SCIRES it
SCIentific RESearch and Information Technology
Ricerca Scientifica e Tecnologie dell'Informazione
Vol 15, Issue 2 (2025), 233-250
e-ISSN 2239-4303, DOI 10.2423/i22394303v15n2p233

CASPUR-CIBER Publishing, http://www.sciresit.it

FROM STUDIO TO STREAMING: A PRACTICAL FRAMEWORK FOR BINAURAL MUSIC PRODUCTION USING "MANTRA" BY ESTADOS ALTERADOS

*César Alonso Cardona-Cano**

* University of Medellín– Medellín, Colombia.

Abstract

This artistic research article documents the end-to-end production of a fixed-binaural remix of "Mantra" by Estados Alterados. We designed a reproducible, hybrid workflow—dummy-head re-amping, HRTF-based spatial synthesis, and conventional multitrack mixing—to translate narrative goals (Reality ↔ Beyond) into perceptible spatial events for headphone listening. Materials included original stems, time-charts and spatial diagrams; decisions emphasized axial anchors with selective periphery, preserving timbral identity while enhancing externalization and presence. Validation covered multi-headphone listening, stereo fold-down checks and streaming-target mastering. Deliverables comprise the published track and a browser-based interactive interface that supports knowledge transfer between academia and industry; computational mood/timbre profiling with Essentia/MusiCNN complements the qualitative analysis.

Keywords

Binaural Audio; Immersive Music; Spatial Sound; Music Production; Binaural Music; Immersive Narrative

1. Introduction

Binaural audio is based on two-channel signals that encode interaural time differences (ITD), interaural level differences (ILD), and spectral filtering, allowing headphone listeners to perceive sources as externalized beyond the head (Blauert, 1997; Begault, 1994, Sun, Zhong, & Yost, 2015). In practice, this principle provides producers with a three-dimensional field where instruments can be positioned above, behind, or to the sides of the listener—a resource particularly valuable when narrative intent must be translated into concrete mixing decisions (Møller, 1992; Roginska & Geluso, 2017). More broadly, immersive systems that combine immersion and interaction can measurably increase user engagement and support learning in cultural settings (Carrozzino et al., 2013).

Technological advances and evolving listening habits have fostered an increasingly immersive sound environment. Binaural techniques, which emulate the auditory system through interaural differences of time, level and phase, enable more precise localization of acoustic events (Møller, 1992; López et al., 2022). For music producers, this translates into a three-dimensional canvas where instruments can circulate across multiple spatial planes, shaping narrative trajectories and emotional dynamics that exceed the limits of conventional stereo mixing (De Gregori Astrici, 2018). Empirical studies confirm that deliberate control of level, timbre, and spatial positioning enhances both immersion and affective response (Fontana et al., 2007). At the same time, greater streaming bandwidth and widespread access to high-fidelity headphones have made large-scale distribution of such content feasible (Rumsey, 2021).

Commercial signals reinforce this trend: Grand View Research (2022) valued the global 3D and immersive-audio market at roughly USD 4.5 billion in 2021 and projected an annual growth rate of about 17% through 2030. After enabling Dolby Atmos content, Apple Music reported double-digit growth in spatial-audio streams (Singleton, 2021). This data suggests that listener demand is keeping pace with technological supply; but the practical challenge remains: how can binaural techniques be systematically integrated into everyday studio workflows?

Within this context, we examine how the perceptual constructs of presence, externalization, and co-immersion can be translated into tangible creative choices (Slater & Wilbur, 1997; Reardon et al., 2018; Fantini et al., 2023). Here, *presence*

denotes the listener's sense of being there; *externalization* refers to the placement of sound outside the head; and *co-immersion* to the fluid integration of virtual and recorded sources within a single scene. To operationalize this aim, we combine dummy-head re-amping, HRTF-based synthesis, and traditional multichannel mixing, thereby expanding the expressive repertoire of contemporary music production.

The methodological integration of these techniques enables further enquiry. Specifically, it supports (i) formulating best-practice guidelines for hybridizing dummy-head captures with multitrack sessions; (ii) devising strategies to translate lyrical narratives into spatial gestures; and (iii) producing headphone-optimized masters that retain loudspeaker compatibility.

This project was made possible through a collaboration between two Colombian universities and *Estados Alterados,* an electronic-rock band from Medellín, combining psychoacoustic research with the artistic demands of a professional ensemble and culminating in a publishable musical artefact. This article presents a reproducible workflow that integrates narrative analysis with a hybrid chain of binaural re-amping, HRTF-based spatialization, and conventional mixing; it illustrates how these choices translate lyrical motifs into perceptible spatial events associated with presence, externalization, and co-immersion; and it showcases the binaural release of the single *Mantra*, with its interactive interface, as evidence of knowledge transfer and creative innovation in contemporary music production.

Beyond the popular-music case, the proposed binaural production framework speaks directly to the journal's scope on Digital and Multimedia Technologies for Cultural Heritage fruition. Its transferable rules—anchoring, periphery, and height—together with the documented workflow (time charts, scene diagrams, parameter logs) can inform headphone-first and in-situ immersive sound design for museums, archaeological sites, and digital art installations. The same pipeline enables reproducible, low-footprint deployments (web interfaces, fixed-binaural assets) that enhance visitor engagement while preserving timbral coherence and narrative clarity across heterogeneous playback contexts.

2. Literature Review

Immersive music production rests on the capacity of three-dimensional (3D) audio to recreate source origin, depth, and motion around the listener. In this field, binaural audio is the dominant distribution format. It packs into two channels the key interaural cues—ITD and ILD—and HRTF-based spectral filtering that the auditory system uses to estimate azimuth (left–right), elevation (up–down), and distance (front–back/depth) (Blauert, 1997). When reproduced over headphones, these cues create the illusion of externalization—sounds perceived outside the head—supporting motion along three spatial axes and fostering presence and envelopment (Best et al., 2020). Spatial audio can increase engagement and the sense of being there, evidenced by faster orienting responses and reduced head motion, markers of more focused attention (Warp et al., 2022, as cited in Alonso Cardona-Cano, Calle, & López Díez, 2024).

This psychoacoustic framework aligns with spatial-production taxonomies. In Ambisonics, the Cartesian axes of B-format formalize the scene: X (front–back), Y (left–right), and Z (up–down) (Gerzon, 1973). This coordinate system provides an operational basis for designing and rendering coherent 3D sound images. Recent literature links this technical infrastructure to sound-experience design principles (Roginska & Geluso, 2017). In that view, the acoustic scene is not mere accompaniment; it becomes a substantive part of musical discourse and its spatial dramaturgy.

Recent work extends beyond technical proofs of concept to broader questions about mixing practices, production methodologies, and listener perceptual responses in commercial contexts (Dewey, Moore & Lee, 2024). The studies reviewed here are organised into four axes: (1) disciplinary landscape and emerging questions; (2) empirical findings and mix lessons; (3) technological innovation; and (4) contemporary production methodologies.

2.1 Binaural musical production: disciplines, subjects and recent questions.

Binaural music production remains rooted in sound engineering and signal processing, yet it now intersects productively with musicology, interaction design, and user-experience studies. At this intersection, Dewey et al. (2024) examine the adoption of Dolby Atmos and binaural renderers from the perspective of popular-music mixers. They highlight tensions between stereo tradition—especially phantom-center

conventions—and opportunities to exploit genuinely three-dimensional space. In parallel, Grundhuber and Lovedee-Turner (2024) propose an end-to-end neural upmix pipeline that transforms stereo mixes into binaural releases. Their approach opens a dialogue between AI methods and mixes aesthetics. From an ethnomusicological perspective, Paik et al. (2024) combine Korean traditional instruments with fifth-order Ambisonics to observe how spatiality reshapes timbral perception and narrative reading. At the systems level, Howie, Kamekawa, and Morinaga (2023) explore the use of bottom channels in 9+10+8 and 22.2 formats—areas with less development. Their capture and mixing choices can strain subsequent binaural translation. In performance contexts, Michael (2024) shows that octophonic spatialization and performer gesture co-determine live immersion.

