $\eta = 8$. $D_{k,\vartheta}(\theta, \eta)$ assigns a flat response at lower frequencies. For $\eta = 3$, as in Figure 5.5, even at the highest frequency, the directivity remains pretty omni-directional as compared to a much sharper directivity at the highest frequency for $\eta = 8$.

Figure 5.6 illustrates how frequency response evolves across angles θ for fixed η. Here we observe that for angles closer to the $\theta = 0^o$ direction the frequency response is relatively flat. Towards the $\theta = 90^o$ direction, D evolves towards a low-pass filter with lower SLL. Figure 5.7 illustrates how frequency response evolves across extent of directivity η for fixed θ. We can see how directivity gets sharper for higher values of η. We illustrate just the following two figures in normalized magnitude scale for better visualization. Hereafter we present plots in dB scale to easily observe spectral modifications carried out by frequency responses due to directivity. Frequency is presented in linear scale (in Hz).

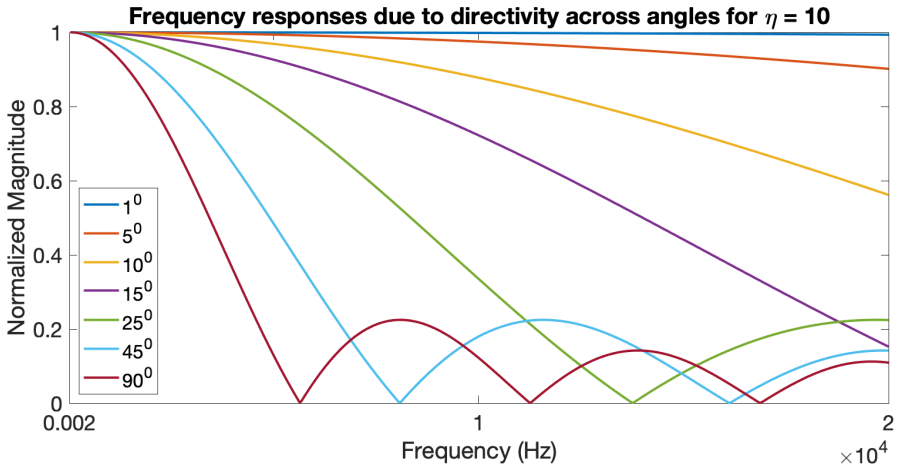

Fig. 5.6 Frequency responses across θ for fixed $\eta = 10$

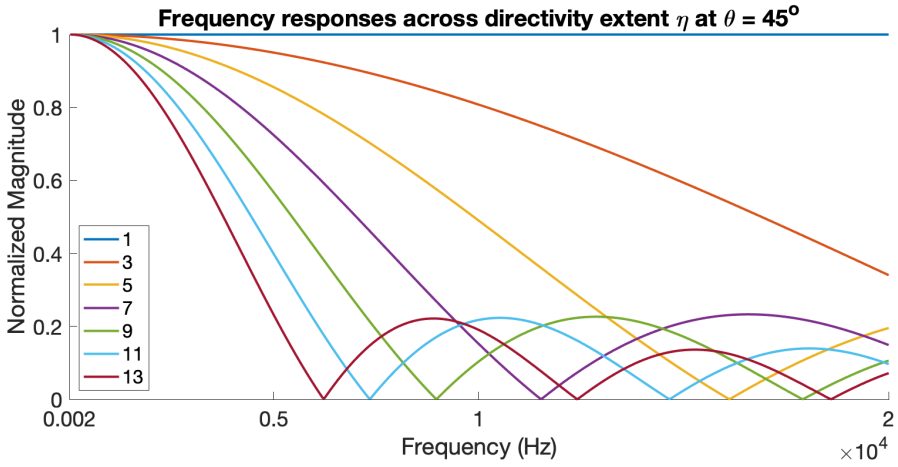

Fig. 5.7 Frequency responses across η for fixed $\theta = 45^o$

From the following figures (5.8 and 5.9) it is possible to observe how the frequency responses vary for different angles around the source. The effect of distance attenuation is removed in the following plots for ease of comparison.

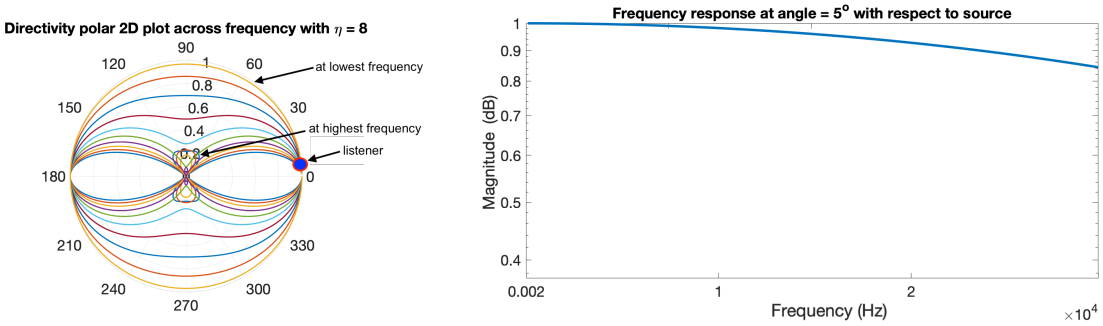

Fig. 5.8 Frequency response of the source at location of higher directivity

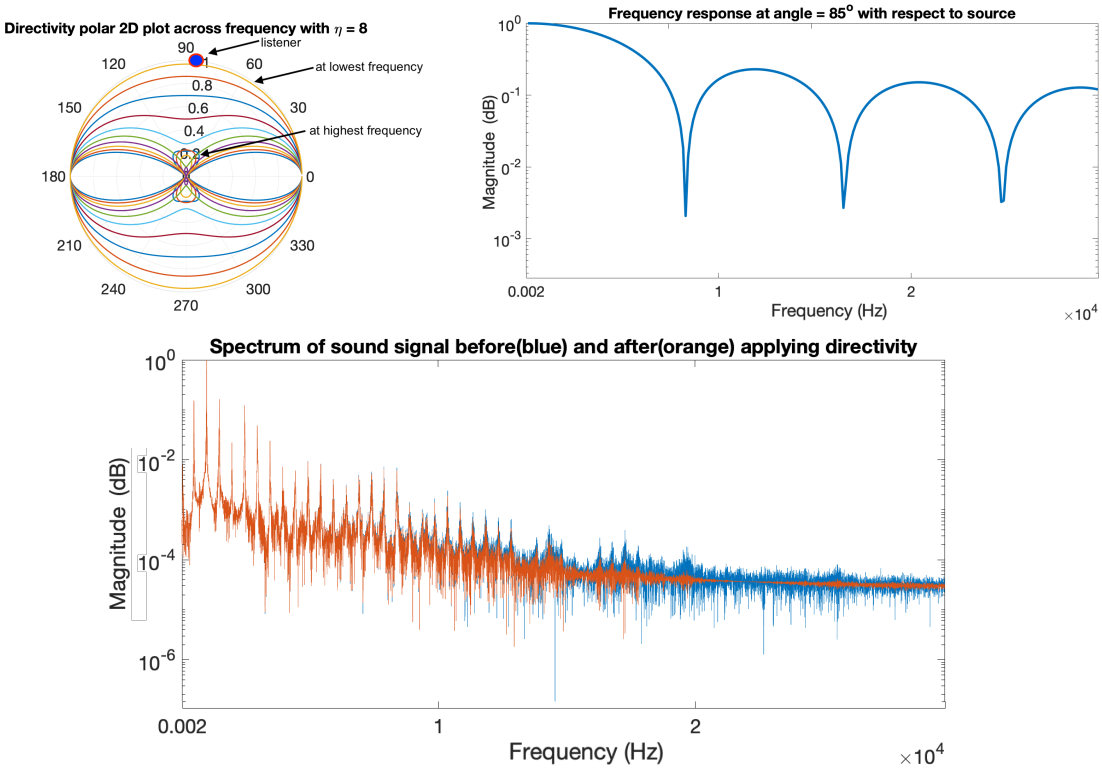

Fig. 5.9 (Top right figure): Frequency response of the source at location of relatively lower directivity at higher frequencies, (bottom figure): spectrum of sound signal (violin tone - B3) before (blue) and after (orange) applying frequency response due to directivity

5.3 Dynamic rendering of sound field

Directivity patterns of the various sources are used to determine the sound field of the sound stage. Eventually, the sound field produced by all the sources on the sound stage can be computed for one or more listening points. At a given point (observer/listener) in the sound field, the frequency response is computed based on the angle of the listener with respect to the source's horizontal axis. This angle θ will determine the power produced in that direction due to directivity. Distance r will determine signal attenuation. Therefore, from the directivity D, angle θ and distance r from the source, it is possible to determine the spatialization filter that is applied as a spectral modification to the STFT frame of the monophonic signal. One or more of these 3 parameters can vary over time, to carry out spatialization and unmasking across output tracks (discussed in detail in Section 5.4). The

idea to carry out spatialization is implemented by determining listening points (around the sound stage) whose resultant sound signals can be mapped to spatial audio reproduction systems. This section describes the various aspects involved in developing these time-varying spatialization filters.

5.3.1 System stability and limits

Since these spatialization filters are assigned for every STFT frame, it is important to consider time and spatial aliasing aspects and associated limits on the evolution of D. The frequency response due to directivity is calculated for every STFT frame. It is important to consider the limits on the evolution of the frequency response due to changes in directivity and/or orientation over time. To avoid time aliasing, we make sure that the STFT hop size is within the limits discussed in Section 3.5.2.

