DISENTANGLING AUTHORED SPACE IN 7.1.4 MUSIC: A LARGE-SCALE CORPUS AND A NEURAL SPATIAL FILTERING CODEC

Anonymous authors
Paper under double-blind review

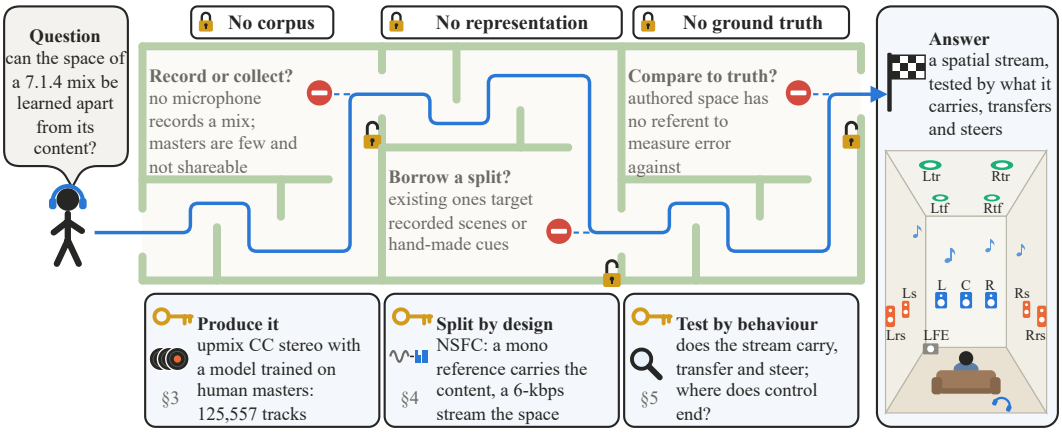


Figure 1: Each chamber is one of the three difficulties of Section 1: one branch ends where an obvious route fails, the other reaches a door opened by the method below. The listener who enters with headphones ends in a 7.1.4 room: L, C, R at the front (blue), four surrounds (orange), four height channels on the ceiling (green) and the LFE (grey).

ABSTRACT

Spatial intelligence has so far dealt with space that has a physical referent; the space people hear in music is authored, since every channel assignment in a 7.1.4 master is a mixing decision. Learning it has lacked a corpus, a representation that holds the space apart from the content, and a ground truth to check one against; to our knowledge, the largest channel-based 7.1.4 music dataset available for research has ten tracks. We produce 125,557 tracks and 8,392 hours of 7.1.4 music from Creative Commons–licensed stereo with an upmixing model trained on human masters, which is closer to them than a commercial upmixer on all five objective metrics and is rated by listeners between the two. Since a mix shapes its content through the relations between channels, the space can be carried as the filtering that turns a mono reference into twelve channels: our neural spatial filtering codec (NSFC) leaves the reference to a pretrained mono codec and transmits the space at a nominal 6 kbps (13.75 kbps in total), optionally with a flow-matching postfilter. Without ground truth, we test the separation by its behaviour, in downstream tasks built from the corpus that fix the reference and change only the spatial codes, so that they apply to any representation learned on it that carries the space apart from the content. Within the domain of the corpus, the stream carries spatial structure beyond static per-channel gains, transfers across musical content and, when substituted, moves the mix in the intended direction. Authored space thus becomes something that can be learned, separated from content and tested.

1 INTRODUCTION

The turn of AI research towards the physical world has made spatial intelligence, the capacity to perceive, represent and reason about space, a shared agenda across world models, embodied agents

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and 3D perception (Liu et al., 2026; Yang et al., 2025; Bruce et al., 2024). The space these systems deal with comes in two kinds. One exists, or is simulated to exist, and sound or imagery is captured or rendered to match it. The other is authored: people create it as part of a work and it depicts nothing beyond itself, as in the composition of a painting or the phantom stage of a stereo mix. Spatial audio's part in this agenda is comparatively small and built on the first kind (Section 2). Authored auditory space is the side still to enter the agenda. It is also the side people actually hear: the spatial audio of cinema, streaming music, games and headsets is mixed rather than recorded, and reaches listeners through channel-based loudspeaker layouts.

Channel-based music in 7.1.4 is the purest case: no picture to follow, no geometry to render. The layout is defined by loudspeaker positions, seven at ear height, four overhead and one low-frequency channel (ITU-R, 2022).

A 7.1.4 music mix embodies the mixing engineer's creative decisions about the placement, perceived width, and movement of musical elements. The resulting twelve-channel signal carries this authored spatial organisation. For representation learning, bringing it into the agenda takes three things: a corpus to learn from, a representation that holds the spatial organisation apart from the musical content, and a way to check what it captures and how far it can be steered. None exists yet at a scale that representation learning can use.

Such a corpus cannot be recorded: a space with a physical referent can be captured again and again (Shimada et al., 2023; Guo et al., 2025), whereas authored space exists only in mixes that people have already made. Those mixes are recent and few. Music with height channels reached listeners between 2019 and 2021,[1] and each title needs its own specialist mixing pass from the stems, usually after the stereo mix (Dewey et al., 2024; Bourbon et al., 2023), so immersive masters cover a small fraction of the catalogue. What exists is largely out of reach for research: commercial masters reach listeners as lossy, rights-managed object-coded streams (Purnhagen et al., 2016) under licences that exclude redistribution, and the Creative Commons–licensed catalogues that stereo music research relies on (Defferrard et al., 2017; Bogdanov et al., 2019) are stereo, having never gone through that production step. The channel-based 7.1.4 music formally published as a research dataset is AlbumDB, ten tracks and about 34 minutes (McKenzie et al., 2026); authored spatial organisation in music has been observed at that scale.

Even given data, the object is not directly observable. Spatial cues live in the relations among the twelve channels, encoded together with the content in the same samples of every channel; no channel is the space. Existing separations address other objects: hand-designed parametric cues, microphone arrays, Ambisonic scenes and binaural signals, or synthetic five-channel panning without height (Section 2). Nothing supervises the split: there are no labels for the space. Even a separated representation has no ground truth to be checked against. Authored space has no physical referent against which error could be measured; the space is the mix. Disentanglement metrics likewise presuppose known generative factors (Locatello et al., 2019); here there are none.

This paper approaches the three difficulties from the data side. The corpus is produced rather than collected: a stereo-to-7.1.4 upmixing model trained on 7.1.4 masters mixed by hand by musicians is applied to Creative Commons–licensed stereo catalogues (Section 3). Trained on human mixes rather than on a geometric panning rule, the model imposes a spatial organisation that targets human mixing practice; and because each stereo source carries a Creative Commons licence, each output track can be redistributed under the terms of its source's licence. To separate space from content we use the fact that, in a mix, spatial organisation acts on content through relations between channels: given the content as a monophonic reference, the space can be carried as the filtering that turns that reference into twelve channels, which is small and can be handled on its own. We build a 12-channel **neural spatial filtering codec (NSFC)** on this principle (Section 4): the content travels through a pretrained monophonic codec (Kumar et al., 2023), used unchanged, and the spatial structure as a separate stream, optionally refined by a flow-matching postfilter (Lipman et al., 2023) that gains spatial quality at a consistent cost in waveform fidelity. With no ground truth to compare against, we test the separation by its behaviour: disentanglement here means that the spatial stream can be separated from the content, moved across content and steered, not statistical independence. Within the domain of the corpus, the tests show the separation to be real and usable, with bounded control.

[1] Launch announcements: Dolby Laboratories, 25 September 2019 (Amazon Music HD) and 12 December 2019 (TIDAL); Apple, 17 May 2021 (Apple Music, available June 2021).

Contributions. (i) A corpus of 125,557 tracks and 8,392 hours of 7.1.4 music under per-track Creative Commons licences, refined by licence screening, deduplication, quality filtering and a frozen split (Section 3). (ii) A stereo-to-7.1.4 upmixing model trained on hand-mixed human masters, which is closer to the human masters than a commercial upmixer on all five reported objective metrics and is rated by listeners between the two (Section 3.2). (iii) The NSFC, which carries the spatial organisation as a 6-kbps stream apart from a mono reference, for a 13.75-kbps complete system, in a deterministic and a flow-matching configuration, with an ablation of its bounded complex mask (Section 4). (iv) Three downstream tasks that evaluate, without ground truth, spatial representations learned on the corpus: whether they carry the spatial organisation, transfer across content and steer the mix, and where that control ends (Section 5).