Three operational questions emerge for this artistic research project. First, which elements should remain centre-anchored—i.e., minimally or not binauralised—to preserve perceptual familiarity? Practice points out: lead vocal, bass, and rhythmic axes as primary candidates, given their impact and transient sharpness (Dewey et al., 2024). Second, how timbral fidelity and localization precision can be balanced within a manual mixing workflow using commercial binaural renderers? Although the review includes neural-upmix proposals, the piece was conceived as streaming-first and produced through manual mixing with binaural headphone renders to ensure translatability across lossy codecs and typical headphone listening (Dewey et al., 2024). Findings by Grundhuber and Lovedee-Turner (2024) indicate that automated spatialization may gain positional accuracy at the cost of high-frequency loss; by analogy, processing choices here prioritized spatial clarity without sacrificing brightness and timbral detail. Third, how can spatialization function as a narrative device—assigning roles, trajectories, and contrasts that reinforce the dramatic arc? Literature on sound narrative suggests that spatial design guides attention, encodes recurrent motifs (a "leit-space"), and modulates emotional activation without visual support (Collins & Dockwray, 2018; Kerins, 2011; Wingstedt, Brandström, & Berg, 2010), and that head-tracked spatial reproduction measurably increases arousal while modulating affective tone by content type (Alonso Cardona-Cano et al., 2024). This final question organizes the design and steers the production decisions for the piece.

2.2 Recent findings and practical lesson for binaural mixing

The most consistent finding is the persistence of the stereo paradigm in critical mixing decisions. In popular music, lead vocals, bass, and drums are typically kept on the center axis to preserve impact and transient sharpness while ensuring translation on small loudspeakers and conventional headphones (Dewey et al., 2024). Dewey et al. (2024) characterize this pattern as a *stereo-plus* approach: expanding the three-dimensional field without dismantling the stereo reference that anchors listening.

Within mixing environments, engineers report dissatisfaction with commercial renderers (Dolby, Apple) owing to limited externalization and restricted control over critical binaural parameters. This lack of fine-tuning is a recurrent concern among users (Dewey et al., 2024). As well, machine-learning-based approaches present benefits and risks: neural binaural upmixers can position up to four sources with a margin azimuth error of ~11.3° and improve spatial clarity for trained listeners. Some implementations, however, attenuate high frequencies, potentially dulling brightness and air if not carefully monitored (Grundhuber & Lovedee-Turner, 2024).

When production originates in high-order Ambisonics, fixed-binaural versions outperform stereo in spatial mapping and perceived depth. Beyond rendering choices, pre-production narrative planning—and a functional classification of sources as *musical*, *spatial*, or *movement effects*—emerges as a good practice, improving transparency, selectivity, and timbral legibility (Paik, Han, Kim, & Lee, 2024). In multichannel systems with bottom layers, placing loudspeakers below the listener's horizontal plane extends the vertical dimension. When these channels carry direct sound and early reflections, listeners orient more easily and report greater presence and realism; overly diffuse content, by contrast, can blur spatial definition and reduce overall mix clarity (Howie, Kamekawa, & Morinaga, 2023, p. 11).

The literature converges on a selective spatial principle: (1) centralize reference sources—lead vocal, bass, kick and snare—to preserve impact

and playback compatibility (Dewey et al., 2024); (2) reserve the 3D field for atmospheres, textures, and secondary elements that benefit from spatial placement and motion (Paik et al., 2024); and (3) maintain spectral integrity when applying binauralization or automated upmixing so that brightness, articulation, and air are not lost in the final mix (Grundhuber & Lovedee-Turner, 2024; Howie et al., 2023). This selective approach sustains perceived immersion while preserving musical readability, ultimately anchoring the work's spatial narrative.

2.3 Technologies and features innovation

To organize the state of the art in binaural music production, recent literature converges on two methodological families with complementary logics. On one side, data-driven/AI approaches propose end-to-end upmix pipelines: starting from paired stereo/binaural examples (often created via HRTF convolution), HDemucs-derived models perform source separation and spatial placement in a single step and are optimized with compound losses that balance positional accuracy and timbral preservation; evaluation typically combines objective metrics (e.g., azimuth error) with listening tests involving expert and non-expert participants (Grundhuber & Lovedee-Turner, 2024). This route accelerates the conversion of stereo catalogues to binaural, yet it demands close monitoring of quality trade-offs (e.g., high-frequency content) and repertoire-specific perceptual validation criteria, and it remains largely agnostic to the musical narrative.

On the other side, higher-order Ambisonics (HOA) framework encompasses fifth-order capture or mock-ups and B-format mixing. It also organizes stem classification by function and intended binaural render (e.g., in-ear monitors, IEM), yielding a high-resolution spatial-audio format compatible with headphone delivery. Studies report a preference for binaural over stereo on spatial mapping, depth, and selectivity (Paik et al., 2024; Malecki et al., 2020). This advantage is strongest when space is treated as a compositional and narrative parameter from pre-production. Additionally, 3D multichannel layouts with lower (bottom) channels guide microphone placement and balance, allocating direct sound and early reflections transferable to binaural renders (Howie, Kamekawa, & Morinaga, 2023). Performance-oriented studies emphasize coherence between spatial diffusion and stage action (Michael, 2024). Some proposals include head-tracking; this lies outside the scope of the present study.

Practical protocols for 3D multichannel production complement these approaches (Howie et al., 2023). Performance-oriented guidance privileges the interplay between physical action and spatial diffusion (Michael, 2024).

2.4 Contemporary Innovations in Immersive Music Production

Recent innovations in audio technology have significantly expanded the toolkit for immersive music production. Object-based audio formats—such as Dolby Atmos, MPEG-H 3D Audio, and Sony 360 Reality Audio—allow engineers to position musical elements as spatial objects within a 360° virtual field. These formats are inherently scalable: a single immersive mix can be deployed over multichannel loudspeaker arrays or binaurally rendered for headphone playback, adapting to the listener's setup (Gould, as cited in Kopp, 2022). Originally introduced in cinema, Dolby Atmos is now standard in music studios, with hundreds worldwide producing albums and streaming releases in immersive audio. Together with the adoption by major platforms (Apple Music, Tidal, Amazon Music, Deezer), these technologies signal a new phase in mainstream music distribution (Kopp, 2022).

Beyond formats, hardware and software have improved playback: modern headphones with integrated gyroscopes support real-time head-tracking, ensuring the virtual soundstage remains externally anchored as the listener moves. This enhances spatial stability and presence (Kelly, Woszczyk, & King, 2020). Producers are also experimenting with hybrid workflows that merge ambisonic soundfield recordings with close-miked stems, subsequently rendered through binaural panning. Such workflows balance natural spatial cues with precise object-based control, enriching the immersive palette.

The case study of *Mantra* (Estados Alterados, 2018) exemplifies this trend. The project—developed in collaboration with the Universidad de Medellín and the Universidad de San Buenaventura—applied a hybrid binaural methodology to design a fully enveloping sound environment. The workflow integrated 3D sound design tools and psychoacoustic techniques

(ITD/ILD-based panning, custom HRTFs, immersive reverbs) to evoke immersion, presence, and co-presence.

2.5 Identified gap and synthesis

The reviewed frameworks converge on a delicate balance: expanding the stereo paradigm without eroding its perceptual heritage, while improving control over binaural rendering and validating the end-listener experience. The agenda is thus broader than technology; it spans perceptual and editorial concerns and demands explicit accounting of trade-offs—what is gained and what is surrendered—when 3D space interacts with narrative conventions, intelligibility, and cross-device translation.

Within this framework, several gaps emerge: (1) a need for more artistic–technical case studies that document, step by step, creative and mixing decisions in real popular-music projects, with a level of detail that enables auditability and transferability; (2) a lack of comparative evaluations that integrate sound narrative—scene design, spatial functions, and narrative progression—together with common technical constraints (translatability to consumer headphones, stereo compatibility, and spectral losses introduced by processing) under a unified production protocol; and (3) a need for replicable methodologies that integrate pre-production, capture, and binaural rendering with clear criteria for timbral control, dynamic balance, and listening tests.

As an outcome, *Mantra—Immersive Experience* (Estados Alterados, 2023) presents a case study integrating experiential design, binaural capture, and creative mixing in collaboration with its musicians. It proposes a replicable methodological guideline and contributes with perceptual data to academic and professional discourse on headphone-oriented immersive music, with sound narrative as the organizing principle for design and production.