It is important to make the connection between physical units and sampling rates on the DSP side of things. We study the required spatial sampling using the case of a listener moving/scanning through the sound field. The frequency response at every point in the sound field is determined by computing the pressure at locations across a spatial grid. The spatial grid is sampled with minimum distance between spatial samples λ_{min} as follows:

$$\lambda_{min} = \frac{c}{2f_{max}} \tag{5.11}$$

$$\approx \frac{343}{2 \times 22050} \tag{5.12}$$

$$\approx 0.0077m \tag{5.13}$$

The trajectory of the listener is sampled at every STFT frame, every I time samples. Let us assume that the maximum speed a listener can move in the sound field is $v_{max} = 0.77$ m/s. To avoid aliasing, the listener can move at most 1 spatial sample per time sample ($1/f_s$ seconds).

$$\lambda_{min} = cT_s = vT \tag{5.14}$$

$$T = \frac{c}{v}T_s = \frac{\lambda_{min}}{v} = \frac{0.0077}{0.77} = 100 \text{ spatial samples} \tag{5.15}$$

Therefore, for an FFT size $N = 2^{15}$, the maximum hop size (discussion in section 3.5.2) equals $I_{max} = N/4 = 8192$ time samples. The target displacement $I_{s_{max}}$ in the sound field for each hop can be computed as follows to avoid aliasing:

$$I_{s_{max}} = \frac{v_{max}}{c} I_{max} = \frac{0.77}{343} 8192 \approx 18 \text{ spatial samples} \tag{5.16}$$

The following figures (5.10 and 5.11) illustrate examples of a moving listener across the sound field (the starting and finishing position is the leftmost and rightmost pink circle, respectively) rendered with the limits discussed above. We can notice how the frequency responses evolve over time. In Figure 5.10 (array of elementary sources), the listener is moving from left to right, towards the $\theta = 90^o$ direction. Since the elementary sources are arranged horizontally, the maximum power is in the $\theta = 90^o$ direction. The frequency response is evolving from a low pass filter towards a flat response.

As discussed in Section 5.2.2, D is designed to have maximum power in the horizontal direction near $\theta = 0$. In Figure 5.11 (template directivity) the listener is moving from left to right, so the listener is now close to this horizontal direction. Hence the frequency response evolves towards low-pass filters with lower SLLs across STFT frames.

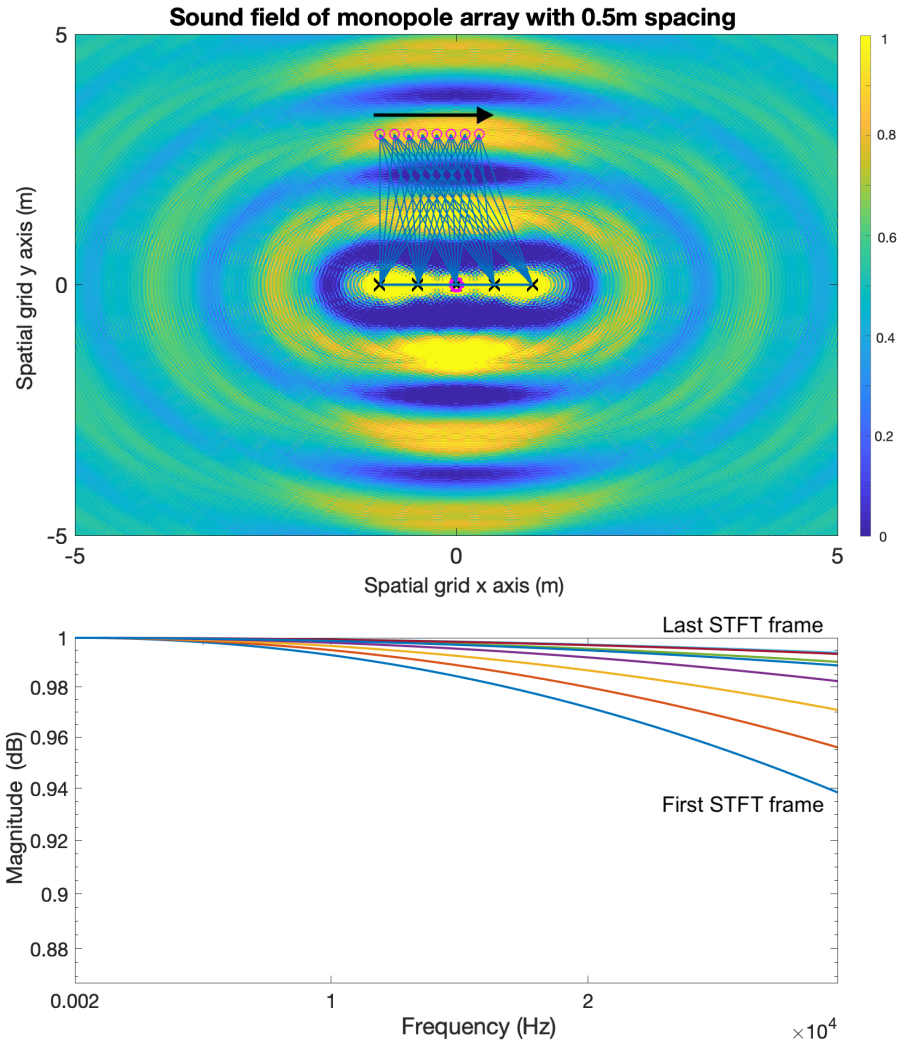

Fig. 5.10 Frequency responses across STFT frames for a moving listener rendered using monopole arrays

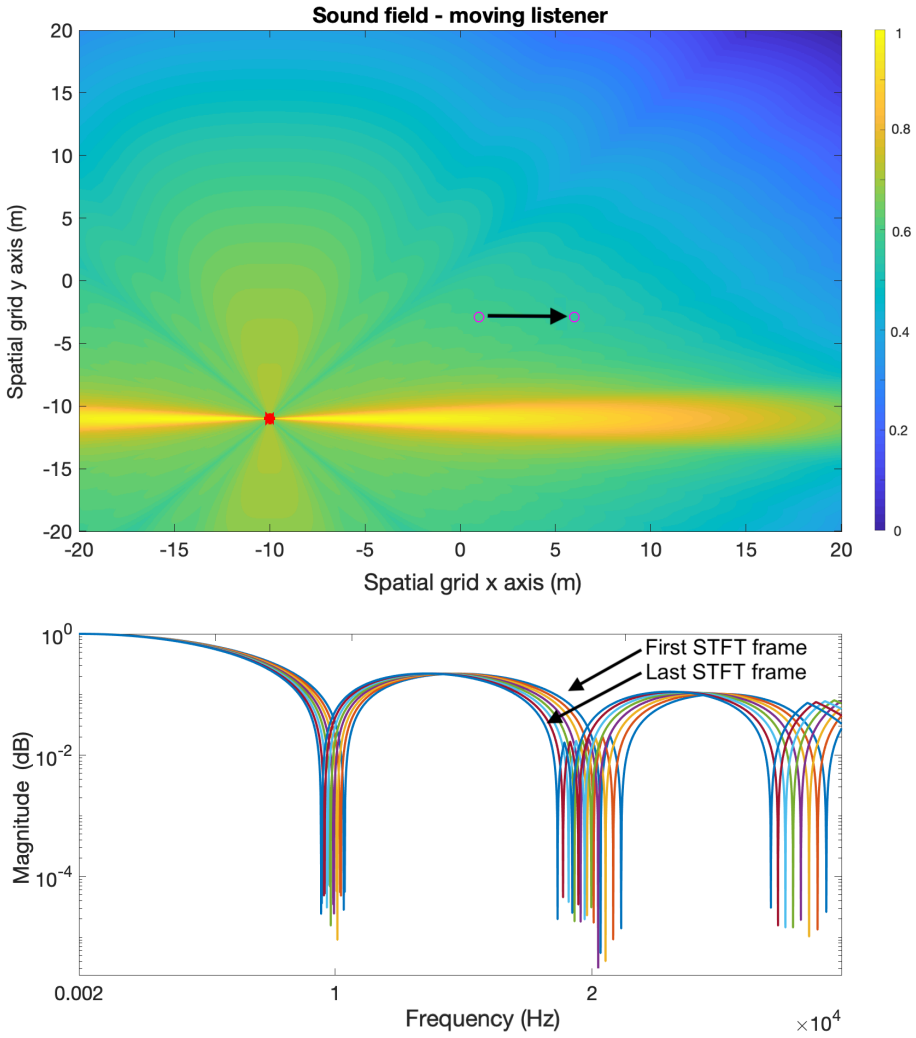

Fig. 5.11 Frequency responses across STFT frames for a moving listener with directivity (η=9) rendered using template radiation

5.3.2 Time-varying source directivity and orientation

In this section we present the main technique used to carry out time-varying spectral spatialization. In our system, the listener is fixed and the sources are assigned fixed positions on the sound stage. The frequency response H at the listener position evolves over time due to change in source directivities. D can evolve over time due to either a change in orientation of the source about its centroid or change in the directivity function itself. According to the linear system discussed in Chapter 3 (Figure 3.2), we have a spatialization filter

(spectral modification) for every orientation change ($D_{k,\vartheta}(\theta(qI), \eta)$) or directivity change ($D_{k,\vartheta}(\theta, \eta(qI))$) of the source in a given STFT frame. In the physical domain, time-varying orientation could be thought of as an acoustic instrument making rotations about its origin, for example a clarinet player pointing their clarinet in varying directions over time. An example for the second case of directivity changing over time is a trumpet performer using a mute at the bell causing a change in the timbre and directivity of the source itself (due to change in geometry). In either case we have a frequency response (directivity D) evolving over time. The limit applied on the speed of temporal evolution of D is established in Appendix B.

Since the spatialization filter discussed in this chapter revolves around a natural phenomenon such as directivity, we limit the maximum orientation change (angular velocity) made by a source with respect to a listener. We use the limits discussed earlier (Equation 5.15) with $v_{max} = 0.77$ m/s. The maximum displacement will take place at the point farthest from the centroid of the source. The magnitude of the displacement is the length of the chord of the circle between the initial and final positions of the farthest point [99]. The displacement by this farthest point (located at a distance r from the centroid of the source) caused by the angular rotation $\Delta\theta$ per STFT frame should be less than the maximum spatial displacement possible Δr_{max} per frame (refer [99] for derivation):

$$2rsin(\Delta\theta/2) < \Delta r_{max} \tag{5.17}$$

The maximum angle of rotation possible per frame is given as follows 5.18:

$$\Delta\theta_{max} = 2 * sin^{-1}(\Delta r_{max}/(2r)) \tag{5.18}$$

In the next section we illustrate simple examples of the frequency responses of these time-varying spatialization filters.