2 Related Work

Authored space as an object. Audio enters the agenda set out in Section 1 through audio-visual navigation (Chen et al., 2020), spatial sound reasoning with language models (Zheng et al., 2024), sound event localisation and detection benchmarks (Shimada et al., 2023), spatial audio generated to match video (Liu et al., 2025; Xie et al., 2026) and the wider field of sound-field capture and room acoustics (Koyama et al., 2025). In all of these the space is given, physically or by simulation, and the task is to perceive or reproduce it. Immersive audio formats are classified as channel-, object- or scene-based (ETSI, 2022); whatever the format, object- and scene-based signals are reproduced through loudspeakers configured for channel-based signals (ITU-R, 2022), so the channel-based layout is the common denominator of produced spatial audio. Outside machine learning, authored space is studied: composers organise and notate it deliberately (Catena, 2025; Kim-Boyle, 2025; Frisk & Schacher, 2025), practitioners build the stereo mix first and then spread it into the immersive layout (Dewey et al., 2024), and commercial 7.1.4 mixes are analysed by per-loudspeaker loudness, correlation and usage (Bourbon et al., 2023). As an object of learning it has stayed outside that agenda; the practitioners' account, content first and space afterwards, motivates separating the two.

Data, and how it is produced. Multichannel corpora exist on the referent side, as microphone-array and Ambisonic recordings of real scenes, music and singing (Shimada et al., 2023; Paik & Lee, 2025; Mróz, 2026; Meza et al., 2025) and binaural recordings of instrumental performance (Guo et al., 2025). They grow by recording, and the room they were made in fixes their spatial content. On the authored side the formally published corpus is AlbumDB (McKenzie et al., 2026), ten songs with stems and both a stereo and a 7.1.4 master; nothing larger exists, and upmixing offers a route around this. Upmixing has been learned as latent-domain upsampling and upmixing (Bralios et al., 2025) and as generative mono-to-stereo or spatial synthesis (Zhang et al., 2025; Heydari et al., 2025), and evaluated as a listening product against stereo and dedicated mixes (Poets et al., 2023); here it is a data-production engine, trained on human masters and compared with them and a commercial upmixer.

Representation and disentanglement. That space can travel separately from content has a long history in coding: binaural cue coding (Faller & Baumgarte, 2003) and MPEG Surround (Herre et al., 2008) transmit a downmix with inter-channel level, time and coherence cues, joint object coding reconstructs objects from a multichannel downmix and parametric side information and delivers commercial immersive music (Purnhagen et al., 2016), and MPEG-H unifies channels, objects and Ambisonics (Herre et al., 2015). These cues are hand-designed perceptual parameters that serve reconstruction, not representations to be read or steered, and learned codecs matured on mono and stereo (Zeghidour et al., 2022; Défossez et al., 2023; Kumar et al., 2023). Learned spatial codecs address captured space: a reference channel plus a quantised spatial stream decoded into filters, for microphone-array speech (Xu et al., 2024); content and spatial cues coded separately for binaural speech (Ratnarajah et al., 2025); and latent codecs for first-order Ambisonics inside generative systems (Heydari et al., 2025). For music, upmixing as style transfer separated the spatial image from content on synthetic five-channel VBAP mixes without height (Yang et al., 2022), and source-level disentanglement inside a codec shows that factorised codebooks can carry separate streams (Bie et al., 2025). This paper keeps the structure of the parametric coders, learns the spatial stream over a pretrained monophonic reference, and applies it to authored 7.1.4 music, judged by what it enables.

3 Dataset Construction

3.1 Internal Source Data

The upmixing model learns from an internal collection of 2,000 7.1.4 masters that musicians of our team mixed by hand over several years, used here with permission. The masters are sampled at 48 kHz, and a stereo version of each is obtained by a standard ITU-R downmix, so that every stereo input has a human 7.1.4 counterpart. The collection holds 115.24 hours with a median track duration of 3.36 min and is split by track into 1,600 training, 200 validation and 200 test tracks. The collection serves only to train and evaluate the upmixing model: it is not released and enters none of the experiments of Sections 4 and 5 (duration and genre statistics in Appendix A).

3.2 Upmixing Model

The upmixing model estimates the twelve channels of a 7.1.4 mix jointly from a stereo input $s = [s_L; s_R]$, in the manner of multi-stem source separation with the channels as stems. Following band-split transformers for source separation (Lu et al., 2024), the complex STFT of both input channels is divided into overlapping bands on the mel scale, and transformers alternate across time within every band and across bands within every frame (Appendix A). For every output channel c a per-band network then estimates a complex mask $\mathcal{M}_c$, averaged where bands overlap, and applies it to the spectrum of the mid signal:

$$Y_c(t, f) = \mathcal{M}_c(t, f)\, \tfrac{1}{2}\big(S_L(t, f) + S_R(t, f)\big). \tag{1}$$

Every channel is thus a time–frequency filtering of one carrier, and the spatial organisation the model produces is the set of per-channel complex gains, the view of a mix that the codec of Section 4 takes; the masks are estimated from the two input channels separately, so the stereo image guides it.

3.3 Training and Evaluation

The model is trained on 12-s segments of the 1,600 training masters (Section 3.1), with their stereo downmix as input, using waveform and complex-spectrum reconstruction losses on every channel, a perceptually weighted multi-resolution STFT loss on a mel scale, and two penalties that keep the sparse surround and height channels from collapsing to silence (Appendix A). One frozen model upmixes all four catalogues. We compare it with the human masters and with the commercial upmixer Halo Upmix (NUGEN Audio, 2026), run with its Music 3D preset on the same input.

The objective comparison uses the 200 internal test tracks, with the human master as reference: SI-SDR (Le Roux et al., 2019) and the log-spectral distance (LSD), both averaged over the twelve channels; the spatial and residual signal ratios SSR and SRR of SPAUQ (Watcharasupat & Lerch, 2024), which separate inter-channel gain and delay errors from the remaining distortion; and the Fréchet audio distance (FAD) (Kilgour et al., 2019) with CLAP music embeddings (Wu et al., 2023; Gui et al., 2024), computed per channel and averaged over the eleven non-LFE channels (SNR and a multi-resolution mel distance, close to the training losses, in Appendix A). The model is closer to the human master than Halo Upmix on all five metrics (Table 1), with under a third of its FAD, and most of all in the rear surround channels (Appendix A).

Table 1: Upmixing model: objective metrics against the human master (FAD in 10^{-3}) and listening-test MOS.

	SI-SDR	SSR	SRR	LSD	FAD	MOS ↑	
	dB↑	dB↑	dB↑	dB↓	↓	Qual.	Spat.
Human master	reference					3.61	3.61
Our model	5.12	9.00	18.07	13.67	19.47	3.44	3.41
Halo Upmix	−1.23	4.89	12.42	14.29	64.50	3.39	3.28

The listening test uses 30 tracks: 20 of the internal test tracks and the ten masters of AlbumDB (McKenzie et al., 2026), whose stereo versions are matrix downmixes of their 7.1.4 mixes. Each track is presented in three versions, the human 7.1.4 master and the outputs of the two upmixers, as 20-s excerpts of the same chorus section, to 24 trained listeners on a calibrated 7.1.4 loudspeaker system (Appendix A). All three versions, the human master included and not presented as a reference, were rated on a five-point scale for three attributes of audio quality and three of spatial rendering, averaged into quality and spatial mean opinion scores (MOS). The model is rated between the commercial upmixer and the human master in both domains (Table 1); spectral balance is the one attribute on which Halo

Upmix is rated higher. The human master remains ahead by 0.17 in quality and 0.20 in the spatial MOS, with 95% intervals of the paired differences excluding zero. The analysis was defined after the test and is descriptive (Appendix A).

3.4 REFINEMENT

Every candidate track receives one decision combining four independent examinations: licence, duplication, technical quality and content, and task scope (Appendix B). The four catalogues publish each track under its own licence: CC0, CC BY, CC BY-SA, CC BY-NC and CC BY-NC-SA, and on Dogmazic the Free Art License; none forbids derivative works. The declared licence is not sufficient on its own. An upmix is an adaptation, so every obligation of the source licence carries over to the 7.1.4 derivative: attribution requires an author and a source for each track, the copyleft terms of share-alike and of the Free Art License require the exact licence version so that the derivative can carry it, and the terms of a remix depend on the works it samples. We therefore verify, per track, the licence family and version, the author and source fields and, for remixes, the declared chain of upstream sources, and exclude tracks for which these cannot be established. This excludes, among others, every Free Art License track, none of which states its version, so each retained track, and with it its 7.1.4 derivative, carries CC0 or a Creative Commons licence of a known version. The result is a per-track licensed research dataset with conditional redistributability, predominantly restricted to noncommercial use.

Duplicates inflate the apparent size of a corpus and leak between training and evaluation. A duplicate is the same complete musical content, the same performance with the same structure (Appendix B); candidates come from byte-exact hashing and acoustic fingerprinting with Panako (Six & Leman, 2014), and each duplicate group keeps one representative. No track matches AlbumDB.

The upmixing model passes on whatever its input contains, and a spatial output can carry no more inter-channel information than its input supplies. Each track is therefore examined for technical quality and, by whole-track inference with Discogs-EffNet (Alonso-Jiménez et al., 2022), for non-musical content; a subset of these checks are curation rules that exclude a track, the rest are recorded as per-track statistics (Appendix B). Outputs that exceed full scale only briefly and slightly are kept, corrected by one gain shared by all twelve channels that brings the true peak to 0 dBTP and leaves the level relations between channels, and so the spatial cues, unchanged; larger overload is excluded.