3. Materials and Methods: Research-Creation

3.1 Research Design

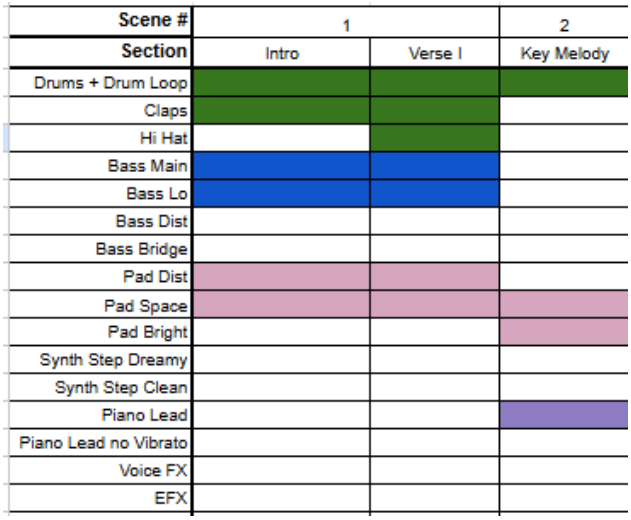
The project was developed as a fixed-binaural remix for headphone delivery using the original stems and materials from *Mantra* (Estados Alterados, 2018). Our approach comprised three strands: (i) applying binaural synthesis directly to the original stems; (ii) mono and binaural re-amping of selected instruments and stems to enhance externalization and acoustic depth; and (iii) targeted overdubs—including a female vocal—to enrich texture and narrative meaning. Artistic planning was organized around the conceptual opposition Reality ↔ Beyond, grounded in literature on presence, externalization, and co-immersion, a dialectical framework that balances perceptual stability with transcendence in immersive listening. This opposition was operationalized as a set of spatial rules for immersive sound design (e.g., Dolby Atmos): anchoring, periphery, and height within the spatial scene.

3.2 Spatial workflow

The project was conducted in collaboration with co-investigator Juan Pablo Hormiga Leal (Universidad de San Buenaventura, Medellín). The source material comprised the band's instrumental and vocal stems, organized in a DAW session with tempo and lyric markers. To support morphological and narrative analysis—and to guide spatial design prior to production and mixing—we prepared detailed time charts, complemented by scene diagrams specifying azimuth, elevation, and distance for each source. An inventory of available plugins and devices accompanied this preparatory phase. Finally, binaural re-amp captures were obtained using the patented binaural-capture device described by Cardona Cano, Moreno Viasus, & Tafur Jiménez (2022).

Morphology (Figure 1). *Mantra* (Estados Alterados, 2018) is a trance-like piece—a mantra built on themes of human uncertainty and alienation—whose hypnotic repetition and ascent/escape imagery (e.g., "llévame al sol," "donde existe el mar") frame the design. It comprises nine sections defined purely by musical criteria; the lyrics provide no explicit formal divisions. The sequence unfolds as:
Introduction (a), Verse 1 (A), Electric-piano melody (B), Verse 2 (A), Electric-piano melody + guitar (B′), Interlude (C), Mantra (D), Mantra + electric-piano melody + guitar (B″), Mantra (voices only) (D′)

This corresponds to the pattern a A B A B′ C D B″ D′. Section D functions as a turning point: a female voice enters, supported by a chorus repeating verse 26, and remains present through B″ and D′ until the close.



Scene #	1		2
Section	Intro	Verse I	Key Melody
Drums + Drum Loop			
Claps			
Hi Hat			
Bass Main			
Bass Lo			
Bass Dist			
Bass Bridge			
Pad Dist			
Pad Space			
Pad Bright			
Synth Step Dreamy			
Synth Step Clean			
Piano Lead			
Piano Lead no Vibrato			
Voice FX			
EFX			

Fig. 1: Time-chart fragment (screen capture) showing the distribution of sound elements in the original song

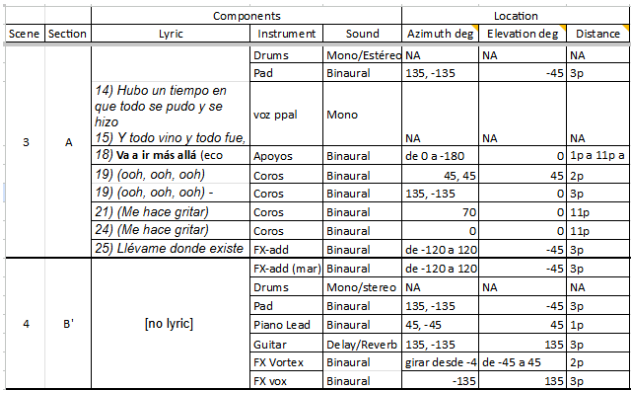
3.3 Formal planning and spatial narrative

It was established a section-level morphology and lyric-derived set-pieces (Figure 2). The time chart marked zones of semantic transition—for example, references to "sun," "sea," and "desire"—that trigger spatial events such as motion, plane shifts, and proximity changes. Scene maps specified target azimuth, elevation, and distance for each cue, linking every lyric trigger to a concrete spatial gesture. These mappings were consolidated in a technical table (positions, angles, levels) to ensure replication and traceability across takes and automation. Each source family was assigned a spatial function:

(1) Anchors (lead vocal, bass, rhythmic axes) ensure stability and legibility within the *Reality* dimension; prioritizing impact clarity and tonal balance, adding micro-variations of proximity to preserve the perceptual anchor.

(2) Periphery (pads, textures, backing vocals, sequences) shapes the enveloping field, extending lateral spread and depth; in passages oriented toward *Beyond*, trajectories and externalization expansions used to gently destabilize listener orientation and evoke a sense of transcendent breadth.

(3) Height (effects, reverberation, processed voices) introduces vertical layers above the horizontal plane, producing perceptual elevation; this dimension operates as a symbolic vector toward the transcendent and emphasizes narrative climaxes.



		Components			Location		
Scene	Section	Lyric	Instrument	Sound	Azimuth deg	Elevation deg	Distance
3	A		Drums	Mono/Estéreo	NA	NA	NA
			Pad	Binaural	135, -135	-45	3p
		14) Hubo un tiempo en que todo se pudo y se hizo *15) Y todo vino y todo fue,*	voz ppal	Mono	NA	NA	NA
		18) **Va a ir más allá** (eco	Apoyos	Binaural	de 0 a -180	0	1p a 11p a
		19) (ooh, ooh, ooh)	Coros	Binaural	45, 45	45	2p
		19) (ooh, ooh, ooh) -	Coros	Binaural	135, -135	0	3p
		21) (Me hace gritar)	Coros	Binaural	70	0	11p
		24) (Me hace gritar)	Coros	Binaural	0	0	11p
		25) Llévame donde existe	FX-add	Binaural	de -120 a 120	-45	3p
4	B'	[no lyric]	FX-add (mar)	Binaural	de -120 a 120	-45	3p
			Drums	Mono/stereo	NA	NA	NA
			Pad	Binaural	135, -135	-45	3p
			Piano Lead	Binaural	45, -45	45	1p
			Guitar	Delay/Reverb	135, -135	135	3p
			FX Vortex	Binaural	girar desde -4	de -45 a 45	2p
			FX vox	Binaural	-135	135	3p

Fig. 2: Time-chart screen capture showing the spatial distribution of sources}

3.4 Session preparation and operational criteria

The session used bus routing to split processing into two lines: Line A (mono/binaural re-amping) and Line B (HRTF synthesis and panning), converging at a final mix bus. Before any spatialization, de-noise and de-click cleaning, phase alignment on parallel material and gain normalization were applied to preserve useful dynamic range during binaural panning. Stems were classified as anchors, periphery, height, or overdubs; this classification determined their spatial function and processing.

Scene maps guided placement and trajectories: Fig. 3 (plan view) and Fig. 4 (elevation view) show the layout by planes (front–side–rear–height), movement paths (arrows), and element labels (e.g., B, D, PL, FXadd, CR/CL). For abbreviations, see the code definitions in Table 2, used throughout the session and in automation.