5.4 Optimization of the spatialization filters

In the previous chapter we presented how the sinusoidal panning filters (approach one) were optimized for unmasking by varying their parameters (frequencies υ and phase offsets δ). In the second approach of using directivity patterns as spatialization filters, the parameters to play with are the features of directivity D: extent of directivity η and angle θ of the

listener with respect to the source. As mentioned in Chapter 3, the spatialization filter parameters are optimized for each STFT frame. In the first approach the initialization of the parameters before the PSO was run are as follows: the amplitude envelope of the panning filters was determined by measuring masking of the track with respect to the rest of the mix (Section 3.6.2, Figure 3.4(b)). We also determined tracks that would undergo opposition panning based on inter-track spectral similarity (correlation matrix in Table 3.1). For the second approach of using directivity, we initialize the parameters as follows:
1) The sources that experience high masking with respect to the rest of the mix are assigned a higher value of η. This would mean that these sources will have sharper directivity at mid and high frequencies. Therefore, the optimized orientations (angular displacements) of the source over time carry out unmasking of the final signal out y. Each orientation is linked to its respective frequency response / spatialization filter which cause spectral modification of the original signal. Users can decide the range of η, or can choose it based on the radiation nature of the source itself. By default, the range is set to [1,10], 1 being omni-directional and 10 having higher directivity (informal listening suggested no perceivable difference beyond η=10), as shown in Figure 5.5.
2) The equivalent of opposition panning (carried out in approach one) is carried out by placing the sources on either sides of the sound stage with respect to the listener's axis. Users can either determine the positions of the source, or by default the sources would be placed on either sides using the Siegel-Tukey type ordering [87] mentioned in section 4.4.1.

Here we present a simple example of two sources (Figure 5.12), a flute and an oboe tone (A-4) each of 1 second duration. We let the PSO choose the directivity and position (within the bounds of the sound stage) for a listener placed at the origin of the grid. The PSO was run for the FFT of the entire signal. The sources were assigned parameters mentioned in Figure 5.13 after the PSO optimization. Figure 5.14 illustrates how the PSO cost over iterations decreases while reaching optimal directivity, source positions and orientations.

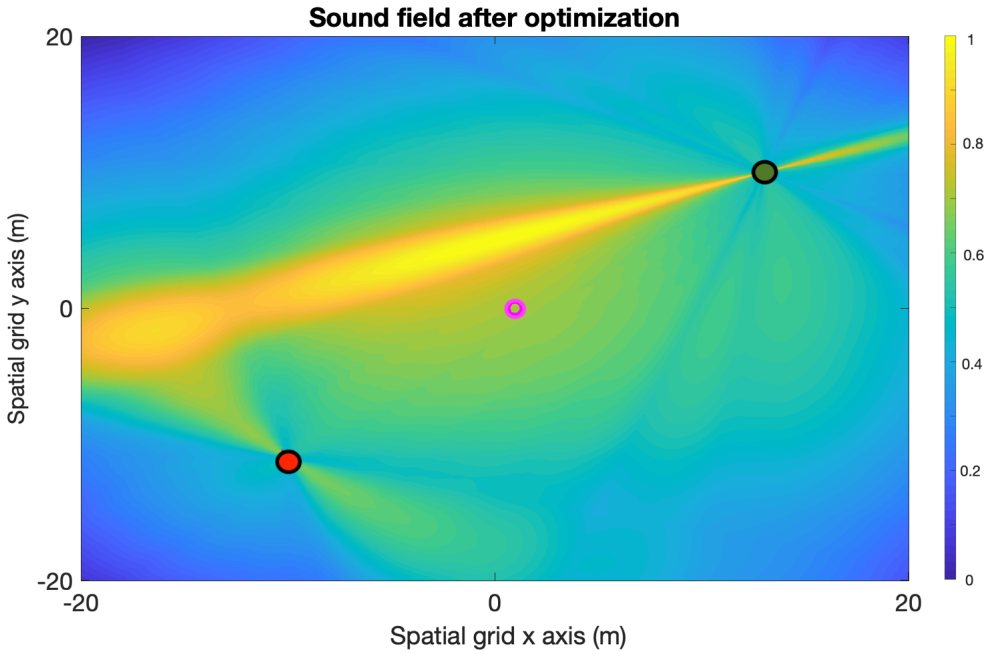

Fig. 5.12 Sound field after assigning directivity and optimization (green dot - oboe, red dot - flute, pink circle - listener)

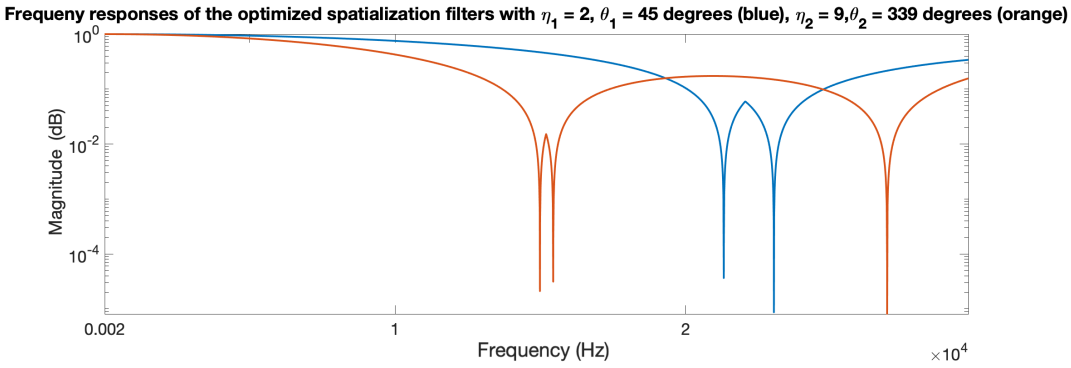

Fig. 5.13 Frequency responses of the optimized spatialization filters

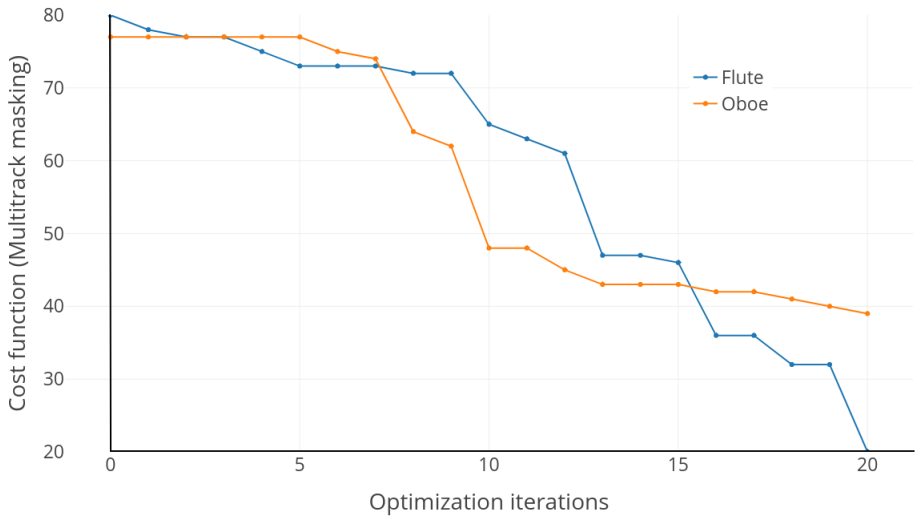

Fig. 5.14 PSO cost over iterations to optimize directivity, source positions and orientations

We investigated how the PSO would assign source orientation changes over time (STFT frames) for the two sources. The PSO assigned $\eta = 2$ and $\eta = 9$ for the flute and oboe respectively. Figure 5.15 illustrates the sound fields before and after the rotations. Figure 5.16 illustrate how the frequency responses evolves (due to orientation changes by the source) for both the sources across STFT frames at the listener location. From Figure 5.17 we can see how the PSO optimized unmasking such that spectral content at higher frequencies at and after 10 kHz is taken by either of the two sources. This is the similar behaviour we observed in approach one, in which alternate spectral bands occupied different parts of the spectral regions at the output tracks.

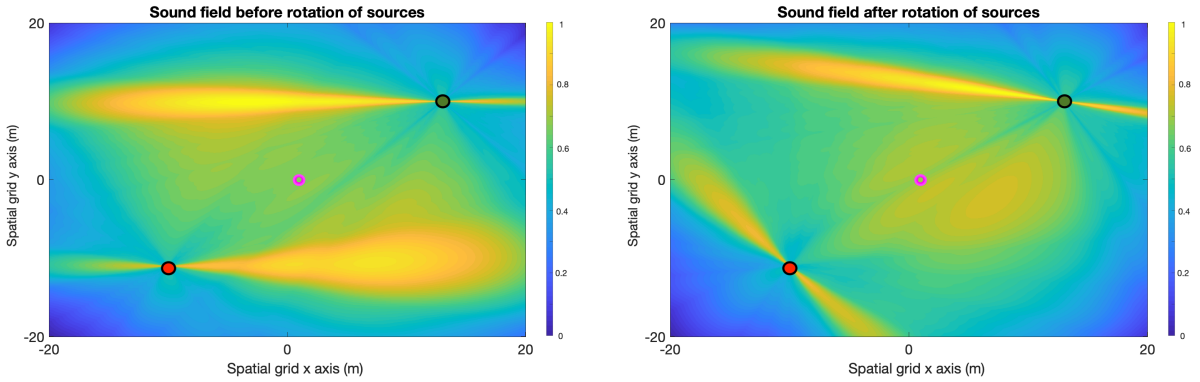

Fig. 5.15 Sound field before and after source rotations using PSO (green - oboe, red - flute, pink circle - listener)

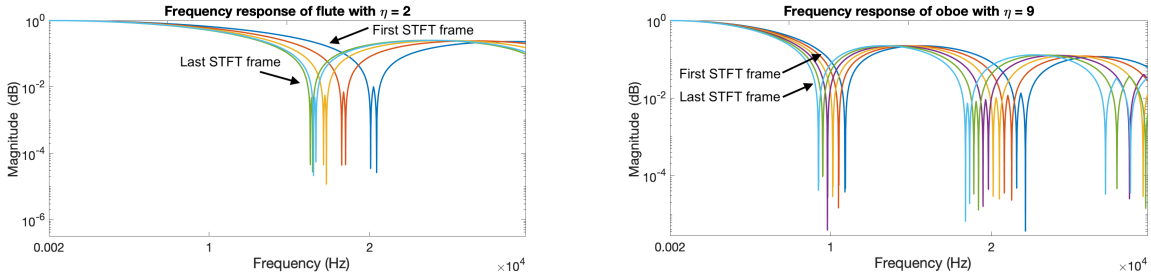

Fig. 5.16 Frequency responses of the rotating sources across STFT frames at listening point