The evaluation question is generalisation to unseen tracks from the same domain, so the split unit is the complete track, and no copy or shared material of a test track may appear in training. The task scope keeps tracks between 30 s and 15 min, which guarantees complete training segments from every track and bounds the long tail of durations. Duplicate groups are never divided, tracks that share partial material with another track stay in the training split, and AlbumDB lies outside all splits. The validation and test splits are sized in hours and stratified by source, with six further attributes balanced deterministically (Appendix B); the test split is used neither for tuning nor for model selection. The resulting split holds 125,011 training, 183 validation and 363 test tracks.

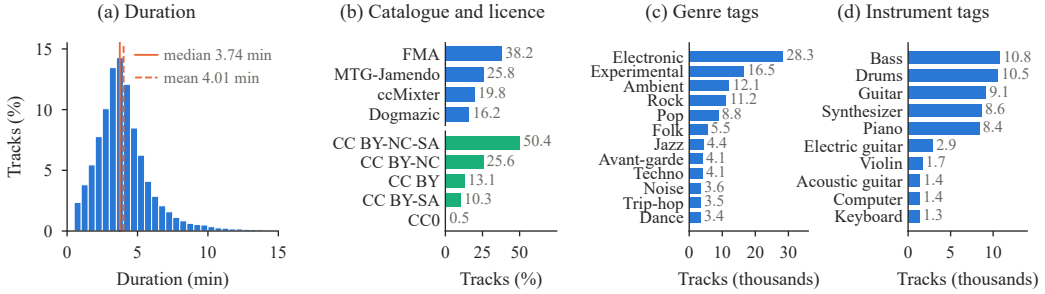


Figure 2: Statistics of the corpus. (c) The 12 most frequent of 27 genre terms shared by the source sites; a track may carry several, and 65.6% carry at least one. (d) The 10 most frequent instrument tags, recorded for 23.7% of the tracks and for none on Dogmazic; tags are source metadata, not verified acoustic content.

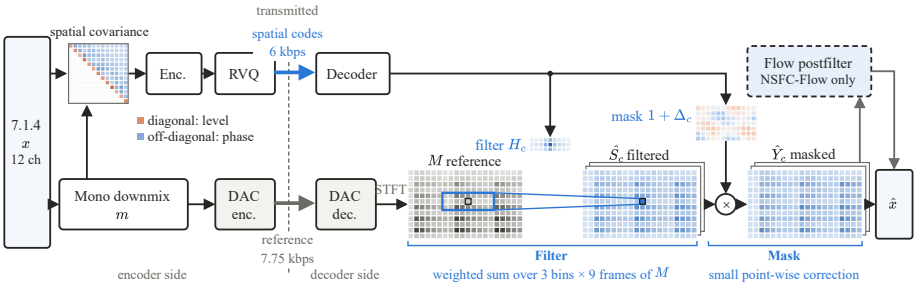


Figure 3: The NSFC. Only the 6-kbps spatial stream is transmitted besides the reference (DAC-coded or uncoded). Per channel c the decoder predicts a complex ratio filter H_c and a bounded mask $1 + \Delta_c$ (Eqs. 2–3); NSFC-Flow adds a flow-matching postfilter. Spectra are schematic.

3.5 Corpus Statistics

The corpus holds 125,557 tracks and 8,392.55 hours; each track pairs its Creative Commons stereo source with the twelve channels of its 7.1.4 upmix (Figure 2). After acceptance, the corpus will be released track by track under the licence of its source, with the attribution that the licence requires.

4 Spatial Music Representation

4.1 Neural Spatial Filtering Codec

Our starting point is the hypothesis that, in spatial music, the content and its spatial cues can be separated: the content can be held in a single reference signal, and the spatial organisation carried apart from it. A codec built on this split puts the hypothesis to a direct test, since the twelve channels must be recovered from the reference and a stream that carries only the spatial side. Parametric coders (Section 2) make the split plausible, having long carried inter-channel cues as side information alongside a downmix. Here the spatial side information is learned instead: an encoder compresses the inter-channel relations of a 7.1.4 mix into a discrete spatial stream, and a decoder turns the stream into filters that distribute the reference over the twelve channels (Figure 3). We call this system a **neural spatial filtering codec (NSFC)**.

The split also keeps the codec cheap and modular. Only the spatial layer is trained: the reference is left to an existing high-fidelity 48-kHz mono codec, here DAC (Kumar et al., 2023), so no codec is trained for the content, and any such codec can carry the reference without retraining.

Let $x \in \mathbb{R}^{12\times T}$ be a 7.1.4 mix. The reference is a fixed mono downmix $m = w^\top x \in \mathbb{R}^{1\times T}$, with standard downmix weights $w \in \mathbb{R}^{12}$ and no LFE contribution (Appendix C). At every time–frequency point of an STFT the encoder receives a **normalised spatial covariance** of the thirteen signals $[m; x]$: the off-diagonal entries are cross-spectra divided by the geometric mean of the two powers, which leaves the inter-channel phase, and the diagonal carries the log power of each signal, which leaves the inter-channel level. Unlike the raw covariance used for speech arrays (Xu et al., 2024), it separates the relations between channels from how loud the music is (Appendix C). The log-compressed spectrum of the reference is appended as a further input. A convolutional encoder emits one vector per STFT frame, quantised by residual vector quantisation with eight stages of 1024 entries as in DAC (Kumar et al., 2023); at 80 bits per frame and 75 frames per second, the stream runs at 6 kbps.

The decoder mirrors the encoder and outputs, for every channel c, a complex ratio filter H_c that spans three frequency bins and nine frames of the reference spectrum M:

$$\hat{S}_c(t,f) = \sum_{\nu=-1}^{1} \sum_{\tau=-4}^{4} H_c(t,f;\nu,\tau)\, M(t+\tau,\ f+\nu). \tag{2}$$

The filter expresses what a mixer does to a source: it assigns, delays and decorrelates the same content across channels. A filter of this size, however, cannot represent every local spectral detail. The

Table 2: Test-set results against the original 7.1.4 mix: objective metrics on 363 tracks (4-s segments; FAD in 10^{-3}), MOS from 24 listeners on 10-s excerpts, 15 tracks per NSFC system. Rates are nominal; uncoded rows count only the spatial stream. Bold: the better decoder configuration under each reference.

		Rate	SI-SDR	SSR	SRR	LSD	FAD	MOS ↑	
Reference	System	kbps	dB ↑	dB ↑	dB ↑	dB ↓	↓	Quality	Spatial
Uncoded	NSFC	6	**9.58**	13.39	**18.95**	**5.46**	0.15	3.30	3.31
	NSFC-Flow	6	8.34	**13.61**	15.63	7.25	**0.07**	**3.40**	**3.41**
DAC	NSFC	13.75	**6.35**	10.85	**10.28**	**13.47**	**33.94**	3.02	3.14
	NSFC-Flow	13.75	5.67	**10.97**	9.56	14.08	34.45	**3.26**	**3.26**

decoder therefore also predicts a complex mask per channel and time–frequency point, estimated jointly in its real and imaginary parts as in mask-based source separation (Lu et al., 2024):

$$\hat{Y}_c(t,f) = \big(1 + \Delta_c(t,f)\big)\,\hat{S}_c(t,f), \qquad \Re\Delta_c = 0.25\tanh(\cdot),\ \ \Im\Delta_c = 0.25\tanh(\cdot). \tag{3}$$

The mask is centred on the identity, initialised at zero and bounded in both components, so it can only correct the filter output locally and cannot take over the reconstruction; without the bound, the encoder learns to shrink its latents and let the mask restore the gain, and the codebooks collapse (Appendix C). The mask adds 3,072 parameters. The codec is trained end to end on waveform L1 and L2 losses, a multi-resolution STFT loss and the quantisation losses.

Flow-matching (Lipman et al., 2023) postfilters have been shown to raise the perceptual quality of coded audio from the output of a frozen codec, at some cost in waveform fidelity (Welker et al., 2025). What the NSFC has to preserve is itself largely perceptual, the spatial image rather than the exact waveform, which makes such a postfilter a promising addition. We train one on the output of the frozen NSFC and call the result **NSFC-Flow**. The postfilter is a flow in a compressed complex STFT domain that starts from the NSFC spectrum perturbed by noise of frequency-dependent scale $\sigma(f)$; for multichannel audio, $\sigma(f)$ is shared by all twelve channels, so the noise introduces no systematic level difference between channels (Appendix C).

Both configurations decode the same spatial codes, so the postfilter changes neither the stream nor its rate, and Section 5 applies to both. For transmission, the reference is coded by DAC with nine codebooks at 7.75 kbps, for a complete system of 13.75 kbps. The spatial layer is trained on the uncoded reference and its coding errors propagate through the filters, so both conditions are reported.