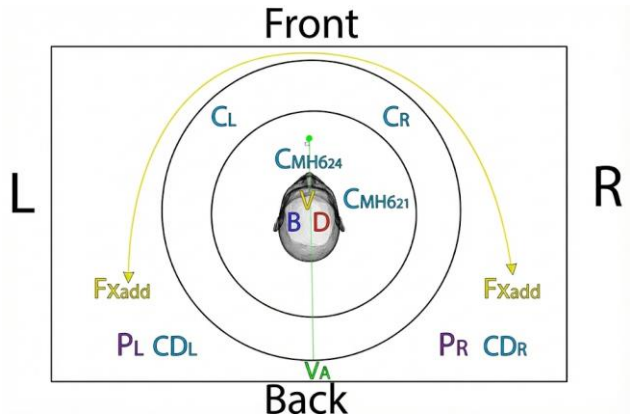


Fig. 3: Screen capture showing example diagrams of trajectory and proximity automation (plan view)

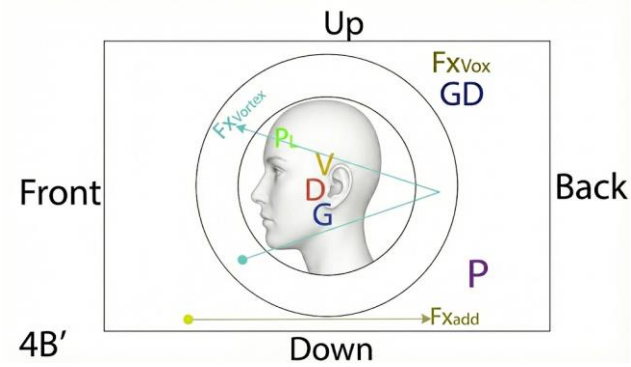


Fig. 4: Screen capture showing example diagrams of trajectory and proximity automation (elevation view)

Table 1 lists the codes used in the diagrams and the session for instruments/elements and supporting layers (e.g., B—bass; D—drums; PL—piano lead; FXadd—added effects; CR/CL—chorus L/R; CD—chorus delay; GD—guitar delay; VF—female voice). This coding enables each action—re-amp, panning, and automation—to be traced from the scene map to the corresponding take or bus, and it was used during headphone translation tests to validate plane coherence (front/side/rear/height) and intelligibility. The same labels appear consistently in the scene diagrams, track names, and automation lanes, improving readability during editing and export. Maintaining a single legend across artefacts reduces annotation errors and speeds cross-checks between design and session. In practice, the scheme supports auditability and straightforward replication of the mix decisions.

Tab. 1: Coding of sound elements and instruments across the song's morphological sections

Cod	Instrument / Element
B	Bass
D	Drums
P	Pad
Fxadd	Fxadd
C	Chorus
A	Backing element
CD	Chorus Delay
V	Main Vox
PL	Piano Lead
CMH6	Chorus MH6 Line
FXVortex	FXVortex
G	Guitar
GD	Guitar Delay
Fx Vox	Fx Vox
FXS	FX sweep
S	Secuence
ScL	Secuence Crescendo
V	Vocoder
VF	Female vox

Line A: mono and binaural re-amping. To externalize sources and enrich room cues, key stems and instruments were re-amplified with amplifiers/monitoring systems placed in front of a binaural head. Mono passes (point source) and binaural passes (nearfield/ambient) were executed, varying distance and angle according to the arcs and positions defined in the scene maps (Figs. 3–4). Each take was logged (position, level, observations). When comb filtering was detected, mitigation combined micro-displacements of the source/head (≈1–3 cm) with narrow-Q notch cuts at the resonant peaks identified by sweeps (typically −3 to −6 dB, Q≈8–12 in the 2–8 kHz region), avoiding attempts to "fill" spectral nulls. Where re-amping produced brightness loss on cymbals and sibilants, a gentle high-shelf on the return bus restored air (+2 to +4 dB above 9–10 kHz), while occasional broad bell attenuation around 2.7–3.2 kHz (−2 to −3 dB, Q≈0.7–1.2) reduced harshness before binaural rendering. Capture employed the patented binaural device US11445298B2 (Cardona Cano, Moreno Viasus, & Tafur Jiménez, 2022).

Figures 5. These figures illustrate vocal-capture setups: (a) rear-lateral placement relative to the binaural head, useful for contrasting the lead vocal and staging entries from the rear field; and (b) frontal placement when axial anchoring and presence were required. These captures were used for overdubs (e.g., VF, female voice) and for re-amping of scene-prioritized elements. The placements mirror the arcs in the scene maps, maintaining consistency with time-chart cues. Positions and levels were logged for session traceability. Labeling matches Table 2 so editing and export remain coherent across artifacts. In both setups the performer used closed-back headphones and maintained a working distance (~20–30 cm) to the binaural head to control leakage and proximity. Height and yaw were adjusted so the mouth aligned with the pinnae plane, stabilizing elevation cues. The studio's absorptive slat treatment reduced early reflections in the capture area, improving externalization consistency.

Fig. 5: Vocal capture setups with a binaural head—(a) rear-lateral and (b) frontal—using auxiliary devices Zoom H3-VR (Ambisonics) and Sennheiser AMBEO Smart Headset.

Line B: HRTF synthesis and panning. Also, selected stems were binauralized via HRTF with both static and automated panning of azimuth and elevation. Distance was modeled by combining early delays, coherent early reflections, and low-density late reverberation, while preserving intelligibility in the lead vocal and rhythmic anchors. Per-source parameters—target position, predelay, early/late balance, and air-band filtering—were taken from the design maps (Figs. 3–4) and consolidated in an internal technical table for traceability.

3.5 Scene assembly and movement logic

The assembly integrated Line A and Line B materials by the time-chart sections. Movements—lateral sweeps, progressive lifts, and rear-field entries—were reserved for semantically marked lyric points and executed along the arcs indicated in Figs. 3–4, avoiding continuous trajectories as mere ornamentation. The female voice (VF) followed targeted paths that interact with the lead vocal, maintaining timbral contrast and avoiding overlap in the presence band. For visual examples of trajectory and proximity automation in the DAW session, see Fig. 4 (curve captures), which shows the relationship between cue points, target positions, and micro-level adjustments used to stabilize perception.

3.6 Mixing and spatial selectivity control

The mix adopted a stereo-plus anchoring strategy: the lead vocal, bass, and rhythmic axes were kept center-focused with controlled proximity, while the periphery occupied height and depth. To maintain selectivity, localized dynamic EQ and creative filtering were applied to moving sources, and parallel compression on buses preserved microtransients. Timbral coherence was verified iteratively after each re-amp or HRTF stage. When comb filtering appeared, mitigation combined micro-displacements (≈1–3 cm) with narrow-Q notch cuts at detected resonant peaks (typically −3 to −6 dB, Q≈8–12 in the 2–8 kHz region), avoiding attempts to boost spectral nulls. For brightness loss from extreme panning/spatialization, a gentle high-shelf on the affected return or object restored air (+2 to +4 dB above 9–10 kHz). To reduce harshness in dense harmonic material before binaural rendering, a broad bell around 2.7–3.2 kHz (−2 to −3 dB, Q≈0.7–1.2) was used; when width needed reinforcement without center smear, M/S processing applied a side high-shelf (+1 to +2 dB >7 kHz) and a side high-pass (up to ~150 Hz).

3.7 Quality control and export

Validation of the binaural mixed track included listening to it across multiple commonly commercial headphones, basic stereo fold-down review and mono checks to minimize artifacts. The final delivery comprised a binaural master and reference derivatives, with levels compatible with streaming distribution. Sessions and binauralized stems were archived for traceability.

3.8 Documentation and traceability

A consolidated package was assembled: the final time chart, per-scene spatial diagrams, automation screenshots, re-amp reports (positions, angles, levels), and an inventory of plugins/devices. This material supports workflow replicability and informs a critical reading of mixing decisions. Table 2 details the procedures and coding scheme applied across production and post-production of the recording.

Tab. 2: Technical and creative procedures applied at Mantra binaural re-mixing

Procedures	Description
Re-amp	Re-records tracks through amplifiers/effect chains to capture additional tone and spatial cues, using a binaural head for nearfield and ambient takes.
Overdub	Adds new layers of voices, instruments, or

	effects to enrich texture and narrative nuance.
Stem mixing	Mixes subgroups (drums, bass, vocals, etc.) to process and balance complete blocks in a controlled manner.
Sampling / resampling	Extracts and reuses treated audio fragments as rhythmic or ambient elements.
Time-stretching	Adjusts duration without altering pitch to reinforce transitions or emphasize atmospheres.
Pitch-shifting / transposition	Changes pitch to create harmonies and melodic contrasts.
Beat slicing & rearrangement	Segments and reorganizes rhythmic patterns to generate variations and unusual percussive effects.
Layering	Superimposes takes or sounds to obtain more complex, consistent timbres.
Creative filters / dynamic EQ	Applies filter sweeps and dynamic equalization to shape spectral perception across the piece.
Parallel dynamics processing	Blends compressed/saturated signal with dry signal to retain detail and increase impact.
Re-synthesis / granular	Decomposes and reconstructs audio spectrally into new textures, used in ambient passages.
Spatial processing	Redesigns the acoustic scene with reverbs, delays, and automated panning to place elements on different planes.
Reverse / backmasking	Uses reversed fragments for transitions and psychedelic effec

3.9 Computational analysis with MusiCNN

The two versions of *Mantra*—stereo (MANTRA_ST_44.1_16.wav) and binaural (MANTRA_BI_44.1_16.wav)—were processed with Essentia (Bogdanov et al., 2013) using the pretrained MusiCNN models (Pons & Serra, 2019) integrated via its TensorFlow framework (Alonso-Jiménez, Bogdanov, Pons, & Serra, 2020). In popular-music analytics, spectrum-derived features and large datasets have proven effective for mapping stylistic and popularity trends, as shown in a recent SCIRES-IT study of reggaeton using audio-spectrum metrics and correlation analyses (Arango-Lopera, Escobar-Sierra, & Cardona-Cano, 2024).