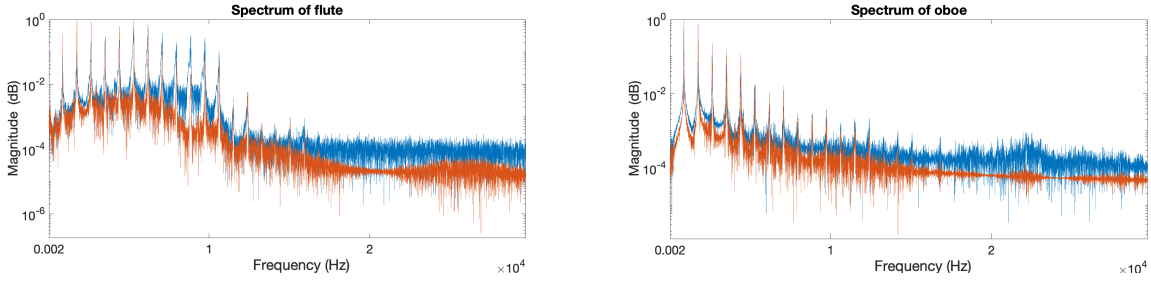

Fig. 5.17 Spectrum of sound signal before (blue) and after (orange) applying the frequency responses

5.5 Evaluation

5.5.1 Implementation

In the final implementation, we use the multitrack musical excerpts from approach one and study how the PSO works with directivity as spatialization filters. A semi-autonomous approach would be to let the user chose the initial values for parameters, η, θ and position for each source, as discussed in previous section, and let the automix system carry out optimized time-varying spatialization. In the fully autonomous approach, the algorithm determines both the initialization as well as optimization aspects. Since the proposed automix system of using directivity patterns as spatialization filters is aimed for multispeaker systems, we let the user choose the output tracks / speaker locations. These locations are basically the listening points which will be placed across the sound stage. Once the speaker locations are determined and the optimization is carried out, we have a matrix of $[L, M_o]$ signal values, where L is the length of each signal and M_o is the number of output tracks. The output signals y constitute the unmasking-optimized spatialized multitrack mix. It is useful to encode the optimized multitrack audio to B-format signals or Higher Order Ambisonics (HOA) [100] to avoid re-optimizing the mix for a new speaker layout. Once we have the mix optimized and encoded in B-fromat, it can be decoded for any given speaker layout / target playback format. For binaural listening[†], the user can choose a single location and the corresponding HRTF processing will produce the binaural left and right output. In figure 5.18, we can see how the sound stage has been arranged by the PSO with directivities assigned to each source.

[†]The binaural model and the implementation used in this algorithm is available at https://www.ajintom.com/618

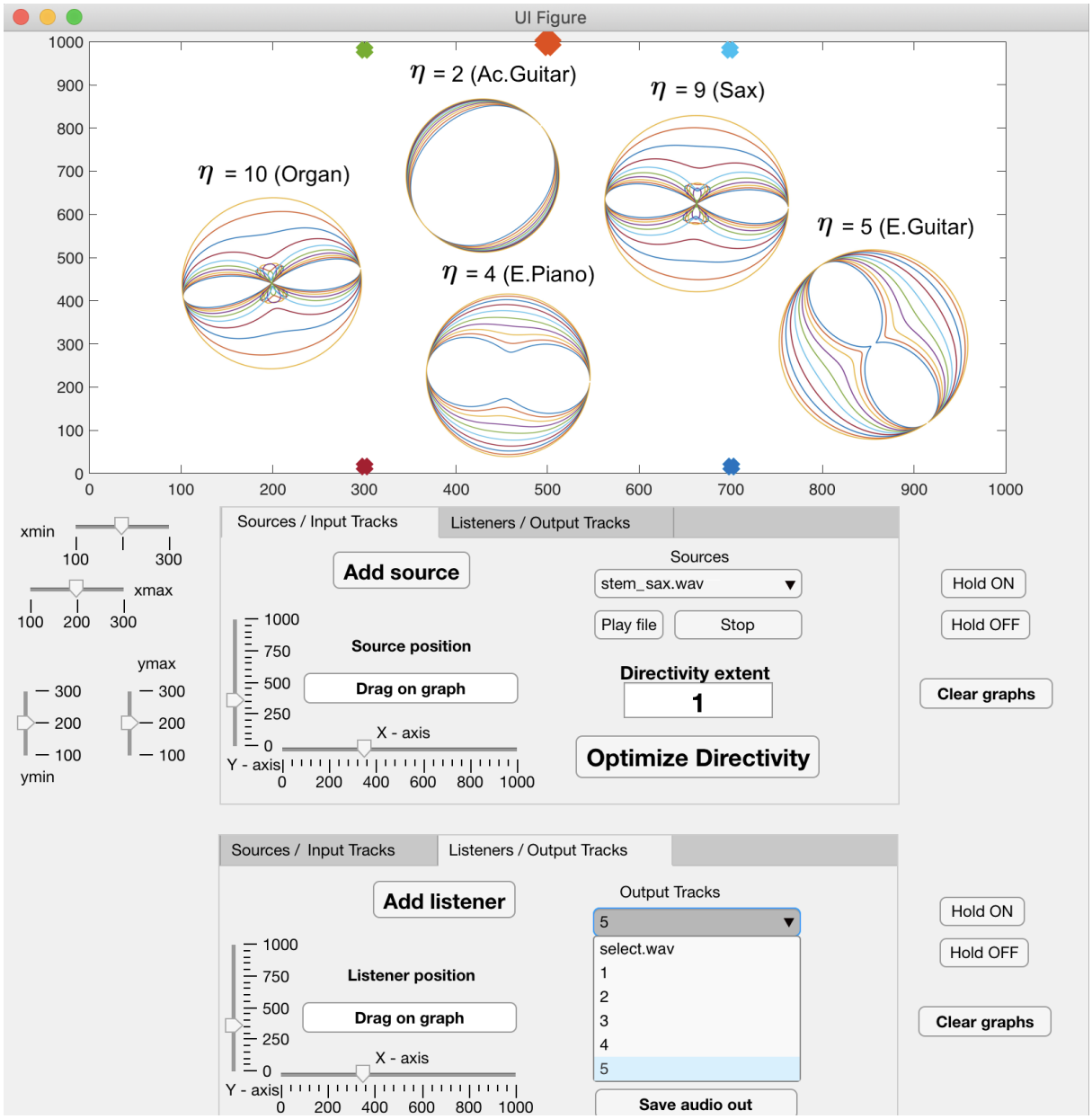

Fig. 5.18 PSO optimization of sound stage of 5 sources with 5 output tracks arranged similar to a surround system (green - front left, orange - centre, light blue - front right, maroon - rear left, blue - rear right)

5.5.2 Results and discussion

In this section we present the results of the proposed algorithms based on quantitative scores of spatialization and masking metrics, just as in Chapter 4. Figure 5.19 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:

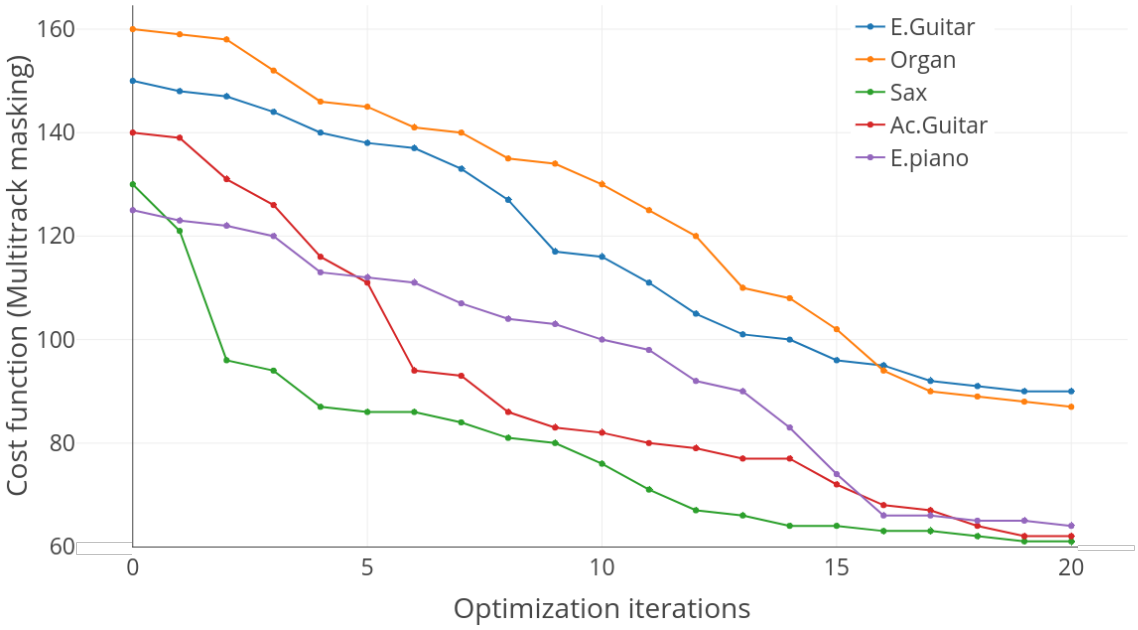

Fig. 5.19 PSO cost over iterations

In Table 5.1 we present the change in masking (20 PSO iterations) that occurred as a result of time-varying directivity for fixed orientation ($D_{k,\vartheta}(\theta, \eta(qI))$) as well as for varying orientation with fixed directivity ($D_{k,\vartheta}(\theta(qI), \eta)$) for the same 6 songs chosen from the Open Multitrack dataset [88] in Chapter 4. The perceptual masking metric (MPEG Psychoacoustic Model 1 [2]) in [1] is compared with the proposed multitrack masking measure M_m. Both metrics follow the same trend. The mix with time-varying source orientations ($D_{k,\vartheta}(\theta(qI), \eta)$) gave higher masking reduction than the mix with time-varying source directivity ($D_{k,\vartheta}(\theta, \eta(qI))$).