4.2 Experimental Setup

Training draws random 4-s segments from the 125,011 training tracks (Section 3.4); the 183 validation tracks select the checkpoint within the first 100k updates, and the 363 test tracks (21,427 segments) are used only for the final evaluation. The NSFC-Flow postfilter is trained for a further 60k updates on the frozen NSFC (details in Appendix C). We report the metrics of Section 3.3, SI-SDR over the twelve channels, SSR and SRR, with LSD over the non-LFE channels and FAD on a mono downmix. Differences are tested by a track-level cluster bootstrap with Holm correction in three pre-specified families on SI-SDR and SSR only; the others carry 95% intervals (Appendix D). The listening test uses the listeners, playback system and six attributes of Section 3.3, with 10-s excerpts of 30 tracks, 20 from the four catalogues and the ten AlbumDB masters: after the original 7.1.4 mix as an unrated reference, listeners rated two of the four NSFC systems per track for how well each reproduced the reference (Appendix D).

Ablation. The ablation isolates the bounded complex mask, the component that the NSFC adds to the complex ratio filter: the same codec is trained without it for the same 100k updates, the matched comparison, and for 160k updates, to test whether longer training replaces the mask (Appendix D).

4.3 Results

Mask ablation. Trained for the same 100k updates, the codec with the mask exceeds the codec without it by 0.58 dB in SI-SDR and 0.82 dB in SSR (both Holm-corrected $p < 10^{-4}$) and improves

seven of eight metrics; SRR falls by 0.67 dB. It also exceeds the codec without the mask trained for 160k updates, by 0.18 dB in SI-SDR and 0.28 dB in SSR with 62.5% of the updates, while LSD, SRR and FAD remain slightly behind that longer-trained reference (Appendix D).

The postfilter trades distortion for perception. On the uncoded reference, NSFC-Flow raises SSR by 0.22 dB (one-sided $p < 10^{-4}$) and halves FAD, at a cost of 1.24 dB in SI-SDR, 3.32 dB in SRR and 1.79 dB in LSD. Listeners gave NSFC-Flow a higher MOS than NSFC on all six attributes, with both the uncoded and the DAC-coded reference. This is the perception–distortion trade-off reported for flow-matching postfilters in mono coding (Welker et al., 2025), here extended to the spatial image. With the coded reference the SSR gain persists (0.13 dB), but FAD no longer improves.

5 Disentangling Spatial Cues

Without ground truth for how a mix should separate into content and space, we test the separation by its behaviour: we call the spatial stream disentangled from the content if it carries the spatial organisation, transfers across musical content, and, when substituted, steers the decoded mix in the intended direction. Statistical independence is not required. This departs from the classical notion, in which latent dimensions align with known generative factors (Locatello et al., 2019); no such factors are defined for authored space. The material of the tests is built from the corpus itself, so the tasks apply to any representation that carries space apart from content and is learned on the corpus; here they evaluate the NSFC.

5.1 Setup

The tests rest on the mono reference: its downmix weights are equal for every left–right channel pair, so mirroring a mix, or redistributing it with gains that preserve $w^\top a = 1$, leaves the reference unchanged sample by sample, and the decoder sees only different spatial codes. All tests use the frozen NSFC and NSFC-Flow of Section 4, which share one set of spatial codes, with the uncoded reference and 4-s windows, on two kinds of material: *real music*, 32 validation tracks in their original and mirrored form; and *controlled conditions*, the reference of each track redistributed so that one of Ls, Rs, Ltf or Rtf lies 6 dB above each other non-LFE channel (Appendix E). As a reference point, **oracle amplitude panning** fits one least-squares gain per non-LFE channel to the target; it bounds any static per-channel gain and is not a codec at equal rate. The analysis was fixed before the runs; we report per-endpoint 95% bootstrap intervals and no confirmatory tests.

5.2 Results

The stream carries the spatial organisation. A ridge classifier fitted on training tracks reads the four controlled conditions from the time-averaged spatial codes of validation tracks with 87.5% accuracy (95% CI [81.3, 93.0]; chance 25%; controls in Appendix E). On real music, NSFC reconstructs the non-LFE channels with a normalised error of 0.059, against 0.132 for oracle amplitude panning (difference 0.073 [0.056, 0.091]): the 6-kbps stream conveys spatial structure varying over time and frequency, beyond one gain per channel.

The stream transfers across content. Decoding a track's reference with the codes of another track under the same controlled condition reconstructs it no worse than its own codes do (error difference -0.0024 [-0.0044, -0.0004]), whereas codes of a different condition lower the emphasis, as quantified below.

Substituting the stream steers the mix. Encoding the mirrored mix of real music and decoding it on the unchanged reference moves the channel energies towards the mirrored target (Figure 4(b)): the cosine between output and target energy changes is 0.86 [0.76, 0.94], and the output agrees 4.09 dB [3.30, 5.01] better in SSR with the correct target than with the mismatched one; NSFC-Flow keeps this response (Appendix E). The mirrored target is the mirror image of the corpus mix itself, so this tests steering within the domain of the corpus, not fidelity to a human master. Under the controlled conditions (Figure 4(a); Table 15), codes of another condition lower the emphasis, and the drop excludes zero for all four channels. For Rs, the emphasis falls from 6.5 dB with the own codes, against a 6-dB target, to 4.9 dB: the de-emphasis is clear but incomplete.

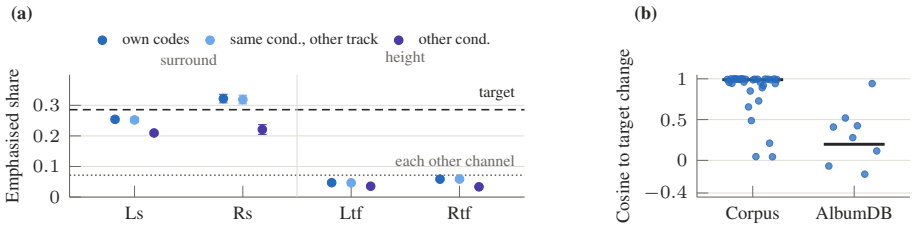


Figure 4: Steering and its limits (NSFC, uncoded reference, validation tracks). (a) Energy share of the emphasised channel under the four controlled conditions (95% block-bootstrap intervals). (b) Mirroring: cosine between output and target changes in channel energy per track; bar, median.

5.3 Where Control Ends

Within the domain of the corpus, the 6-kbps stream thus carries spatial structure beyond static gains, transfers across musical content and moves the mix when substituted. Control has two bounds. First, under height emphasis the output stays below even the share of each other channel (Figure 4(a)): the stream responds to the condition, but the emphasis is never reached. In the corpus the four height channels carry a median 1.9% of the non-LFE energy, so control over the height layer is limited by how little of it the training data contains. Second, all tests run within the corpus domain. The public human masters of AlbumDB, ten tracks of a single album, follow a centre-heavy mixing style far from common 7.1.4 music practice, with a median 71.7% of non-LFE energy in the centre against 0.26% in the corpus; too few and too atypical to characterise human mixes in general, they are reported in Appendix E.

6 Conclusion and Discussion

Authored auditory space has lacked what representation learning needs: a corpus, a representation, and a way to check the representation without a ground truth. This paper addresses all three from the data side. The corpus is produced rather than collected: an upmixing model trained on human 7.1.4 mixes, whose output listeners rated between that of a commercial upmixer and the human masters, turns Creative Commons–licensed stereo into 125,557 tracks of 7.1.4 music. Given the content as a mono reference, the spatial organisation can be carried apart from it as a 6-kbps filtering stream; only this spatial layer is trained, and the content is left to an existing mono codec. Without a ground truth, the separation is tested by its behaviour: the stream carries the spatial organisation, transfers across musical content and, when substituted, moves the mix in the intended direction, within bounds.

The spatial organisation of the corpus is produced by one model, trained on the mixes of one team; its closeness to human mixing has been measured only on internal test tracks and in a listening test, where it remains below the human masters. The codec and the tests of disentanglement are learned and evaluated within this domain.

Both bounds point to what comes next. Control over the height layer is limited by how little of it the training data contains (Section 5.3), so data and objectives that give the height channels more weight are the direct route to steering the complete 7.1.4 layout. The tests of Section 5 need no ground truth and apply unchanged to any collection of 7.1.4 mixes; human masters in greater number and variety than are publicly available today would show how far the separation holds beyond the domain in which it was learned. And because only the spatial layer is trained, it can be paired with better mono codecs as they appear.

Spatial intelligence has so far dealt with space that has a physical referent, whereas the space people hear in music, cinema and games is mostly authored. For authored space to enter that agenda, it needs data to learn from, a representation to learn, and a way to evaluate what has been learned; this paper provides each of the three for channel-based music. They open the way to models that not only perceive or reproduce space but take part in how people organise it, for instance by upmixing or generating music from spatial codes.