Audio at 44.1 kHz/16-bit was segmented into 3-second windows with 50% overlap (hop ≈ 1.5 s), producing frame-level predictions aggregated to track level by mean, median, and interquartile range (IQR); the mean is reported as the primary value, with the other metrics retained as controls. The extractor also computed basic descriptors (energy_rms, loudness, bpm, beats_confidence) and estimated 0–100 probabilities for the MusiCNN axes: danceable, happy, relaxed, aggressive, party, acoustic, electronic, instrumental/voice, and gender. When required by a given model, Essentia resampled internally; default extractor settings were preserved (e.g., mel filter bank and the models' internal normalizations).

To prevent level differences from explaining mood changes by themselves, the analysis was executed under two conditions: RAW (files as delivered) and LUFS-normalized to −14 LUFS-I (ITU-R BS.1770/EBU R128), applying the same gain offset to the entire file prior to feature extraction. Result tables and figures indicate, where relevant, the condition used.

Interpretation did not rely on numbers alone: each difference was weighed against the mix's technical and narrative context so that mood shifts could be reasonably attributed to format (stereo vs. binaural) and acoustic-scene decisions, rather than to level or compression alone.

Beyond simple level effects, the direction and rank-ordering of key MusiCNN tags remained stable when moving from RAW to −14 LUFS-I normalization. In both conditions, the binaural render shows a coherent shift relative to stereo—higher probabilities on relaxed, voice, and acoustic axes, and lower on party, danceable, and aggressive. Normalization attenuates absolute magnitudes (as expected) but preserves the between-version contrasts, indicating that the affective redistribution reflects spatial-scene decisions (externalisation, depth, elevation) rather than level or compression alone.

4. Results

4.1 Binaural Track Outcome

Mantra—Immersive Experience (Estados Alterados, 2023) is a three-dimensional reinterpretation of the song "Mantra" (Estados Alterados, 2018), intended for headphone listening and optimized to elicit immersive spatial perception. The production draws on psychoacoustic principles of externalization and spatial positioning that allow listeners to perceive sources within an enveloping physical space beyond the traditional stereophonic plane (Blauert, 1997).

This version preserves the original work's formal structure and core musical elements, while reorganizing placement, depth, and motion within a binaural field. Its spatialization criteria align with proposals by Warp et al. (2022, as cited in Alonso Cardona-Cano et al., 2024) for creating three-dimensional soundscapes in immersive musical experiences. The distribution of voices, instruments, and textures establishes spatial contrasts that reinforce the lyrical opposition Reality ↔ Beyond, consistent with Fontana et al. (2007) account of space as a narrative device in music production.

The 3′46″ recording is available on streaming platforms (see Figure 7)—including Spotify, Deezer, YouTube Music, and Apple Music—and on an interactive microsite on the artist's official website (Estados Alterados, 2022). It constitutes an artistic research output in the sense outlined by Alonso Cardona-Cano et al. (2024, citing Warp et al, 2022), that technical and aesthetic experimentation leads to applied knowledge about binaural production in popular-music contexts. The result is both a finished sound artefact and a proof of concept that integrates theory, experimentation, and professional practice.

Technical delivery sheet. Duration: 3:46. Master: 48 kHz / 24-bit, −14 LUFS-I (integrated), −1 dBTP (true peak). Release date: November 9, 2023.

Credits. Composition: Fernando Sierra Rodríguez, Felipe Carmona Montoya, Natalia Valencia Zuluaga, Ricardo Restrepo Guzmán, and Amir Derakh. Production: Amir Derakh. Binaural remixing / re-amp / overdubs / mixing / mastering: César Cardona-Cano, Juan Pablo Hormiga-Leal, and Felipe Carmona-Montoya.

Collaborative production methodology. The work involved coordinated collaboration between the investigator-producers, producer Felipe Carmona, and the band, following a linear sequence in which each stage informed the next, with targeted creative iterations to refine timbre, balance, and spatialization. Key challenges included preserving the original timbral character under intensive spatial processing and achieving smooth integration among re-amped, synthesized, and newly recorded layers.

The goal was to retain the piece's aesthetic identity, enhance three-dimensional auditory depth, and explore new possibilities for immersive sound.

Principal constraints concerned the coherent integration of binaural-head captures with HRTF-synthesized elements and the balance between timbral density and spatial clarity in a multichannel environment ultimately rendered for binaural headphone playback.

Mixing criteria — stereo-plus anchoring. The scene was organized into three operational zones. *(1) Anchor:* bass and kick/snare were kept centered (phantom center) and frontmost to ensure impact, transient definition, and reliable translation on headphones and small loudspeakers—consistent with practices reported by Dewey et al. (2024). The lead vocal remained the narrative anchor in the foreground, with controlled lateral excursions and depth automation as expressive cues for the dramatic arc. *(2) Periphery:* backing vocals, textures, and ambiences were arranged along lateral and rear arcs via binaural spatialization to promote externalization and envelopment, supporting presence and co-presence between virtual sources and real captures (Reardon et al., 2018; Fantini et al., 2023). *(3) Height:* selected spectral elements and reverberation were lifted using binaural panning with elevation cues and decorrelated early reflections, integrating the vertical axis described in with-height reproduction and HOA workflows (without head tracking), in line with Gerzon (1973) and recent Ambisonics mixing practice (Paik, Han, Kim, & Lee, 2024). This anchor/periphery/height allocation preserved stereo-plus intelligibility while materializing the Reality ↔ Beyond narrative through contrasts of depth, lateral spread, and elevation.

Spatial selectivity. Pads, backing vocals, and effects occupied periphery and depth to provide envelopment and narrative "leit-spaces," while avoiding brightness loss associated with certain

automated processes (Grundhuber & Lovedee-Turner, 2024).

Timbral coherence. Spectral balance was monitored after each binaural stage—capture and/or synthesis—to sustain clarity, air, and transient legibility.

Narrative transfer into spatial decisions — Reality vs. Beyond. Spatial choices materialize the Reality/Beyond opposition through continuous sound dramaturgy. As an example: kick, snare, and bass were axis-anchored. In *"llévame al sol"* (~01:00), pads were raised 20–30° and azimuth widened to ±60°, with short early reflections suggesting ascent. In *"donde existe el mar,"* slow diffuse sounds were densified in the rear plane to stage a gradual transition. This differential use of the sound space increases presence—the subjective sense of "being inside" an auditory environment (Fantini et al., 2023)—and leverage externalization, whereby sources are perceived outside the head, deepening immersion (Reardon et al., 2018). Alternation and contrast between spatial planes guide attention, structure the narrative, and amplify expressive impact.

Marker moments. In *"llévame al sol"* (~01:04–01:05, original version), harmonic-spatial flashes (pads/FX) are emphasized via elevation and depth; in *"llévame donde existe el mar,"* a slow, diffuse landscape suggests immersion and transfer.

Trajectories and roles. Male backing vocals as "inner voices" were distributed along a low-density lateral–rear arc; vocoder textures/pads occupied high and rear layers to reinforce dramatic progression and formal pacing.

Compatibility and distribution. The master was created for streaming platforms and parameters of loudness and true peak were followed, ensuring its translation on consumer headphones and stereo fold-down without evident phase artefacts.

It was released on Spotify/Apple Music and other platforms as an independent single, supporting technology transfer across academia, the band, and industry.