$\Delta Mask$		Folk	Country	Jazz	Funk	Pop	Rock
$D_{k,\vartheta}(\theta, \eta(qI))$	*[1]*	13.8	8.3	7.0	10.2	9.2	8.8
	ΔM_m	144	65	72	125	72	45
$D_{k,\vartheta}(\theta(qI), \eta)$	*[1]*	21.6	11.6	12.4	22.7	14.4	8.6
	ΔM_m	155	72	90	154	78	38

Table 5.1 Change in masking (unmasking amount): MPEG Multitrack Masking [1] and multitrack masking M_m

The amount of spatialization achieved was measured using a goniometer [10]:

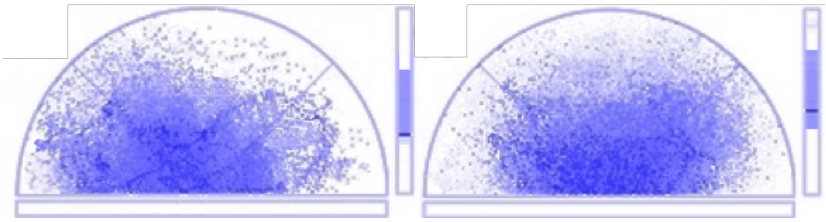

Fig. 5.20 Goniometer output - (left): $D_{k,\vartheta}(\theta, \eta(qI))$, (right): $D_{k,\vartheta}(\theta(qI), \eta)$

The sound examples discussed in this chapter are presented online at:
http://ajintom.com/auto-spatial

The discussion for the sound output and directivity optimization is presented along with the sound examples in the above link. The sound examples are available in multi-speaker formats such as 5.1 surround; however for convenience of listening and comparing with approach one (chapter 4), the sound output examples are down-mixed to binaural. In general, from informal listening we observe that the proposed automix system carries out significant amount of spatialization (also visible from the goniometer plots in Figure 5.20). In this approach of using directivity and spreading the sources across the sound stage, we hear binaural effects and achieve a good amount of externalization, thus making the mix sound natural. We can clearly localize the sources and hear them with high clarity. However, in some cases the PSO assigns extreme locations to spectrally dense sources thus making the mix sound extremely wide. In a dense multitrack with more than five tracks, it is difficult to perceive orientation changes of the sources distinctly; we perceive the orientation changes more as level variations. When the number of tracks are small (< 5),

it is easier to observe the orientation changes of sources, perceived as low-pass filters whose cut-offs vary over time. Overall, the proposed automix system performs well in terms of unmasking and spatialization. We can conclude that using source directivity helps create a plausible recreation of radiation properties of acoustic instruments; it is also an innovative and effective audio effect/tool to produce well-spatialized mixes for multispeaker playback systems.

Chapter 6

Conclusion

We will now summarize the findings and outcomes of this dissertation and suggest directions for future research. This includes possible improvements to the automatic spatialization tools presented in this work and scenarios which can make use of such systems.

6.1 Summary

In this dissertation, we developed a novel automatic mixing technique to carry out spatialization and unmasking of multitrack audio content. We first presented relevant background and concepts related to recording, multitrack mixing and spatial audio systems. We made links between the various concepts in acoustics/physics, signal processing and sound recording for readers from either of these backgrounds to understand the context of this work. We then gave an overview of automatic mixing, various approaches and building blocks to designing such systems. Later we present previous work in automatic spatialization and discussed a few shortcomings in the existing automix works. We discussed the idea of carrying out spectral panning/spatialization and then presented the framework of the proposed automatic mixing system which is a linear system (STFT framework) to which multitrack content (monophonic time domain signals) is fed in, processed (unmasking and spatialization) and fed out. We presented two novel approaches for carrying out multitrack spatialization, and our findings are discussed in the following section.

6.2 Discussion

From the first approach of using ERB-based sinusoidal panning filters, we learn that spectral panning is indeed an effective tool in carrying out spatialization provided that the filters comply to best practices. Without setting constraints on the panning filter, the algorithm would end up panning spectral content in an unusual manner; for example, low frequency content tied to one side of the stereo field. This would end up creating an unstable centre image and can be disturbing/distracting in headphone listening context. We noticed significant difference between the unoptimized and optimized mixes: the optimized mix complied strictly to the constraints set by the masking metric and panning rules, thereby producing a well-spatialized unmasked mix. The simplicity of the masking model enabled to build a real-time version of the algorithm of the automix system with the same framework, while still complying to perception. The takeaway is that it is not always necessary to use complex psychoacoustic models to achieve a desirable perception in an automix system. Instead we designed filters that comply to perception. In our initial experiments of trying to pan every alternate frequency bin, we could not perceive any amount of panning. Also we needed to address the fact that we have decreasing frequency resolution. This was our main motivation towards using ERB-based filters which exaggerated the width of the lobes of the filter at higher frequencies, thereby panning larger spectral bands at higher frequencies. The subjective results from the listening test conducted at Queen Mary University suggested that the proposed automix system produced mixes comparable to professional sound engineer mixes. This work drew lot of interest after our peer-reviewed paper was accepted and presented at the Audio Engineering Society (AES) Convention 2019 in Dublin, Ireland.

We then designed the second spatialization approach, in which we use the same optimization framework but with source directivities as spatialization filters, instead of handcrafted filters. This technique addressed how spatialization systems could place sound sources in other formats beyond stereo, such as ambisonics or 5.1 surround for example. This spatialization system developed directivity modelling using the following techniques: template directivity drawn from acoustics / geometry of the source as well as user-defined directivity for producing specific spatial effects. We illustrated simple examples to show how directivity patterns technically provide a frequency response at a listening point located at a certain distance and angle from the source. We studied how the sound field in

the sound stage is computed using directivities, orientations and positions of the sources. This system technically is a spectral panning linear system just like the first approach, except here the filters are drawn from the natural phenomenon of radiation of acoustical instruments. We learn that this system was useful not only as plausible recreation of this acoustical property of sources, but could be used as an audio effect / tool to carry out spectral spatialization. Though the objective scores for unmasking in the second approach were slightly lower than that of the first approach, informal listening suggested significant improvement in perceived spatialization and realism in the final mix while testing on various playback systems (mainly due to binaural effects and externalization).

The proposed automix systems 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 (clarity and intelligibility). Both spatialization approaches proved to give good sense of unmasking and spatialization. The automix sound output for both approaches are available online (http://ajintom.com/auto-spatial). The spatialization filter in approach one is more flexible for the PSO to work on, since it has a sinusoidal nature and is designed to carry out specific tasks in the three spectral regions. However, using source directivity as spatialization filters are restricted to low-pass structures for the model used in this dissertation. Indeed the considered filter is constrained by geometry and properties of acoustic radiation. On the other hand, since this approach considers the physical placement of sources across a sound stage, the final mix and the localization of sources is perceived to be more realistic because of binaural effects. This system is useful to carry out spatialization and upmixing for multispeaker systems. From this research, we achieved our general goal of carrying out unmasking and spatialization for multitrack audio material to be played back on various spatial audio systems.

6.3 Future work

As automatic mixing is a relatively new field of research, there are numerous directions the research could take. Current state-of-the-art in deep neural networks (DNNs) and related concepts in artificial intelligence (AI) could produce reliable automix systems that emulate traditional sound engineering practices. Such systems would use professional sound engineering projects (multitrack content) as datasets to learn how the various mixing parameters are assigned and/or change over time in a mix. AI can also be utilized to build

innovative audio effects and automix systems using the latest technologies like GANs [101] (generative adversarial nets). As we develop and share more multitrack datasets which include parameters used for individual tracks, their evolution over time in the mix, subjective evaluation of mixes of the same song with a good link between perceptual and objective scores, we should be able to produce reliable and robust automix systems. For example, in the context of this dissertation, in approach one (Chapter 4), we observed a variation of ratings across mixes of different genres. It might be worth letting the amount of panning for specific instruments across frequency depend on the genre of the multitrack being mixed. In the case of approach two (Chapter 5), it might be useful to learn the movements that musicians make on their acoustic instruments to have a more plausible evolution of directivity over time.

The aim of automix systems is not to solely replace sound engineers from the sound recording and production chain, rather it is to help sound engineers deal with the more creative aspects in mixing by letting the automix systems take care of the technical aspects in a mix. Sound engineers can now use a pre-mix to start with; the mixing parameters could be assigned (by the automix system) based on chosen / desirable metrics (by the user) like unmasking, spatial balance across output tracks, or even feed in a target (already existing) mix and let the automix system adjust EQ, panning, compression, etc. to match the automix to the target mix. Another motivation towards developing automix systems is to further develop existing traditional effects available on a mixing desk. For example, this dissertation deals with spectral panning which is not a very common feature on DAWs or mixing consoles (they usually have just a panpot to carry out panning). Now that we see from the results of this research that carrying out spectral panning could be useful in producing a relatively unmasked and spatialized mix, it might be worth building actual audio effects which sound engineers can further study and utilize in their mixing practices. Building up on approach one, sound engineers could choose from a palette of curves such as linear, Bezier, sinusoidal, etc. to draw panning filters on the input tracks, while analyzing output metrics like goniometer, unmasking measures and so on.

In the case of approach two, the idea would be to provide a palette of directivities which the user can choose for different frequencies of a source. Building such a system would make use of spherical harmonic structures. For example, the user can choose an omnidirectional directivity for 20 Hz, a dipole structure at 1000 Hz and a sharp directivity towards a desired angle at 20 kHz, and the audio effect / system would interpolate across all

frequencies to render the source directivity. These directivities could now be adapted for a given mix either by an optimization step like the one proposed in this dissertation or can be manually varied by the user for the specific scenario. Using spherical harmonic structures is also beneficial in the context of importing an existing measured directivity and rendering the same for a source in a mix. Another use-case would be to build a video context-aware source localization and spatialization system, in which the orientations of the source would be tied to the movement of sources in a video (of an acoustic performance for example) and the frequency response due to directivity would change for the output tracks in the sound stage. This would be an informed mix as we rely on external parameters such as videos, annotations, and/or scores. Most of the proposed work in automatic mixing considers only stereo playback. Expanding the current knowledge and implementations to surround sound, object-based audio, scene-based audio and related formats would be useful for the latest advent of VR/AR technologies. Audio content creators, hosts, consumer electronics manufacturers and broadcasters require technological frameworks to capture and render true-to-life immersive 3D audio experiences. The scene-based audio format [102] is designed to represent the audio scene as a field of pressure values at all points in a space over time. This is engineered to be an absolute and true representation of the 3D sound-scape. Latest formats like the MPEG-H which support scene-based audio technology are designed to overcome key limitations of traditional audio formats [90]. The proposed spatialization technique can be beneficial to design systems that create plausible 3D sound scapes. Using innovative audio effects/tools like source directivity coupled with optimization techniques can address how music can be meaningfully upmixed from the more common stereo to other playback formats like 5.1, 22.2, Ambisonics, etc.