AI USE STATEMENT

In this work, we used generative AI tools (Claude models through Claude Code, and OpenAI Codex) for the following tasks with required disclosure: feedback on the design of the corpus refinement and of the experiments, including the operational definition of duplicate recordings, the split design and the evaluation protocols; debugging and optimising data-processing, evaluation and analysis code; and cleaning and reformatting dataset metadata (scope filtering, split assignment and provenance joins), without altering any audio. We have not used generative AI tools to generate data: the 7.1.4 corpus is produced by the upmixing model of Section 3.2, and the listening tests were designed and conducted by the authors. Theoretical models, mathematical propositions and proofs, and qualitative or thematic analysis are not applicable to this work. Additionally, we used generative AI tools to polish the writing. We have reviewed all AI-assisted work, and every decision on scope, data, experiments and claims was taken by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The listening tests involved 24 listeners trained in music production, electronic composition and recording, who took part with informed consent under a protocol approved by an institutional ethics review; stimuli were played at a calibrated level of 83 dB SPL, and results are reported only as aggregates. The internal 7.1.4 masters were mixed by musicians of our team and are used with permission, only to train and evaluate the upmixing model; they are not released. The released corpus is derived only from tracks whose licence family and version, author and source could be verified, so that the terms of each source carry over to its adaptation (Section 3.4); each upmix keeps the licence and attribution of its source, and the corpus is a per-track licensed research dataset with conditional redistributability, predominantly restricted to noncommercial use. The spatial organisation of the upmixes follows the practice of the mixers who produced the internal masters, and the music follows the genres of the four catalogues (Section 3.5); models trained on the corpus inherit both.

REPRODUCIBILITY STATEMENT

The construction of the corpus is specified in Section 3.4 and Appendix B, which give the licence, duplication, quality and scope criteria with their thresholds, the number of tracks excluded for each reason and the split procedure. The supplementary material gives, for each of the 125,557 tracks, its split, its selected audio asset and the licence and attribution it carries, together with a 30-s excerpt of one training track with its stereo source and twelve 7.1.4 channels, and a verifier that checks the archive against its checksums; the full corpus will be released after acceptance (Section 3.5). The upmixing model and its evaluation are described in Sections 3.2–3.3 and Appendix A. Appendix C gives the architecture of the NSFC and of its postfilter, the quantiser, the losses, the optimiser, the training budget and the seed; Appendix D the test population, the reference codec, the metrics, the statistical procedure with its resampling counts and seeds, the pre-specified test families and the configuration of per-channel Opus; and Appendix E the sample selection, the controlled conditions, the read-outs, the error measures and the statistics of the disentanglement tests. Code for training and running the NSFC and NSFC-Flow is available at `https://anonymous.4open.science/r/nsfc-review-327B`.

REFERENCES

Pablo Alonso-Jiménez, Xavier Serra, and Dmitry Bogdanov. Music representation learning based on editorial metadata from Discogs. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*, pp. 825–833, Bengaluru, India, December 2022. URL `https://ismir2022program.ismir.net/poster_243.html`.

Xiaoyu Bie, Xubo Liu, and Gaël Richard. Learning source disentanglement in neural audio codec. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, Hyderabad, India, April 2025. doi: 10.1109/ICASSP49660.2025.10888065.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Work-*

shop, International Conference on Machine Learning (ICML 2019), Long Beach, CA, USA, 2019. URL http://hdl.handle.net/10230/42015.

Andrew Bourbon, Matthew Watson, Dawid Stelmach, Michael Champion, Ansh Chomal, Steven Fenton, and Hyunkook Lee. An investigation into immersive contemporary popular music mixing practice through analysis of released music. In *AES 2023 International Conference on Spatial and Immersive Audio*, Huddersfield, UK, August 2023. URL https://aes.org/publications/elibrary-page/?id=22160. Paper 58.

Dimitrios Bralios, Paris Smaragdis, and Jonah Casebeer. Learning to upsample and upmix audio in the latent domain. In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, Tahoe City, CA, USA, October 2025. doi: 10.1109/WASPAA66052.2025.11230977.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623, Vienna, Austria, July 2024. PMLR. URL https://proceedings.mlr.press/v235/bruce24a.html.

Stefano Catena. Understanding and analysing 'organised space'. *Organised Sound*, 30(2):108–115, August 2025. doi: 10.1017/S1355771825100423.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-visual navigation in 3D environments. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, volume 12351 of *Lecture Notes in Computer Science*, pp. 17–36. Springer, Cham, 2020. doi: 10.1007/978-3-030-58539-6_2.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, pp. 316–323, Suzhou, China, October 2017.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=ivCd8z8zR2.

Christopher Dewey, Austin Moore, and Hyunkook Lee. 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, July 2024. doi: 10.17743/jaes.2022.0153.

ETSI. 5G; subjective test methodologies for the evaluation of immersive audio systems (3GPP TS 26.259 version 17.0.0 release 17). Technical Specification ETSI TS 126 259 V17.0.0, European Telecommunications Standards Institute, Sophia Antipolis, France, May 2022. URL https://www.etsi.org/deliver/etsi_ts/126200_126299/126259/17.00.00_60/ts_126259v170000p.pdf.

Christof Faller and Frank Baumgarte. Binaural cue coding-part II: Schemes and applications. *IEEE Transactions on Speech and Audio Processing*, 11(6):520–531, November 2003. doi: 10.1109/TSA.2003.818108.

Henrik Frisk and Jan Schacher. Experience of space in sound: Perceptions, concepts and methods. *Organised Sound*, 30(2):98–107, August 2025. doi: 10.1017/S1355771825100678.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting Fréchet audio distance for generative music evaluation. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1331–1335, Seoul, Korea, April 2024. doi: 10.1109/ICASSP48485.2024.10446663.

Wenxiang Guo, Changhao Pan, Zhiyuan Zhu, Xintong Hu, Yu Zhang, Li Tang, Rui Yang, Han Wang, Zongbao Zhang, Yuhan Wang, Yixuan Chen, Hankun Xu, Ke Xu, PengFei Fan, ZheTao Chen, Yanhao Yu, Qiange Huang, Fei Wu, and Zhou Zhao. MRSAudio: A large-scale multimodal recorded spatial audio dataset with refined annotations. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d6ce4d1c760694b4cd9879710f9e7cd0-Abstract-Datasets_and_Benchmarks_Track.html.

Jürgen Herre, Kristofer Kjörling, Jeroen Breebaart, Christof Faller, Sascha Disch, Heiko Purnhagen, Jeroen Koppens, Johannes Hilpert, Jonas Rödén, Werner Oomen, Karsten Linzmeier, and Kok Seng Chong. MPEG Surround—the ISO/MPEG standard for efficient and compatible multichannel audio coding. *Journal of the Audio Engineering Society*, 56(11):932–955, 2008. URL https://aes.org/publications/elibrary-page/?id=14643.

Jürgen Herre, Johannes Hilpert, Achim Kuntz, and Jan Plogsties. MPEG-H 3D Audio—the new standard for coding of immersive spatial audio. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):770–779, August 2015. doi: 10.1109/JSTSP.2015.2411578.

Mojtaba Heydari, Mehrez Souden, Bruno Conejo, and Joshua Atkins. ImmerseDiffusion: A generative spatial audio latent diffusion model. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, Hyderabad, India, April 2025. doi: 10.1109/ICASSP49660.2025.10889311.

ITU-R. Recommendation ITU-R BS.2051-3: Advanced sound system for programme production, May 2022. URL https://www.itu.int/rec/R-REC-BS.2051-3-202205-I/en.

ITU-R. Recommendation ITU-R BS.1770-5: Algorithms to measure audio programme loudness and true-peak audio level, November 2023. URL https://www.itu.int/rec/R-REC-BS.1770-5-202311-I/en.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of Interspeech 2019*, pp. 2350–2354, Graz, Austria, September 2019. doi: 10.21437/Interspeech.2019-2219.

David Kim-Boyle. 3D scores – affordances and applications. *Organised Sound*, 30(3):242–256, December 2025. doi: 10.1017/S135577182510068X.

Shoichi Koyama, Enzo De Sena, Prasanga Samarasinghe, Mark R. P. Thomas, and Fabio Antonacci. Past, present, and future of spatial audio and room acoustics. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, Hyderabad, India, April 2025. doi: 10.1109/ICASSP49660.2025.10890366.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/58d0e78cf042af5876e12661087bea12-Abstract-Conference.html.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR – half-baked or well done? In *ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 626–630, Brighton, UK, May 2019. doi: 10.1109/ICASSP.2019.8683855.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Disheng Liu, Tuo Liang, Zhe Hu, Jierui Peng, Yiren Lu, Yi Xu, Yun Fu, and Yu Yin. Spatial intelligence in vision-language models: A comprehensive survey. *Artificial Intelligence Review*, August 2026. doi: 10.1007/s10462-026-11671-x.