Known limitations. As a fixed-binaural master, the scene is head-locked (no dynamic re-referencing). By design, the production prioritized narrative legibility and timbral fidelity over continuous hyper-movement. Externalization and affect may vary across HRTFs and headphones, limiting generalization across devices and listener anatomies.

4.2 Mood topology

Data and reading criteria. Two versions—stereo (MANTRA_ST_44.1_16.wav) and binaural (MANTRA_BI_44.1_16.wav)—were analyzed with Essentia/MusiCNN in Python (Python Software Foundation, 2023) at 44.1 kHz/16-bit. The pipeline produced basic descriptors (energy_rms, loudness, bpm, beats_confidence) and 0–100 probabilities for semantic axes (danceable/not, happy/non, relaxed/non, aggressive/not, party/non, acoustic/non, electronic/non, instrumental/voice, gender). Values were aggregated by track level and therefore characterize each mix's global profile. Interpretation weights results by the technical and narrative context of each version (stereo vs. binaural with HRTF synthesis, re-amps, and overdubs).

Basic descriptors. The binaural profile shows lower global energy (energy_rms 0.28 → 0.12, −55%) and reduced loudness (2302.6 → 778.9, −66%), with tempo effectively unchanged (161.6 → 160.1) and beats_confidence increased (+14%). This pattern is consistent with a deeper, externalized scene: less overall push (RMS/loudness) and greater local rhythmic legibility due to spatial separation.

Z-score heatmap. The z-score heatmap (Figure 6) indicates opposed profiles. Stereo sits above the set mean for bpm, energy_rms, loudness, and duration, whereas binaural sits below on those descriptors. The exception is beats_confidence, where binaural exceeds the mean and stereo falls below. Practically, the stereo version trends faster, louder, and slightly longer; the binaural version exhibits a more stable perceived pulse. This contrast suggests that binaural processing softened global energy and perceived loudness while consolidating rhythmic anchoring—shifting the experience toward groove focus over sheer impact. Notably, the pattern is antiphasic across descriptors: whenever Stereo scores +1, Binaural scores −1, with beats_confidence inverting that relation. Because values are standardised (z-scores), the plot emphasises relative displacement from the dataset mean rather than absolute magnitudes, highlighting a mirror-profile rather than small random fluctuations. The symmetric colour scale (±1) makes this opposition immediately legible and consistent across all displayed descriptors.

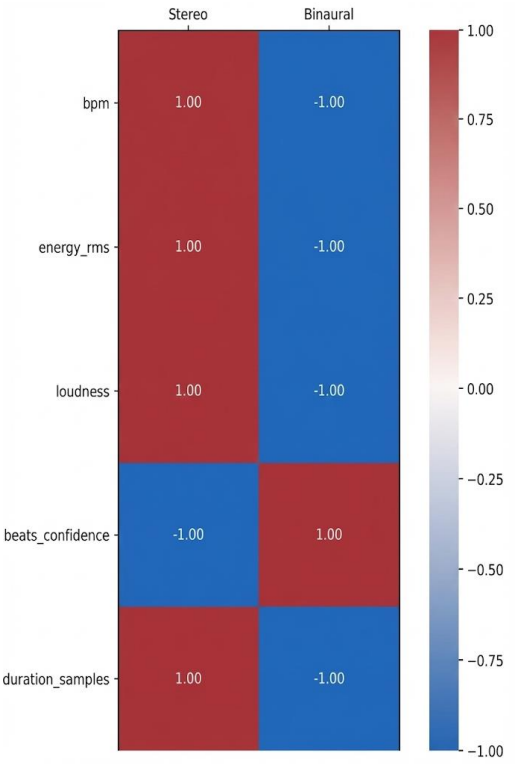


Fig. 6: Heatmap (z-scores) of basic descriptors—energy_rms, loudness, bpm, and beats_confidence—for the stereo and binaural versions

Mood and timbre variables — salient changes (Binaural – Stereo). The lollipop plot in Figure 7 shows a coherent affective shift from stereo to binaural: increases in instrumental_voice and relaxed_relaxed, and decreases in instrumental_instrumental, party_party, danceability_danceable, and aggressive_aggressive. Electronic_non_electronic also increases (with a smaller rise in acoustic), indicating that externalization achieved via HRTF synthesis and re-amps introduces room cues and separation that the model reads as less electronic and more organic. Reinforced vocals—enhanced by the female overdub and directional focusing—act as a semantic anchor, shifting valence toward a calm, contemplative state at the expense of dance-oriented and front-loaded energy markers. This pattern aligns with the basic descriptors: although aggregate loudness decreases, beats_confidence rises, suggesting that spatial selectivity clarifies the pulse and improves temporal legibility. Overall, the binaural version prioritizes depth, vocal clarity, and an enveloping scene, reconfiguring mood interpretation without relying solely on level.

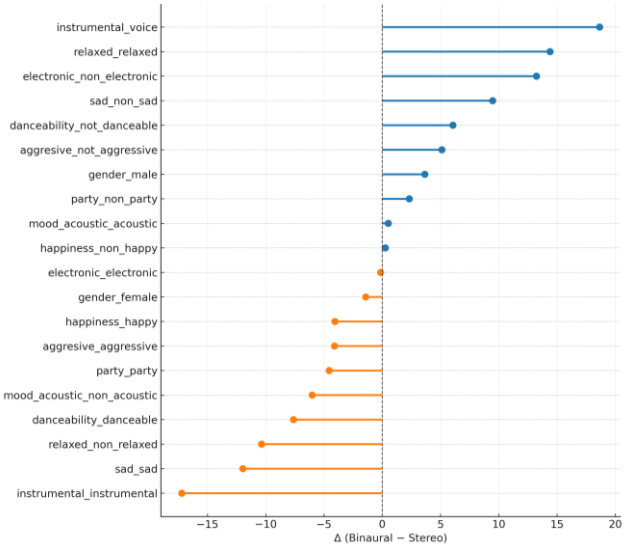


Fig. 7: Lollipop plot of differences (Binaural – Stereo) across MusiCNN mood variables

Compact profile: radar of positive classes. The positive-class radar in Figure 8 profiles a redistribution of probability mass across affective–timbral axes rather than isolated increases or decreases. In the binaural version, the polygon shifts and opens toward relaxed_relaxed and, to a lesser extent, acoustic, while contracting in party_party, danceability_danceable, and aggressive_aggressive. This geometry suggests an arousal–valence readjustment: lower behavioral activation and reduced timbral friction, with a more stable envelope less oriented toward festive drive. The contraction in instrumental_instrumental implies a re-hierarchisation of sources: textural backgrounds cede prominence to semantic carriers, concentrating attention on guide elements and reducing cross-plane interference. It is not merely "more relaxed"; the resulting contour redistributes attentional focus in the auditory field, shifting from motoric markers (party/danceable) toward sustained listening states. For additional quantification, polygon area and angular similarity between axes (e.g., relaxed vs. aggressive) could formalize this shape change into a comparable affective-profile index across mixes.

The centroid drifts toward the relaxed–acoustic quadrant, yielding a left-upper sector weighting and a smoother, less jagged outline. Axis coherence is visible: reductions on high-activation tags mirror the opening toward low-arousal descriptors, reinforcing a global—not isolated—redistribution.

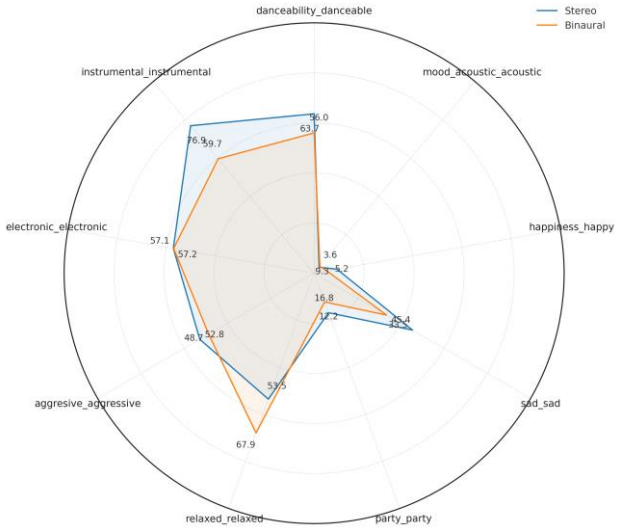


Fig. 8: Compact radar of positive classes (MusiCNN) comparing the stereo and binaural versions

5. Discussion

The study addresses a central question—how to translate *Mantra*'s lyrical dramaturgy into a reproducible three-dimensional scene for headphones—by showing that a hybrid workflow (dummy-head re-amps, HRTF-based synthesis, and multitrack mixing) appears to achieve presence, externalization, and co-immersion without sacrificing narrative legibility. A stereo-plus anchoring decision—lead vocal, bass, and rhythmic axes on the center line, with textures and backing vocals in the periphery—prioritized intelligibility and focus control. Motion was reserved for semantically marked turns in the time chart, and the female overdub served as a cue for attentional re-hierarchization rather than a blanket expansion of spatial extent.