Appendix A

Particle Swarm Optimization

Particle swarm optimization (PSO) is a population-based stochastic optimization technique developed by Eberhart and Kennedy in 1995 [20], inspired by social behaviour of bird flocking or fish schooling. PSO is a powerful tool with considerable flexibility, simplicity of use and implementation and versatility. PSO is a global optimization algorithm which is well suited to solve non-linear non-convex problems where the optimal solution is a point in a multidimensional search space of the variables (real-valued optimization). The PSO algorithm forms central technique in this dissertation for carrying out multitrack masking minimization. The following sections give a brief overview of mathematical optimization and describes the PSO algorithm.

A.1 Mathematical Optimization

An optimization problem consists of minimizing* a real function (objective function/fitness), $f(x)$ by systematically choosing input values/variables x (fitness values) in an efficient manner, from within a set χ (solution/search space bounded by lower and upper limits (b_l, b_u)) to determine the extremum value of the fitness function, subject to rules/constraints (Equation A.1):

$$\min{}^*{}_{x \subset \chi} f(x) \texttt{ subject to constraints} \tag{A.1}$$

Within this broad framework, optimization problems can have different mathematical

*or maximizing

properties. An important step in the optimization process is classifying the optimization model, since algorithms for solving optimization problems are tailored to their specificities. One possible criteria for classifying these optimization problems are direct and model-based. Direct algorithms determine search directions by computing values of the function f directly, whereas model-based algorithms construct and utilize a surrogate model of f to guide the search process. We further classify algorithms as local or global, with the latter having the ability to refine the search space arbitrarily. We also classify algorithms as stochastic or deterministic, depending upon whether they require random search steps or not. Another class of optimization algorithms are the derivative-based ones, which are used when the derivatives of the function are available. The most popular ones include the gradient descent, steepest descent, Newton methods and so on. These algorithms iteratively find the local minimum of the function using its derivatives and calculates the optimal step size and next direction to head in the solution space, until convergence. However, it is not always possible to theoretically extract the derivative information efficiently, and even when it is the case, sometimes the associated implementation procedures are non-trivial and time-consuming [103]. Derivative-free optimization has lately received considerable attention within the optimization community, including the establishment of solid mathematical foundations for many of the methods considered in practice. This is why in the context of this dissertation, we use one such derivative-free technique called the PSO.

A.2 PSO : Background concept

Particle swarm optimization (PSO) is a stochastic, population-based search method, inspired by social behaviour of bird flocking and fish schooling [104]. In the context of the algorithm we refer to these birds and fishes as 'particles' of a 'swarm'. This algorithm emulates the interaction between these particles to share information. There are a number of particles which move through the search space in search of the best solution. Every particle has a position that represents a potential solution and the goodness/fitness of that solution is measured by an objective/fitness function (the function being optimized). These particles also have velocities which direct the movement of particles in the search space.

Among the various kinds of optimization techniques, PSO has proven to give fairly convincing results in the context of audio, automatic mixing in particular [1, 17]. Mix engineers iteratively keep adjusting panning and EQ amounts of individual tracks until they

achieve a well spatialized clear mix. Similarly, the algorithm proposed in this dissertation relies on PSO with the same objective/fitness: to minimize multitrack masking and to create a well spatialized mix with high perceived quality. The variables/fitness values in this context refers to the parameters of the panning filters (frequency υ_m and phase offsets δ_m) in Chapter 4 and the directivity patterns (orientation θ_m and directivity η_m) in Chapter 5. Due to the complexity and the nonlinearity of this iterative process, the optimization process tends to have mutual multitrack influences (processing one track can affect the perception of another track(s)). Solving this multi-target non-convex optimization problem calls for the need of an evolutionary algorithm like the PSO, which is a heuristic-based approach to solving problems that cannot be easily solved in polynomial time. Unlike simple optimization techniques which involve moving a single individual around in the search space, the PSO algorithm involves moving a population of individuals or swarm particles around looking for a potential solution. PSO has been successfully applied in many areas: artificial neural network training, fuzzy system control, and other areas where Genetic Algorithms can be applied. One of the advantages of PSO over other derivative-free methods is the reduced number of parameters to tune and constraints acceptance.

A.3 PSO algorithm

The particle swarm algorithm begins by initializing a group of particles by assigning them random positions and velocities. At every iteration, it evaluates the objective function at each particle location, and determines the best position that gave the best (lowest) function value. It chooses new velocities, based on the current velocity, the particles' individual best locations, and the best locations of their neighbours. It then iteratively updates the particles' positions, velocities, and neighbours. Iterations proceed until the algorithm reaches a stopping criterion.

In PSO, the movement of the particles through the search space is governed by three factors: an *inertia weight* component, a *cognitive* component and a *social* component [105]. The inertia weight component allows a particle to maintain some momentum between iterations. This inertia prevents the particle from drastically changing direction by keeping track of previous flow of direction. The cognitive component allows the particle's movement to be influenced by its memory of good positions that it has found in earlier iterations. The social component will cause the good positions found by other members of the swarm to

influence the given particle's movement. The PSO model also has a stochastic component which appears as factors with the latter two components; it widens exploration of the search space.

Each particle i in the swarm is associated with two D-dimensional vectors: the current position $\underline{x}_i$ and the velocity $\underline{v}_i$. D is the dimensionality of the variables in the solution/search space in each direction. The search space is bounded by (b_l, b_u). The performance of each particle at position $\underline{x}_i$ is evaluated for the given problem, using an objective function. At each iteration k the local best position $\underline{L}_i^k$ is located and the global best position $\underline{G}$ is found after comparing all the solutions among all the particles in the set χ. Swarm size, N_s is an input parameter to the PSO algorithm. The best position in this context refers to the variables (position of the swarm particles) that give the minimum of the objective function (fitness). A new population is created based on a preceding one and the particles' velocities and positions are updated by the equation A.2 and A.3 respectively:

$$\underline{v}_i^{k+1} = w_i \cdot \underline{v}_i^k + c_1 \cdot \underline{r}_1 \otimes (\underline{L}^k – \underline{x}_i^k) + c_2 \cdot \underline{r}_2 \otimes (\underline{G} – \underline{x}_i^k) \tag{A.2}$$

$$\underline{x}_i^{k+1} = \underline{x}_i^k + \underline{v}_i^{k+1} \tag{A.3}$$

where:

$\underline{x}_i$ = particle position constrained by bounds (b_l, b_u)
$\underline{v}_i$ = particle velocity constrained by bounds $(-|b_u - b_l|, |b_u - b_l|)$
$\underline{L}_i^k$ = local best particle location (within iteration k)
$\underline{G}$ = global best/most promising location amongst the particles of the swarm
w_i = each particle's inertia
c_1 = personal acceleration coefficient / constant associated to cognitive weight
c_2 = social acceleration coefficient / constant associated to social weight
$\underline{r}_1$, $\underline{r}_2$ = vectors of stochastic components, random values uniformly drawn from $[0, 1]$
$\otimes$ = point-wise multiplication
The limits and choice of parameter values are discussed in the following section. Pseudocode for the PSO algorithm is presented in (Algorithm 1).

Algorithm 1 PSO Algorithm (x, b_l, b_u, N_s, w, c_1, c_2)

1: initialize parameters b_l, b_u, N_s, w, c_1, c_2
2: **for** each dimension d **do**
3: initialize location of global best in each direction G_d as ∞
4: **end for**
5: **for** all particles i **do**
6: **for** all dimensions d **do**
7: initialize random position x_{id} within bounds (b_l, b_u)
8: initialize random velocity v_{id} within bounds $(-|b_u - b_l|, |b_u - b_l|)$
9: **end for**
10: calculate fitness $f(x_i)$
11: $L \leftarrow x_i$
12: **if** $f(x_i) < f(G)$ **then**
13: $G \leftarrow x_i$
14: **end if**
15: **end for**
16: **repeat**
17: **for** each particle i in set χ **do**
18: **for** each dimension d **do**
19: pick random numbers: $r_{1_{id}}$, $r_{2_{id}} \sim U(0,1)$
20: update particle velocity according to A.2:
21: $v_{id}^{k+1} \leftarrow w_i \cdot v_{id}^k + c_1 \cdot r_{1_{id}} \otimes (L – x_{id}^k) + c_2 \cdot r_{2_{id}} \otimes (G – x_{id}^k)$
22: update particle position according to A.3:
23: $x_{id}^{k+1} \leftarrow x_{id}^k + v_{id}^{k+1}$
24: apply constraints/bounds by clamping x within (b_l, b_u)
25: **end for**
26: calculate fitness $f(x_i)$
27: **if** $f(x_i) < f(L)$ **then**
28: update particle's best known position:
29: $L \leftarrow x_i$
30: **if** $f(L) < f(G)$ **then**
31: update swarm's best known position:
32: $G \leftarrow L$
33: **end if**
34: **end if**
35: **end for**
36: **until** maximum iterations or minimum error criteria

A.4 Tuning the PSO: Parameter choice and control

The choice of PSO parameters w, c_1 and c_2 can have a large impact on the optimization performance. The parameters of the PSO algorithm must be chosen so as to properly balance between exploration and exploitation to avoid premature convergence to a local minima yet still ensure a good rate of convergence to the global minimum. The following points discuss the most common initialization strategies, choice of parameters as well as boundaries as seen in several works [105–109]:

- **Position and velocity initialization :** One of the best strategies involve random initialization of particles in which the velocity and position is drawn from a uniform distribution of the entire search space. Another common technique is to set very low random values (close to zero) for velocities; the exploration of the solution space is still guaranteed by choice of the initial positions.

- **Choice of inertia component w :** The inertia component is responsible for keeping the particle moving in the same direction it was originally heading. The value of the inertia coefficient is typically between 0.8 and 1.2, which can either dampen the particle's motion or accelerate the particle in its original direction.

- **Choice of acceleration coefficients c_1 and c_2 :** The cognitive component c_1, acts as the particle's memory, causing it to tend to return to the regions of the search space in which it has experienced high individual fitness. c_1 is usually close to 2, and affects the size of the step the particle takes toward its individual best candidate solution, L. The social component c_2 causes the particle to move to the best region the swarm has found so far. The social coefficient c_2 is also typically close to 2, and represents the size of the step the particle takes toward the global best candidate solution G the swarm has found up until that point.