Huadai Liu, Tianyi Luo, Kaicheng Luo, Qikai Jiang, Peiwen Sun, Jialei Wang, Rongjie Huang, Qian Chen, Wen Wang, Xiangtai Li, Shiliang Zhang, Zhijie Yan, Zhou Zhao, and Wei Xue. OmniAudio: Generating spatial audio from 360-degree video. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39060–39084. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25as.html.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, Long Beach, CA, USA, June 2019. PMLR. URL https://proceedings.mlr.press/v97/locatello19a.html.

Wei-Tsung Lu, Ju-Chiang Wang, Qiuqiang Kong, and Yun-Ning Hung. Music source separation with band-split RoPE transformer. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 481–485, Seoul, Korea, April 2024. doi: 10.1109/ICASSP48485.2024.10446843.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

Thomas McKenzie, Eloi Moliner, and Alec Wright. AlbumDB: A multitrack dataset of a ten-song album with stereo and immersive 7.1.4 masters. In *Audio Engineering Society Convention 160*, Copenhagen, Denmark, May 2026. URL https://aes.org/publications/elibrary-page/?id=23185. Paper 10278.

Gerardo Meza, Mariana Sepúlveda, Adrian Roman, Jorge Rodrigo Sigal Sefchovich, and Iran R. Roman. MULTIVOX: A spatial audio-visual dataset of singing groups. Dataset, Zenodo, version 1, September 2025. URL https://doi.org/10.5281/zenodo.17058101.

Bartłomiej Mróz. A seven-year corpus of higher-order Ambisonics recordings. Dataset, Bridge of Data (Most Danych), Gdańsk University of Technology, 2026. URL https://doi.org/10.34808/W8BX-2094.

NUGEN Audio. Halo upmix. Audio plug-in, version 1.7.1.5, 2026. URL https://nugenaudio.com/haloupmix/.

Seungryeol Paik and Kyogu Lee. AMBISONIC-DML: Higher-order Ambisonic music dataset for spatial AI generation. In *NeurIPS 2025 Workshop on AI for Music: Where Creativity Meets Computation*, San Diego, CA, USA, December 2025. URL https://openreview.net/forum?id=sPMMQy76WD.

Alexander Poets, Stephan Preihs, and Jürgen Peissig. On the evaluation of perceived spatial immersion in the application of automatic upmixing for 3D surround sound systems. In *Fortschritte der Akustik – DAGA 2023*, Hamburg, Germany, March 2023. Deutsche Gesellschaft für Akustik (DEGA).

Heiko Purnhagen, Toni Hirvonen, Lars Villemoes, Jonas Samuelsson, and Janusz Klejsa. Immersive audio delivery using joint object coding. In *Audio Engineering Society Convention 140*, Paris, France, May 2016. URL https://aes.org/publications/elibrary-page/?id=18285. Paper 9587.

Anton Ratnarajah, Shi-Xiong Zhang, and Dong Yu. BANC: Towards efficient binaural audio neural codec for overlapping speech. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, Hyderabad, India, April 2025. doi: 10.1109/ICASSP49660.2025.10888151.

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji. STARSS23: An audio-visual dataset of

spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6c9671ed3b3106b71cafda3ba225c1a-Abstract-Datasets_and_Benchmarks.html.

Joren Six and Marc Leman. Panako – a scalable acoustic fingerprinting system handling time-scale and pitch modification. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pp. 259–264, Taipei, Taiwan, October 2014. URL https://archives.ismir.net/ismir2014/paper/000122.pdf.

Jean-Marc Valin, Koen Vos, and Timothy B. Terriberry. Definition of the Opus audio codec. Request for Comments 6716, Internet Engineering Task Force, September 2012.

Karn N. Watcharasupat and Alexander Lerch. Quantifying spatial audio quality impairment. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 746–750, Seoul, Korea, April 2024. doi: 10.1109/ICASSP48485.2024.10447947.

Simon Welker, Matthew Le, Ricky T. Q. Chen, Wei-Ning Hsu, Timo Gerkmann, Alexander Richard, and Yi-Chiao Wu. FlowDec: A flow-based full-band general audio codec with high perceptual quality. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=uxDFlPGRLX.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, Rhodes Island, Greece, June 2023. doi: 10.1109/ICASSP49357.2023.10095969.

Siyi Xie, Hanxin Zhu, Xinyi Chen, Tianyu He, Xin Li, and Zhibo Chen. Sonic4D: Spatial audio generation for immersive 4D scene exploration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(13):11087–11095, March 2026. doi: 10.1609/aaai.v40i13.38087.

Zhongweiyang Xu, Yong Xu, Vinay Kothapally, Heming Wang, Muqiao Yang, and Dong Yu. SPATIALCODEC: Neural spatial speech coding. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1131–1135, Seoul, Korea, April 2024. doi: 10.1109/ICASSP48485.2024.10445853.

Haici Yang, Sanna Wager, Spencer Russell, Mike Luo, Minje Kim, and Wontak Kim. Upmixing via style transfer: A variational autoencoder for disentangling spatial images and musical content. In *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 426–430, Singapore, May 2022. doi: 10.1109/ICASSP43922.2022.9746978.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, Nashville, TN, USA, June 2025. doi: 10.1109/CVPR52734.2025.00994.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. doi: 10.1109/TASLP.2021.3129994.

Suqi Zhang, Zheqi Dai, Yongyi Zang, Yin Cao, and Qiuqiang Kong. DiffStereo: End-to-end mono-to-stereo audio generation with diffusion transformer. In *Proceedings of Interspeech 2025*, pp. 3150–3154, Rotterdam, The Netherlands, August 2025. doi: 10.21437/Interspeech.2025-1850.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. BAT: Learning to reason about spatial sounds with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61454–61469, Vienna, Austria, July 2024. PMLR. URL https://proceedings.mlr.press/v235/zheng24i.html.

A Internal Collection and Upmixing Model Details

Internal collection. Track durations range from 60.48 s to 1,130.69 s, with a mean of 207.43 s (standard deviation 75.47 s) and a median of 201.50 s; 84.85% of the tracks last 2 to 5 min and 7.60% exceed 5 min. Table 3 gives the split, and Table 4 the genre groups, which were annotated by hand with one group per track. Figure 5 (Appendix B) places the collection in the style embedding of the corpus.

Table 3: Split of the internal collection.

Split	Tracks	Hours	Mean duration (s)
Train	1,600	92.36	207.80
Validation	200	11.52	207.39
Test	200	11.36	204.52
Total	2,000	115.24	207.43

Table 4: Genre groups of the internal collection (manual annotation).

Genre group	Tracks	%	Genre group	Tracks	%
Electronic / Dance	715	35.75	Christian / Gospel	30	1.50
Pop	281	14.05	Soundtrack / Stage	30	1.50
Hip-Hop / Rap	174	8.70	Holiday / Children's / Comedy	23	1.15
New Age / Easy Listening / Instrumental	150	7.50	Jazz	17	0.85
Rock / Alternative	131	6.55	World / Regional	16	0.80
Classical / Crossover	126	6.30	Folk / Singer-Songwriter	14	0.70
R&B / Soul / Funk	103	5.15	Reggae	2	0.10
Country	102	5.10	Vocal	1	0.05
Latin / Brazilian	85	4.25			

Architecture. The input is the complex STFT of the two stereo channels (48 kHz, 2048-sample Hann window, hop 480, 1025 bins). The bins are grouped into 60 overlapping bands whose supports follow a mel filter bank; every bin belongs to at least one band. Each band is normalised and projected linearly to 256 dimensions. Each of the six blocks contains one transformer layer across time and one across bands (8 heads of dimension 64, feed-forward expansion 4, gated attention, rotary position embeddings, no dropout), and the outputs of earlier blocks are added to the input of later ones. For each of the twelve output channels, a per-band two-layer network with a gated output estimates the real and imaginary parts of the mask; masks of overlapping bands are averaged per bin. The mask is applied to the STFT of $(s_L + s_R)/2$, the DC bin is set to zero, and the channel is resynthesised by the inverse STFT. Content that is exactly antiphase between the two input channels vanishes from this carrier and cannot be placed.

Objective and optimisation. The loss combines waveform and complex-spectrum reconstruction on every channel with a multi-resolution STFT loss (FFT sizes 1024, 2048 and 4096 with hops of 147, 256 and 512 samples, 128 mel bins, perceptual weighting, log-magnitude and spectral-convergence terms, weight 1.0). Two penalties act on individual channels: one on outputs that are silent where the target is not (weight 0.02), and one on outputs whose RMS envelope, in 100-ms frames, falls below 0.65 of the target's (weight 0.06). Training uses 12-s segments, Adam with mixed precision, and an exponential moving average of the weights with momentum 0.999.

Inference. Every track is processed in 12-s windows with eightfold overlap and written as twelve 48-kHz floating-point channels. One checkpoint, identified by its SHA-256 hash in the dataset metadata, produced all tracks of the four catalogues.