These choices align with psychoacoustic of localization and externalization, where the combination of direct sound and room cues supports coherent elevation/azimuth and distance judgments (Blauert, 1997; Begault, 1994; Møller, 1992; Roginska & Geluso, 2017). At the experiential level, the case accords with presence/immersion models in which spatial design supports action rather than acting as permanent ornamentation (Slater & Wilbur, 1997; Reardon et al., 2018; Fantini et al., 2023). Computational analyses with MusiCNN point in the same direction: a shift toward relaxed and voice with reduced party/danceable, consistent with a deeper scene and a foreground vocal narrative.

A caveat is warranted: part of the marked drop in energy_rms and loudness arises from the spatialization chain. In stereo, center-panned material sums coherently and maximizes amplitude at the phantom center. In the binaural workflow, HRTF convolution and deliberate inter-channel decorrelation are employed to achieve externalization. The resulting spectral shaping (e.g., elevation notches and pinna-related peaks) and phase offsets disperse energy across the virtual field rather than concentrating on it. Consequently, the algorithm's lower energy readings reflect a technical trade-off: central density gives way to headphone-oriented depth, separation, and usable dynamic range.

The work proposes an operational, transferable pipeline between academia and industry, supported by technical documentation, spatial diagrams, and position reports that facilitate replication, together with a patented binaural-capture device for controlled re-amping (Cardona Cano et al., 2022). This traceability enables adaptation across repertoires and publication contexts.

Limitations include that the master does not include head-tracking and in the mood analysis uses track-level aggregation which does not estimate temporal variability. Future work includes windowed extraction, nonparametric contrasts, and level-normalization checks prior to comparison, as well as controlled listening tests to isolate spatial effects from those of mixing.

6. Conclusions

6.1 Alignment with Objectives and Research Questions

This study addressed two central aims: (i) to operationalize the perceptual constructs of *presence*, *externalization* and *co-immersion* into concrete production decisions, and (ii) to document a reproducible workflow capable of translating lyrical narrative into audible spatial events.

The analysis shows that presence was reinforced through anchor sources (lead voice, bass, rhythmic axis) whose spatial stability-maintained intelligibility and orientation. Externalization was achieved by combining binaural re-amping with HRTF synthesis, a strategy that preserved natural depth cues while allowing flexible manipulation of stems in the hybrid mix. Co-immersion emerged when

peripheral and height elements were integrated seamlessly with the anchors, blurring the perceptual boundaries between natural re-recordings and synthetic layers; listeners reported difficulty distinguishing these layers, which suggests that the hybrid approach succeeded in creating a coherent sound scene.

Computational analysis with Essentia/MusiCNN provided an additional layer of validation. By segmenting the stereo and binaural versions of *Mantra* into 3-second windows, the system quantified descriptors such as RMS energy, loudness and BPM, along with probabilities for affective and stylistic dimensions (e.g., danceable, relaxed, acoustic/electronic). The convergence between perceptual design and algorithmic outputs is notable: binaural mixes exhibited higher probabilities in categories associated with relaxation, immersion and acoustic depth, consistent with the intended sense of externalization and co-presence. In contrast, the stereo version scored higher in instrumental clarity and rhythmic stability, aligning with the anchoring function of presence. These results suggest that machine-learning models can serve as complementary tools to corroborate how immersive design strategies manifest in measurable descriptors.

Regarding the guiding question —*how can a hybrid workflow that combines re-amping with a binaural head and HRTF synthesis be implemented in popular music production to deliver a headphone-optimized master without losing stereo translatability?*— the findings indicate that this integration is feasible under three conditions: (a) parallel monitoring of binaural and stereo outputs throughout the mixing process, (b) calibration of spatial rules (anchors, periphery, height) to balance immersion with clarity, and (c) iterative rendering tests across consumer platforms to guarantee consistency. These conditions allowed the final master to retain spatial richness in headphones while remaining compatible with conventional stereo playback.

6.2 Critical analysis of creative decisions

Creative decision-making was grounded in formal, timbral, and dynamic analysis of the original piece and in design derived from the *Reality ↔ Beyond* concept. Under this premise, a stereo-plus strategy kept lead vocal, bass, and rhythmic axes center-focused to sustain intelligibility and impact, while pads, backing vocals, sequences, and FX were reserved for depth/height layers and for trajectories that expand the scene. Implementation combined mono/binaural re-amps and HRTF panning on selected stems, with targeted overdubs marking semantic inflection points. The outcome is spatial selectivity that preserves stereo fold-down translation and remains traceable in the mix session and the procedures table.

6.3 Relation to theoretical framework and state of the art

Spatialization decisions are anchored in binaural psychoacoustics: deliberate use of ITD/ILD and HRTF-based spectral filtering to place sources in azimuth/elevation/distance under headphone listening (Blauert, 1997; Møller, 1992). They also align with immersive-production practices that treat the acoustic scene as part of musical discourse rather than ornament (Roginska & Geluso, 2017). By translating lyrical motives into spatial gestures, the project situates itself where the field is moving—beyond "how to render" toward designing spatial dramaturgy coherent with musical narrative.

Computational findings with MusiCNN/Essentia—aggregated at track level—show an affective shift in the binaural version toward relaxed/voice and away from party/danceable/aggressive, consistent with greater externalization and re-hierarchization of layers. Interpreted alongside the technical context (re-amps/HRTF/overdubs), this pattern supports the view that spatialization modulates emotional appraisal without sacrificing rhythmic or tonal clarity, as reflected in higher beats_confidence.

6.4 Innovation and knowledge transfer

The project addresses the operational gap identified in the review. Specifically, it responds to the need for auditable case studies that document—step by step—the integration of creative/narrative decisions with binaural capture, re-amping, and synthesis in popular music, under explicit criteria for timbral control and perceptual validation. It delivers a replicable workflow with technical traceability—time chart, spatial diagrams, and position/level metadata—and a public outcome (binaural track plus interactive web interface) that materializes university–industry–artist transfer. The use of a patented binaural-capture device in the re-amps

reinforces applied innovation and technology transfer to professional practice.

6.5 Limitations and considerations

Methodologically, the MusiCNN analysis relied on track-level aggregates using 3-s windows with 50% overlap; temporal variability and inferential significance were therefore not estimated. Potential level bias was mitigated by running a parallel −14 LUFS-I normalized condition alongside the raw extraction. Future work should extend to time-by-variable curves, nonparametric inference, and controlled listening tests. On the production side, the release is fixed-binaural: externalization is head-locked (no head tracking), which invites comparisons with head-tracked renderers and studies of loudspeaker translation that preserve spatial dramaturgy.

6.6 Final synthesis

The hybrid workflow demonstrates that musical legibility can be preserved while narrative scope is expanded through depth, height, and external trajectories. The results are consistent with binaural theory and current practice in immersive production. Complementarily, Essentia/MusiCNN yielded segment-level descriptors and tagging distributions that reliably distinguished the binaural and stereo renders in line with the intended spatial design (e.g., shifts along affective axes and the acoustic–electronic balance), providing convergent, reproducible evidence alongside listening tests. By articulating a reproducible and transferable framework, the study shows that spatial design can reshape perceived valence and arousal, decoupling expressiveness from level. Can spatial design be formalized as a primary compositional axis (a "space score") whose affective outcomes generalize across HRTFs and devices, and are predictable from Essentia/MusiCNN tagging?

REFERENCES

Alonso Cardona-Cano, C., Calle, J. S., & López Díez, J. (2024). Quantifying the impact of head-tracked spatial audio on common user auditory experiences using facial micromovements (Paper 269). Paper presented at the 152nd *Audio Engineering Society Convention.* New York, United States. Retrieved from https://aes2.org/publications/elibrary-page/?id=22727

Alonso-Jiménez, P., Bogdanov, D., Pons, J., & Serra, X. (2020). TensorFlow audio models in Essentia. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 266–270). Barcelona, Spain: IEEE. https://doi.org/10.1109/ICASSP40776.2020.9054688

Arango-Lopera, C. A., Escobar-Sierra, M., & Cardona-Cano, C. A. (2024). Sonority and popularity of reggaeton: From the ghetto to the mass. *SCIRES-IT - SCIentific RESearch and Information Technology, 14*(2), 169–184. https://doi.org/10.2423/i22394303v14n2p169

Begault, D. R. (1994). *3-D sound for virtual reality and multimedia.* NASA Ames Research Center.