- **Random values r_1 and r_2 :** The random values r_1 in the cognitive component and r_2 in the social component cause these components to have a stochastic influence on the velocity update. This stochastic nature causes each particle to move in a semi-random manner heavily influenced in the directions of the individual best solution of the particle and global best solution of the swarm.

- **Swarm size N_s:** The number of particles is another factor that may have an impact on the performances of the PSO. Though a larger population increases the diversity of the swarm and its exploration ability, it may also increase the probability or premature convergence and computational efforts. A common practice is to set swarm size as $N_s = 10 + \sqrt{D}$. A more refined setting is available as a lookup table in [107] which lists ideal swarm sizes for given values of inertial components, acceleration coefficients and problem dimension. The typical range for N_s is 20 - 40.

Appendix B

Directivity evolution over time

Following from Section 5.3.1 in Chapter 5, we derive the upper limit on the speed of the temporal evolution of D over time when θ evolves over time. We recall the template directivity equation (5.10) used in the dissertation:

$$D_{k,\vartheta}(\theta,\eta) = \left| \frac{sin(\pi\eta\vartheta(k)sin(\theta))}{\eta sin(\pi\vartheta(k)sin(\theta))} \right| \tag{B.1}$$

Following are the scaling factors incorporated into the original dirichlet kernel: η determines the extent of directivity and ϑ is a linear mapping of (0, $N-1$) to (0, 0.78). The value 0.78 was found empirically and we use $\pi/4$ henceforth as it is a good approximation. Also, the FFT parameters are as follows: $N = 2^{15}$, $I_{max} = N/4$, $k = 0, 1, 2, 3, \ldots, N\text{-}1$. We consider that for a given k, ϑ is a constant; this way we omit k in our calculations to find an upper limit on the speed of the temporal evolution of the directivity. Since the aim is to compute **upper** bounds and since ϑ is a linear mapping, we carry out the bound calculations for the highest frequency (at $\vartheta_{max} = 0.78 \approx \pi/4$) for the highest extent of directivity considered in this work (at $\eta = 10$). We also omit the absolute function on Equation B.1; this way we do not need to consider discontinuity while computing derivatives to obtain the bound (upper bound will not be at $\theta = 0$). So hereafter we work with the following function:

$$E_\eta(\theta) = \frac{sin(\pi\eta\vartheta(k)sin(\theta))}{\eta sin(\pi\vartheta(k)sin(\theta))} = \frac{sin(\eta z)}{\eta sin(z)} = diric_\eta(z) \tag{B.2}$$

with $z = \pi\vartheta sin(\theta)$

When θ varies over time, θ is a function $\theta(t)$, we then have $z(t) = \pi\vartheta sin(\theta(t))$ and $E_\eta(\theta(t)) = \frac{sin(\eta z(t))}{\eta sin(z(t))} = diric_\eta(z(t))$. Finding the bound on the speed of temporal variation of E_η requires the evaluation of $\frac{\partial}{\partial t}E_\eta$:

$$\frac{\partial}{\partial t}E_\eta(\theta(t)) = \frac{\partial}{\partial t}diric_\eta(z(t)) = \frac{\partial}{\partial t}z(t)\frac{\partial}{\partial z}diric_\eta(z(t)) \tag{B.3}$$

where $\frac{\partial}{\partial t}z(t) = \pi\vartheta cos(\theta(t))\frac{\partial}{\partial t}\theta(t)$
and J_η is a function defined as the derivative of the $diric_\eta$ function, so we have:

$$\frac{\partial}{\partial t}E_\eta(\theta(t)) = \pi\vartheta cos(\theta(t))\frac{\partial}{\partial t}\theta(t)J_\eta(z(t)) \tag{B.4}$$

We see that $\frac{\partial}{\partial t}E_\eta$ is proportional to $\frac{\partial}{\partial t}\theta(t)$.
For finding a bound on the speed of variation of E_η as a proportion of $\frac{\partial}{\partial t}\theta$, it is sufficient to find bounds B_1 and B_J on $\pi\vartheta cos(\theta(t))$ and J_η respectively.
We then have $\left|\frac{\partial}{\partial t}E_\eta(\theta)\right| < B_1 B_J \left|\frac{\partial}{\partial t}\theta(t)\right|$.

Finding B_1: We have $|\pi\vartheta cos(\theta(t))| \leq |\pi\vartheta| \leq |\pi\vartheta_{max}|$ as $|cos(\theta(t))| \leq 1$, then
$B_1 = \pi\vartheta_{max} \simeq \pi\frac{\pi}{4} \simeq 2.467$

Finding B_J: Let us first compute $J_\eta(z)$:

$$J_\eta(z) = \frac{\partial}{\partial z}diric(z) \tag{B.5}$$

$$= \frac{\eta cos(\eta z)\eta sin(z) - sin(\eta z)\eta cos(z)}{\eta^2 sin^2(z)} \tag{B.6}$$

$$= \frac{\eta^2 cos(\eta z)sin(z) - \eta sin(\eta z)cos(z)}{\eta^2 sin^2(z)} \tag{B.7}$$

$$= \frac{cos(\eta z)}{sin(z)} - cos(z)\cdot\frac{sin(\eta z)}{\eta sin^2(z)} \tag{B.8}$$

$$J_\eta(z) = \frac{cos(\eta z)}{sin(z)} - \frac{1}{tan(z)}\cdot\frac{sin(\eta z)}{\eta sin(z)} \tag{B.9}$$

We find the upper bound of $J_\eta(z)$ empirically at https://www.ajintom.com/dir-graphs. It is clear that $J_\eta(z)$ reaches B_J within the range corresponding to the mainlobe of the *diric*. The mainlobe region is deduced by finding the zeros of $diric(z)$ in its first interval, which

are at $-\pi/\eta$ and π/η. At the highest frequency, for $\eta = 10$ the mainlobe region lies between $[-\pi/10, \pi/10]$ and we observe the bound (positive maximum) $B_J = 4.365$.

Finally, $B_1 B_J = 10.768$ and then $\left|\frac{\partial}{\partial t} E_\eta(\theta)\right| < 10.768 \left|\frac{\partial}{\partial t}\theta\right|$. In other words, the largest evolution of D per time sample is computed as follows (just like in Equation 5.15):

$$\Delta D = \frac{\partial D}{\partial t}\Delta t = \frac{\partial D}{\partial t} I \tag{B.10}$$

References

[1] 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).

[2] 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).

[3] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools, 3rd ed.* (NY: Routledge, New York, USA) (2018).

[4] R. Izhaki, *"Panning" in Mixing Audio: Concepts, Practices and Tools, 3rd ed. chapter 14* (Focal Press/Elsevier, Burlington, USA) (2018).

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

[6] J. D. Reiss, "Automation for the people," *Proceedings of the 17th IEEE International Conference on Digital Signal Processing, Corfu, Greece*, (6 pages) (6-8 July, 2011).

[7] A. Pras, C. Guastavino, M. Lavoie, "The impact of technological advances on recording studio practices," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 3, pp. 612–626 (2013).

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

[9] V. Verfaille, U. Zölzer, 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).

[10] 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).

[11] E. Perez-Gonzalez, J. D. Reiss, "A real-time semiautonomous audio panning system for music mixing," *EURASIP Journal on Advanced Signal Processing https://doi.org/10.1155/2010/436895*, vol. 2010, (10 pages) (2010).

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

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

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

[15] B. G. Shinn-Cunningham, "Influences of spatial cues on grouping and understanding sound," *Proceedings of the Forum Acusticum, Budapest, Hungary* (2005).

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

[17] A. Tom, J. D. Reiss, P. Depalle, "An automatic mixing system for multitrack spatialization for stereo based on unmasking and best panning practices," *146th Audio Engineering Society Convention, Dublin, Ireland*, (10 pages) (March 20-23, 2019).

[18] R. Caussé, J. Bresciani, O. Warusfel, "Radiation of musical instruments and control of reproduction with loudspeakers," *Proceedings of ISMA, Tokyo, Japan*, pp. 67–70 (1992).

[19] F. Giron, "Investigations about the directivity of sound sources," *Ph.D. Dissertation, Ruhr-Universitat, Bochum, Shaker Verlag, Aachen, Germany* (1996).

[20] R. Eberhart, J. Kennedy, "Particle swarm optimization," *Proceedings of the IEEE international conference on neural networks (ICNN'95), Perth, Australia*, vol. 4, pp. 1942–1948 (1995).

[21] J. M. Eargle, *Music, sound, and technology, 2nd ed. New York, USA* (Springer Science & Business Media) (2013).

[22] W. Moylan, *Understanding and Crafting the Mix, 2nd ed. Boston, USA* (Focal Press) (2007).

[23] E. Zwicker, H. Fastl, *Psychoacoustics — Facts and Models, 3rd ed. New York, USA* (Springer-Verlag) (2007).

[24] G. Gatzsche, F. Melchior, "Spatial audio Authoring and Rendering: Forward Research through Exchange," *International Computer Music Conference (ICMC), Belfast, Ireland*, (2 pages) (August, 2008).

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

[26] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466 (1997).

[27] V. Pulkki, T. Lokki, D. Rocchesso, *"Spatial Effects" in DAFX: digital audio effects, 2nd ed. chapter 5* (John Wiley & Sons, New Jersey, USA) (2011).

[28] H. Wallach, E. B. Newman, M. R. Rosenzweig, "A precedence effect in sound localization," *The Journal of the Acoustical Society of America*, vol. 21, no. 4, pp. 468–468 (1949).

[29] 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).

[30] M. Frank, "Localization using different amplitude-panning methods in the frontal horizontal plane," *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany*, pp. 41–47 (April 3-5, 2014).

[31] F. Ortolani, "Introduction to Ambisonics," *Ironbridge Electronics (http://www.ironbridge-elt.com/downloads/FrancescaOrtolani-IntroductionToAmbisonics.pdf)* (2015).

[32] A. Berkhout, D. de Vries, P. Vogel, "Wave front synthesis: a new direction in electroacoustics," *93rd Audio Engineering Society Convention, San Francisco, USA* (October 1-4, 1992).

[33] C. P. Brown, R. O. Duda, "A structural model for binaural sound synthesis," *IEEE transactions on speech and audio processing*, vol. 6, no. 5, pp. 476–488 (1998).

[34] J. C. Bennett, K. Barker, F. O. Edeko, "A new approach to the assessment of stereophonic sound system performance," *Journal of the Audio Engineering Society*, vol. 33, no. 5, pp. 314–321 (1985).