Objective evaluation. The metrics are those of Appendix D, computed against the human master on the 200 test tracks, with two differences in aggregation: LSD and the multi-resolution mel distance (MR-Mel) are averaged over all twelve channels, and FAD is computed on each non-LFE

channel as a mono signal and averaged over the eleven channels. SSR, SRR and the RTF angle error are computed jointly over the twelve channels; the RTF angle error is 0.42 rad for our model against 0.68 rad for Halo Upmix. SNR and MR-Mel are kept out of the main text because they are close to the waveform and mel-spectral reconstruction losses on which the model is trained. Table 5 gives the per-channel results. The difference is largest in the rear surround channels, where the SI-SDR of Halo Upmix is -9.6 and -9.0 dB against 0.1 and 0.2 dB for the model; Halo Upmix is closer to the human master only in the FAD of the front pair and in the LSD of the LFE and front height channels.

Table 5: Per-channel objective results of our upmixing model (M) and Halo Upmix (H) on the 200 internal test tracks, against the human master. FAD in units of 10^{-3}, not computed for the LFE channel; its mean is over the eleven other channels, all other means over twelve.

		L	R	C	LFE	Ls	Rs	Lrs	Rrs	Ltf	Rtf	Ltr	Rtr	Mean
SI-SDR (dB) ↑	M	12.31	12.17	3.81	5.18	3.96	4.33	0.07	0.22	6.35	6.59	3.25	3.25	5.12
	H	8.65	8.51	−2.78	−1.09	−2.85	−2.70	−9.57	−9.01	1.50	1.81	−3.80	−3.48	−1.23
SNR (dB) ↑	M	11.41	11.31	1.95	2.42	4.73	4.88	2.47	2.49	5.98	6.21	4.47	4.46	5.23
	H	7.58	7.50	−7.41	−0.97	0.91	0.90	0.41	0.42	1.36	1.48	0.87	0.91	1.16
MR-Mel ↓	M	9.13	9.30	11.91	4.82	4.96	5.25	8.63	8.77	4.79	4.79	6.32	6.31	7.08
	H	15.10	15.26	28.58	9.28	10.80	11.19	25.19	25.46	6.03	5.89	20.31	20.25	16.11
LSD (dB) ↓	M	6.02	6.02	18.56	18.44	10.07	10.22	17.78	17.81	12.92	12.98	16.50	16.72	13.67
	H	6.38	6.42	20.73	14.97	11.22	11.31	21.68	21.73	10.15	10.08	18.39	18.40	14.29
FAD ↓	M	8.58	9.14	30.26	–	16.52	16.70	29.79	29.89	14.19	14.14	22.25	22.64	19.47
	H	5.12	5.50	47.85	–	37.08	36.04	153.06	152.80	19.78	18.56	117.63	116.04	64.50

Listening test. The 24 listeners were trained in electronic music production (15), electronic composition (7) and recording (2), and listened on a 7.1.4 loudspeaker system calibrated to 83 dB SPL. Excerpts were cut from the same chorus section in all three versions, with candidates proposed automatically and confirmed by listening, and stored as 48-kHz, 32-bit floating-point 12-channel files with 0.5-s fades. Each stimulus set was normalised to -18 LUFS integrated multichannel loudness with one gain for all channels, which preserves the level relations between channels. The order of the three versions within a track followed a pre-generated randomisation, and all listeners followed the same playback plan; a practice trial on a separate track preceded the test and was not analysed. All listeners gave informed consent. The commercial upmixer was Halo Upmix version 1.7.1.5 with the Music 3D preset.

Single attributes. Table 6 gives the MOS of the six attributes and the 95% confidence intervals of the two domain means from a listener-by-track bootstrap.

Table 6: MOS of the six attributes in the listening test of the upmixing model. Bold: the higher of the two upmixers.

	Human master	Our model	Halo Upmix
Quality MOS	3.61 [3.42, 3.81]	**3.44** [3.25, 3.64]	3.39 [3.21, 3.57]
Spectral balance	3.49	3.35	**3.40**
Timbral fullness	3.72	**3.50**	3.39
Instrument transients	3.62	**3.46**	3.38
Spatial MOS	3.61 [3.44, 3.79]	**3.41** [3.21, 3.62]	3.28 [3.09, 3.46]
Sound-field balance	3.53	**3.38**	3.36
Localisation clarity	3.75	**3.50**	3.29
Envelopment	3.54	**3.35**	3.18

Paired differences. Table 7 reports the mean paired difference between versions of the same track rated by the same listener. Intervals are 95% percentile intervals from 10,000 bootstrap resamples over listeners and tracks (seed 20260924). The four contrasts were fixed before any paired result was computed; each p-value is the larger of two sign-flip permutation tests (99,999 resamples), one

over tracks and one over listeners, and is Holm-corrected over the four contrasts. The analysis was defined after the test.

Table 7: Paired MOS differences in the listening test of the upmixing model. Position-adjusted: estimated with track fixed effects and playback position.

	Quality MOS		Spatial MOS	
Contrast	Δ [95% CI]	p_{Holm}	Δ [95% CI]	p_{Holm}
Our model − Halo Upmix	$+0.05$ [-0.10, 0.21]	0.42	$+0.13$ [-0.05, 0.33]	0.15
Our model − human master	-0.17 [-0.31, -0.04]	0.007	-0.20 [-0.36, -0.05]	0.005
Position-adjusted				
Our model − Halo Upmix	$+0.07$ [-0.07, 0.22]	–	$+0.17$ [0.00, 0.35]	–
Our model − human master	-0.15 [-0.29, -0.01]	–	-0.15 [-0.30, -0.01]	–

Playback order. Versions played first in a track were rated lower on average (quality MOS 3.41 against 3.50 and 3.53 in the second and third positions; spatial MOS 3.33 against 3.47 and 3.50). Under the fixed playback plan our model was played first on 16 of the 30 tracks, against 9 for Halo Upmix and 5 for the human master, so the unadjusted comparison is, if anything, biased against the model. Among the single attributes, the paired difference to Halo Upmix is largest for localisation clarity ($+0.21$ [0.01, 0.42], uncorrected).

B Corpus Refinement Details

Populations. Refinement acts on three nested populations. The candidate population is every stereo track of the four catalogues together with its 7.1.4 upmix, 157,627 pairs, each copied with a verified content hash. Merging the licence, duplicate, and technical-quality and content decisions retains 128,233 of them. The task scope of 30 s to 15 min then gives the corpus of 125,557 tracks and 8,392.55 hours (Table 8). Every 7.1.4 output has twelve 48 kHz single-channel floating-point files in the order L, R, C, LFE, Ls, Rs, Lrs, Rrs, Ltf, Rtf, Ltr, Rtr, and each track is identified by the content hash of its source and by one hash per decoded output channel.

Table 8: Tracks per source catalogue before refinement and in the final corpus.

Catalogue	Candidates	Corpus tracks	Corpus hours
ccMixter	29,957	24,850	1,658.49
Dogmazic	30,904	20,283	1,408.33
FMA	60,185	48,024	3,205.63
MTG-Jamendo	36,581	32,400	2,120.11
Total	157,627	125,557	8,392.55

Decision rule. Each examination records every reason that applies to a track, and a track is excluded if any reason applies. Table 9 lists the reasons that exclude tracks from the candidate population. Because a track may carry several reasons, the per-reason counts overlap: 34,040 reason memberships fall on 29,394 excluded tracks, and the rows of the table do not sum to the number excluded. The class column separates leakage control, which follows from the definition of a non-duplicated corpus, from licence governance and from curation and scope rules, which are project decisions about what the corpus should contain. Three tracks whose source and output are both digitally silent are excluded by other reasons as well and do not change the population.

Licences. Table 10 gives the licence families of the candidates and of the corpus. Each track keeps its title, author, source URL, licence name, family and version, and the non-commercial and share-alike flags, from which a per-track attribution is generated. Licence examination excludes two groups: remixes on ccMixter whose upstream sources are undeclared or form a cycle, so that the

Table 9: Exclusion reasons applied to the 157,627 candidates. Counts are memberships and overlap (34,040 memberships on 29,394 excluded tracks). The last two rows are the task scope applied afterwards to the 128,233 retained tracks. †A historical minimum-duration rule, superseded by the 30 s scope.