Best, V., Kopčo, N., & Shinn-Cunningham, B. (2020). Sound externalization: A review of recent research. *Trends in Hearing,* 24, 1–14. https://doi.org/10.1177/2331216520948390

Blauert, J. (1997). *Spatial hearing:The psychophysics of human sound localization* (Rev. ed.). Cambridge, MA: MIT Press.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., & Serra, X. (2013). Essentia: An open-source library for audio analysis and music information retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)* (pp. 493–498). International Society for Music Information Retrieval.

Carrozzino, M., Angeletaki, A., Evangelista, C., Lorenzini, C., Tecchia, F., & Bergamasco, M. (2013). Virtual technologies to enable novel methods of access to library archives. *SCIRES-IT - SCIentific RESearch and Information Technology*, *3*(1), 25–34. https://doi.org/10.2423/i22394303v3n1p25

Casales, A. (2024). Aproximaciones a la inmersión y su percepción auditiva. *Calle 14: Revista de Investigación en el Campo del Arte,* 19(36), 288–305. https://doi.org/10.14483/21450706.20572

Collins, K., & Dockwray, R. (2018). Experimental sound mixing for the well, a short film made for tablets. *Leonardo Music Journal*, 28, 65–71. https://doi.org/10.1162/lmj_a_00996

De Gregori Astrici, A. (2018). *Técnicas de diseño sonoro para narrativas inmersivas en cine y realidad virtual* (Tesis doctoral, Universidad Complutense de Madrid). Repositorio Institucional UCM. https://eprints.ucm.es/id/eprint/50052

Dewey, C., Moore, A., & Lee, H. (2024). *Practitioners' perspectives on spatial audio: Insights into Dolby Atmos and binaural mixes in popular music. Journal of the Audio Engineering Society,* 72(7–8), 504–516. https://doi.org/10.17743/jaes.2022.0153

Estados Alterados. (2018). *Lumisphera* [Album]. Estados Alterados. Retrieved from https://open.spotify.com/intl-es/album/4n5NrxwtF4QrnOJhynbGOI?si=qi-oTyoOSX6Wx7FfioHWMQ

Estados Alterados. (2022). *Binaural* [Website]. Retrieved from https://binaural.estadosalterados.net/

Estados Alterados. (2023). *Mantra binaural – experiencia inmersiva* [Song]. Estados Alterados. Retrieved from https://open.spotify.com/album/1TLOx0TYyfiIaNKt91D5mR?si=SI4PACmJTGenQnz-qKIhvA

Fantini, D., Presti, G., Geronazzo, M., Bona, R., Privitera, A. G., & Avanzini, F. (2023). Co-immersion in audio-augmented virtuality: The case study of a static and approximated late-reverberation algorithm. *IEEE*

Transactions on Visualization and Computer Graphics, 29(11), 4472–4482. https://doi.org/10.1109/TVCG.2023.3320213

Fontana, S., Farina, A., & Grenier, Y. (2007). Binaural for popular music: A case study. In *Proceedings of the 13th International Conference on Auditory Display (ICAD 2007)* (pp. 85–90). Montréal, Canada.

Gerzon, M. A. (1973). Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society,* 21(1), 2–10.

Grand View Research. (2022). *3D audio market size, share & trends analysis report,* 2022–2030. Grand View Research. Retrieved from https://www.grandviewresearch.com/industry-analysis/3d-audio-market

Grundhuber, P., Lovedee-Turner, M., & Habets, E. A. P. (2024). NBU: Neural binaural upmixing of stereo content. In *Proceedings of the 27th International Conference on Digital Audio Effects (DAFx24)* (pp. 404–411). Guildford, Surrey, UK. Retrieved from https://www.dafx.de/paper-archive/2024/papers/DAFx24_paper_36.pdf

Howie, W., Kamekawa, T., & Morinaga, M. (2023). Case Studies in Music Production for Advanced 3D Audio Reproduction with Bottom Channels (1.3). Zenodo. https://doi.org/10.5281/zenodo.7710002

Kelly, J., Woszczyk, W., & King, R. (2020, October*). Are you there? A literature review of presence for immersive music reproduction* [Paper presentation]. 149th Audio Engineering Society Convention, Online.

Kerins, M. (2011). *Beyond Dolby (Stereo): Cinema in the Digital Sound Age*. Bloomington, IN: Indiana University Press.

Kopp, B. (2022). What is immersive audio?: How engineers, artists & industry are changing the state of sound. *GRAMMY.com*. https://www.grammy.com/news/what-is-immersive-audio-industry-explainer-dolby-atmos

López, X. X. (2012). Algunas ideas sobre la inmersión sonora [Blog post]. *Un Ruido Secreto*. Retrieved from https://www.unruidosecreto.net/algunas-ideas-sobre-la-inmersion-sonora/

López, M., Kearney, G., & Hofstädter, K. (2022). Seeing films through sound: Sound design, spatial audio, and accessibility for visually impaired audiences. British Journal of Visual Impairment, 40(2), 117–144. https://doi.org/10.1177/0264619620935935

Małecki, P., Piotrowska, M., Sochaczewska, K., & Piotrowski, S. (2020). *Electronic music production in ambisonics—Case study. Journal of the Audio Engineering Society, 68*(1/2), 87–94. https://doi.org/10.17743/jaes.2019.0048

Michael, K. (2024). Towards a taxonomy for immersive music performance. *Music & Practice, 11*. https://doi.org/10.32063/1107

Møller, H. (1992). Fundamentals of binaural technology. *Applied Acoustics,* 36(3-4), 171-218. https://doi.org/10.1016/0003-682X(92)90046-T

Paik, S., Han, J., Lee, T., & Lee, K. (2024*).* Case study on high order ambisonics music production: Music and technology within an ambisonics framework using Korean traditional instruments. In *Proceedings of the AES 5th International Conference on Audio for Virtual and Augmented Reality (AVAR)* (2024, August 19–21). DigiPen Institute of Technology, Redmond, Washington, USA.

Pons, J., & Serra, X. (2019). musicnn: Pre-trained convolutional neural networks for music audio tagging. arXiv. https://doi.org/10.48550/arXiv.1909.06654

Python Software Foundation. (2023). Python (Version 3.11) [Computer software]. https://www.python.org/

Reardon, G., Faller, C., & Frank, A. (2018, August 20-22). Evaluation of binaural renderers: Externalization, front/back and up/down confusions [Conference paper]. *AES International Conference on Audio for Virtual and Augmented Reality,* Redmond, WA, United States.

Roginska, A., & Geluso, P. (Eds.). (2017). *Immersive sound: The art and science of binaural and multi-channel audio.* New York, NY: Routledge.

Rumsey, F. (2021). Whose immersive audio? Technology, media, and markets. *Journal of the Audio Engineering Society,* 69(1/2), 142-148.

Singleton, T. (2021). Apple Music's spatial audio is an opportunity for producers, but challenges remain. *Billboard.* Retrieved from https://www.billboard.com/pro/apple-music-spatial-audio-dolby-atmos-producers-challenges/

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 6(6), 603–616. https://doi.org/10.1162/pres.1997.6.6.603

Sun, L., Zhong, X., & Yost, W. A. (2015). Dynamic binaural sound source localization with interaural time difference cues: Artificial listeners [Conference abstract]. The Journal of the Acoustical Society of America, 137(4), 2226.

Warp, R., Zhu, M., Kiprijanovska, I., Wiesler, J., Stafford, S., & Mavridou, I. (2022). Moved by sound: How head-tracked spatial audio affects autonomic emotional state and immersion-driven auditory orienting response in VR environments. Paper presented at 152nd *Audio Engineering Society (AES).* https://www.aes.org/e-lib/browse.cfm?elib=21703.

Wingstedt, J., Brändström, S., & Berg, J. (2010). Narrative music, visuals and meaning in film. *Visual Communication, 9*(2), 193–210. https://doi.org/10.1177/1470357210369886