[35] T. Lossius, P. Baltazar, T. de la Hogue, "DBAP–distance-based amplitude panning," *International Computer Music Conference (ICMC), Montreal, Canada* (August, 2009).

[36] J. M. Chowning, "The simulation of moving sound sources," *Journal of the Audio Engineering Society*, vol. 19, no. 1, pp. 2–6 (1971).

[37] F. P. Preparata, M. I. Shamos, *Computational geometry: an introduction* (Springer Science & Business Media) (1985).

[38] V. Pulkki, "Uniform spreading of amplitude panned virtual sources," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, USA*, pp. 187–190 (1999).

[39] P. Power, "Future spatial audio: Subjective evaluation of 3D surround systems," *Ph.D. Dissertation, University of Salford, United Kingdom* (2015).

[40] T. M. Holman, "Front loudspeaker directivity for surround sound systems," *US Patent 9,729,992* (2017).

[41] G. Theile, H. Wittek, "Principles in surround recordings with height," *130th Audio Engineering Society Convention, London, United Kingdom* (May 13-16, 2011).

[42] J. R. Stuart, "The psychoacoustics of multichannel audio," *11th Audio Engineering Society Conference : Audio for New Media (ANM), London, United Kingdom* (March 1, 1996).

[43] K. Hamasaki, K. Hiyama, R. Okumura, "The 22.2 multichannel sound system and its application," *118th Audio Engineering Society Convention, Barcelona, Spain* (May 28-31, 2005).

[44] A. Roginska, P. Geluso, *Immersive sound: The art and science of binaural and multichannel audio* (Taylor & Francis) (2017).

[45] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society*, vol. 33, no. 11, pp. 859–871 (1985).

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

[47] S. Braun, M. Frank, "Localization of 3D ambisonic recordings and ambisonic virtual sources," *1st International Conference on Spatial Audio, Detmold, Germany* (2011).

[48] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms, and Acoustic Virtual Reality, 1st ed.* (Springer Verlag) (2008).

[49] J. Breebaart, J. Herre, L. Villemoes, C. Jin, K. Kjörling, J. Koppens, J. Plogsties, "Multi-channel goes mobile: MPEG Surround binaural rendering," presented at the *29th Audio Engineering Society Conference: Audio for Mobile and Handheld Devices, Seoul, Korea* (September 2-4, 2006).

[50] E. Perez-Gonzalez, "Advanced Automatic Mixing Tools for Music," *Ph.D. Dissertation, Queen Mary University of London, United Kingdom* (2010).

[51] 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).

[52] 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*, (4 pages) (2009).

[53] M. J. Terrell, "Perceptual Mixing for Musical Production," *Ph.D. Dissertation, Queen Mary University of London, United Kingdom* (2012).

[54] Z. Ma, "Intelligent Tools for Multitrack Frequency and Dynamic Processing," *Ph.D. Dissertation, Queen Mary University of London, United Kingdom* (2016).

[55] B. De Man, J. D. Reiss, "A semantic approach to autonomous mixing," *Journal on the Art of Record Production (JARP), no. 8* (December 2013).

[56] B. De Man, J. D. Reiss, "The mix evaluation dataset," *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx), Edinburgh, UK*, pp. 436–442 (September 5–9, 2017).

[57] J. Scott, M. Prockup, E. M. Schmidt, Y. E. Kim, "Automatic multi-track mixing using linear dynamical systems," *Proceedings of the 8th Sound and Music Computing Conference, Padova, Italy* (July 6-9, 2011).

[58] M. Senior, *"Mixing secrets for the small studio", 2nd ed.* (NY: Routledge, New York, USA) (2019).

[59] M. B. Cartwright, B. Pardo, "Social-EQ: Crowdsourcing an Equalization Descriptor Map," *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR), Curitiba, Brazil*, pp. 395–400 (November 4-8, 2013).

[60] P. Seetharaman, B. Pardo, "Crowdsourcing a reverberation descriptor map," *Proceedings of the 22nd ACM international conference on Multimedia, Orlando, USA*, pp. 587–596 (November 3-7, 2014).

[61] B. McCarthy, *Sound systems: design and optimization: modern techniques and tools for sound system design and alignment, 3rd ed.* (Focal Press) (2016).

[62] 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*, (6 pages) (7-10 May, 2015).

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

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

[65] H. Glyde, J. M. Buchholz, H. Dillon, S. Cameron, L. Hickson, "The importance of interaural time differences and level differences in spatial release from masking," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. EL147–EL152 (2013).

[66] J. B. Allen, L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564 (1977).

[67] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83 (1978).

[68] 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).

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

[70] Z. Ma, J. D. Reiss, D. A. Black, "Partial loudness in multitrack mixing," *53rd Audio Engineering Society Conference: Semantic Audio, London, UK* (Jan 27-29, 2014).

[71] M. R. Schroeder, B. S. Atal, J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *The Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1647–1652 (1979).

[72] T. Thiede, *et al.*, "PEAQ-The ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29 (2000).

[73] 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).

[74] M. Karjalainen, "A new auditory model for the evaluation of sound quality of audio systems," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Tampa, USA*, vol. 10, pp. 608–611 (1985).

[75] E. Perez-Gonzalez, J. D. Reiss, "Improved control for selective minimization of masking using interchannel dependancy effects," *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx), Espoo, Finland*, pp. 12–19 (September 1-4, 2008).

[76] S. Vega, J. Janer, "Quantifying masking in multi-track recordings," *Proceedings of the 8th Sound and Music Computing Conference, Barcelona, Spain*, (8 pages) (July 21-24, 2010).

[77] E. Perez-Gonzalez, J. D. Reiss, "Automatic equalization of multichannel audio using cross-adaptive methods," *127th Audio Engineering Society Convention, New York, USA* (October 9-12, 2009).

[78] S. Hafezi, J. D. Reiss, "Autonomous multitrack equalization based on masking reduction," *Journal of the Audio Engineering Society*, vol. 63, no. 5, pp. 312–323 (2015).

[79] 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).

[80] J. E. Dennis Jr, R. B. Schnabel, "Numerical methods for unconstrained optimization and nonlinear equations," *SIAM*, vol. 16 (1996).

[81] 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).

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

[83] E. Benjamin, "An experimental verification of localization in two-channel stereo," *121st Audio Engineering Society Convention, San Fransisco, USA*, (14 pages) (Oct 5-8, 2006).

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

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

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

[87] S. Siegal, *Nonparametric statistics for the behavioral sciences* (McGraw-hill, New York), 2nd ed. (1988).

[88] B. De Man, M. Mora-Mcginity, G. Fazekas, J. D. Reiss, "The open multitrack testbed," *137th Audio Engineering Society Convention, Los Angeles, USA*, (4 pages) (Oct 9-12, 2014).

[89] B. De Man, J. D. Reiss, "APE: Audio perceptual evaluation toolbox for MATLAB," *136th Audio Engineering Society Convention, Berlin, Germany*, (4 pages) (April 26-29, 2014).

[90] J. Herre, J. Hilpert, A. Kuntz, J. Plogsties, "MPEG-H audio—the new standard for universal spatial/3D audio coding," *Journal of the Audio Engineering Society*, vol. 62, no. 12, pp. 821–830 (2014).

[91] J. Meyer, *Acoustics and the Performance of Music: Manual for Acousticians, Audio Engineers, Musicians, Architects and Musical Instrument Makers, 5th ed. New York, USA* (Springer-Verlag) (2009).

[92] G. S. Kendall, "Spatial perception and cognition in multichannel audio for electroacoustic music," *Organised Sound*, vol. 15, no. 3, pp. 228–238 (2010).

[93] A. M. Pasqual, "Sound directivity control in a 3-D space by a compact spherical loudspeaker array," *Ph.D. Dissertation, Universidade Estadual de Campinas, Campinas, Brazil* (2010).

[94] A. Chaigne, J. Kergomard, *Acoustics of Musical Instruments*, Modern Acoustics and Signal Processing (Springer-Verlag, New York, USA) (2016).

[95] M. C. Junger, D. Feit, *Sound, structures, and their interaction*, vol. 225 (MIT press Cambridge, MA) (1986).

[96] C. A. Balanis, *Antenna theory: analysis and design* (John wiley & sons) (2016).

[97] A. Bashirov, *Fourier Series and Integrals, chapter 12* (Elsevier) (2014).

[98] Mathonline, "Dirichlet's Kernel," http://mathonline.wikidot.com/dirichlet-s-kernel (2015), last accessed on 30 July, 2019.

[99] B. Hughes, "Dirichlet's Kernel," https://www.ck12.org/trigonometry/Length-of-a-Chord/lesson/Length-of-a-Chord-TRIG/ (2017), last accessed on 15 July, 2018.

[100] A. Farina, "Software implementation of B-format encoding and decoding," *104th Audio Engineering Society Convention, Amsterdam, Netherlands* (May 16-19, 1998).

[101] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, pp. 2672–2680 (2014).

[102] Qualcomm, "Scene-based Audio for MPEG-H," https://www.qualcomm.com/scene-based-audio (2016), last accessed on 18 July, 2019.

[103] O. Kramer, D. Ciaurri, S. Koziel, "Derivative-free optimization," *In: S. Koziel & X.S. Yang (eds.) Computational Optimization, Methods and Algorithms, SCI 356, Berlin, Germany*, pp. 61–83 (2011).

[104] J. Kennedy, "Particle swarm optimization," *In: C. Sammut, G. I. Webb (eds.) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, USA*, pp. 967–972 (2017).

[105] F. Marini, B. Walczak, "Particle swarm optimization (PSO). A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 153–165 (2015).

[106] X. Hu, "PSO Tutorial," http://www.swarmintelligence.org/tutorials.php (2006), last accessed on 30 May, 2018.

[107] M. E. H. Pedersen, "Good parameters for particle swarm optimization," *Hvass Laboratories, Copenhagen, Denmark, Tech. Rep. HL1001*, (12 pages) (2010).

[108] E. Mezura-Montes, C. A. C. Coello, "Constraint-handling in nature-inspired numerical optimization: past, present and future," *Swarm and Evolutionary Computation*, vol. 1, no. 4, pp. 173–194 (2011).

[109] A. E. Olsson, "Particle Swarm Optimization: Theory, Techniques and Applications," *Nova Science Publishers, Inc., Commack, USA* (2011).