Reason	Class	Tracks
Duplicate (non-representative) or incomplete duplicate search	leakage	6,670
ccMixter remix with undeclared or cyclic upstream sources	licence	2,430
Dogmazic Free Art License without a stated version	licence	5,403
Source bit rate, reported or effective, below 128 kbps	curation	4,014
Source sample rate below 44.1 kHz	curation	924
Mono source	curation	1,019
Two-channel source that is false stereo	curation	5,368
Discogs-EffNet non-music activation above 0.1	curation	3,163
Output overload, sustained ($\geq$1 s) or above +1 dBFS	curation	4,764
Downmix consistency in the joint lower tail	curation	150
Low spatial dimension in the joint tail	curation	35
Duration below 5 s†	scope	100
Duration below 30 s	scope	961
Duration above 15 min	scope	1,715

attribution and licence terms of the sampled works cannot be established, and all Free Art License tracks on Dogmazic, none of which states a version, so that the share-alike obligation cannot be honoured. No track in the corpus carries a no-derivatives term. The corpus is a per-track licensed research dataset with conditional redistributability, predominantly restricted to noncommercial use: redistribution follows the terms of each track, not a single licence for the whole corpus, and the platform licence metadata on which it rests is not a clearance of recording, performance or sample rights.

Table 10: Licence families of the candidate population and of the corpus.

Licence family	Candidates	Corpus
CC0	761	635
CC BY	20,311	16,449
CC BY-SA	14,851	12,976
CC BY-NC	38,809	32,176
CC BY-NC-SA	77,492	63,321
Free Art License	5,403	0
Total	157,627	125,557

Duplicates. Two tracks are duplicates when they carry the same complete musical content: the same performance and structure, with differences only in encoding, bit rate, sample rate, container or metadata, gain, equalisation or limited dynamics processing, leading and trailing silence or trimming, medium (digital master or an analogue transfer), transmission or re-recording that preserves the whole content, and constant or slowly varying speed and pitch. Covers, alternate performances and takes, remixes that change structure or arrangement, instrumental or vocal versions, DJ mixes and mashups, excerpts, and tracks that share only an introduction, sample or loop are not duplicates. Byte-identical files are grouped by SHA-256. Acoustic candidates come from Panako (Six & Leman, 2014) queries at 10, 30, 50, 70 and 90% of each track's length. A pair enters a duplicate group when at least four of the five anchors match the same reference, the matches cover both the early and the late part of the track, the mapping from query to reference position is monotonic, the time-scale and frequency factors agree across anchors, and any difference in duration is explained by speed change, silence or trimming. Relations that fall short of this (one track containing another, a shared excerpt, or inconsistent anchors) do not remove either track; both are confined to the training split,

which covers 24,378 tracks of the corpus. The search formed 5,836 duplicate groups with 12,447 members and excluded the 6,611 non-representative members. Tracks whose search coverage was incomplete (55) and four tracks with invalid audio found before the search are excluded as well, for 6,670 exclusions in total. The ten stereo masters of AlbumDB were included in the search as held-out queries, and no candidate matches any of them.

Technical quality and content. All checks read each track once, in full. Structural integrity comprises complete decoding of the source and of all twelve output channels, channel identity against the frozen layout, equal frame counts across the source and the output channels, non-finite values (none were found), content hashes against the copy manifest, and output collapse, meaning silence in all twelve channels for an audible source. Level and loudness comprise the true peak following Annex 2 of ITU-R BS.1770-5 (ITU-R, 2023) and the loudness of the native 7.1.4 layout following its Annex 3, DC offset, and near-full-scale clipping plateaus; these native twelve-channel measurements are quality controls for research use, not a test of compliance with any delivery specification. Inter-channel structure comprises the correlation between channel pairs, the correlation and side-to-mid ratio of the stereo source (a false-stereo source has high correlation and a negligible side signal), the effective rank and the top-two energy share of the 12×12 spatial covariance matrix, per track and per window, the energy shares of the height, LFE and non-front channels, and windowed channel activity. Consistency with the source compares a fixed 12-to-2 downmix of the output with the source stereo by correlation, SI-SDR and NMSE. Transient risk is measured by sample jumps, waveform curvature and click events. Source properties are the reported and effective bit rate, the sample rate and the channel count. Content is assessed by the Discogs-EffNet model (Alonso-Jiménez et al., 2022) applied to every whole source track at 16 kHz mono, with class activations averaged over all patches; the non-music activation is a model output, not a human label.

Of these checks, the curation rules in Table 9 exclude tracks; the two empirical tail rules take the lower 0.1% tail of the population and require two metrics at once, downmix correlation below 0.858 together with downmix SI-SDR below 5.02 dB, and covariance effective rank below 1.038 together with a top-two energy share above 0.9989. All other measurements are kept as per-track statistics. Overload is judged on output samples at or beyond full scale: a track is excluded when, in its worst channel, such samples last 1 s or more in total or the sample peak exceeds +1 dBFS. In 4,746 of the 4,764 tracks excluded in this way the source stereo already exceeds full scale, so the overload is largely inherited from the source. Tracks with briefer and smaller overs are kept; for each of the 14,610 such retained tracks, one static gain is applied to all twelve output channels and to the stereo source alike, bringing the output true peak to 0 dBTP (the largest residual is 2.5×10^{-6} dB) while preserving the level relations between channels. The unprocessed floating-point outputs are kept, and the corrected versions are stored separately with their own hashes; 14,446 of them fall within the task scope, and the corpus uses the corrected version for each.

Scope and split. The task scope keeps tracks from 30 s to 15 min inclusive, which removes 961 shorter tracks (4.78 h) and 1,715 longer ones (761.57 h). The lower bound guarantees every track at least two complete 12 s upmixing segments and seven 4 s codec segments; the upper bound limits the long tail of durations. The split unit is the complete track, and the split obeys four hard constraints: every track in the scope appears exactly once, duplicate groups are never divided, tracks confined to training by a partial relation are placed in the training split, and AlbumDB lies outside all splits. Validation and test are sized to about 12 and 24 hours and stratified by the hours available from each catalogue; six further attributes (duration bin, licence family, Discogs-EffNet top-level genre family, non-music activation band, metadata action and source codec) are balanced by deterministic marginal-deficit sampling. The largest deviation of any balanced attribute from its corpus proportion is 1.9 percentage points in validation and 0.7 in test. Artist and album are not grouping keys, because the artist strings of the catalogues are not canonicalised and album identity is incomplete; the test split therefore measures generalisation to unseen tracks rather than to unseen artists. Table 11 gives the resulting split.

Corpus statistics. Track durations in the corpus range from 30.00 s to 900.00 s, with a mean of 240.63 s (standard deviation 118.53 s) and a median of 224.47 s; 66.20% of the tracks last 2 to 5 min and 22.19% exceed 5 min. The catalogues contribute 38.25% (FMA), 25.81% (MTG-Jamendo), 19.79% (ccMixter) and 16.15% (Dogmazic) of the tracks. Genre is taken from the genre metadata of the source site for 94,823 tracks (75.52%; 65.6% carry one of the 27 terms shared by the source sites,

Table 11: The frozen split. Segment counts are complete 4 s codec segments and 12 s upmixing segments.

Split	Tracks	Hours	4 s segments	12 s segments
Train	125,011	8,356.55	7,462,305	2,446,427
Validation	183	12.00	10,701	3,500
Test	363	24.00	21,427	7,011
Total	125,557	8,392.55	7,494,433	2,456,938

Figure 2(c)); the remaining 30,734 tracks carry no specific genre label and are assigned the top-level family predicted by Discogs-EffNet. Table 12 gives the resulting groups. Figure 5 shows the corpus and the internal collection in one embedding of their Discogs-EffNet style activations. The 400 activations of each track are centred and reduced to 50 dimensions by principal component analysis fitted on the corpus, without whitening, and projected to two dimensions by UMAP (McInnes et al., 2018) fitted on all 125,557 tracks of the corpus (30 neighbours, minimum distance 0.15); the internal collection is mapped into the same projection without refitting. The input is the stereo track for the corpus and the stereo downmix of the 7.1.4 master for the internal collection.

Table 12: Genre groups of the corpus by label source.

Genre group	Source labels	Predicted	Total	%
Electronic / Dance	33,360	18,054	51,414	40.95
Rock / Alternative	16,509	5,124	21,633	17.23
Experimental	10,511	0	10,511	8.37
Pop	7,669	590	8,259	6.58
Hip-Hop / Rap	3,635	3,158	6,793	5.41
New Age / Easy Listening / Instrumental	4,519	0	4,519	3.60
Classical / Crossover	4,141	270	4,411	3.51
Folk / Singer-Songwriter	3,961	0	3,961	3.15
Jazz	1,886	472	2,358	1.88
Blues	1,558	365	1,923	1.53
R&B / Soul / Funk	1,082	564	1,646	1.31
World / Regional	1,616	0	1,616	1.29
Folk / World / Country	0	1,550	1,550	1.23
Reggae	1,183	262	1,445	1.15
Other native descriptors	1,273	0	1,273	1.01
Soundtrack / Stage	658	104	762	0.61
Country	523	0	523	0.42
Spoken / Non-Music	339	52	391	0.31
Latin / Brazilian	262	118	380	0.30
Historic	79	0	79	0.06
Holiday / Children's / Comedy	15	44	59	0.05
Christian / Gospel	28	0	28	0.02
Brass / Military	16	7	23	0.02
Total	94,823	30,734	125,557	100.00

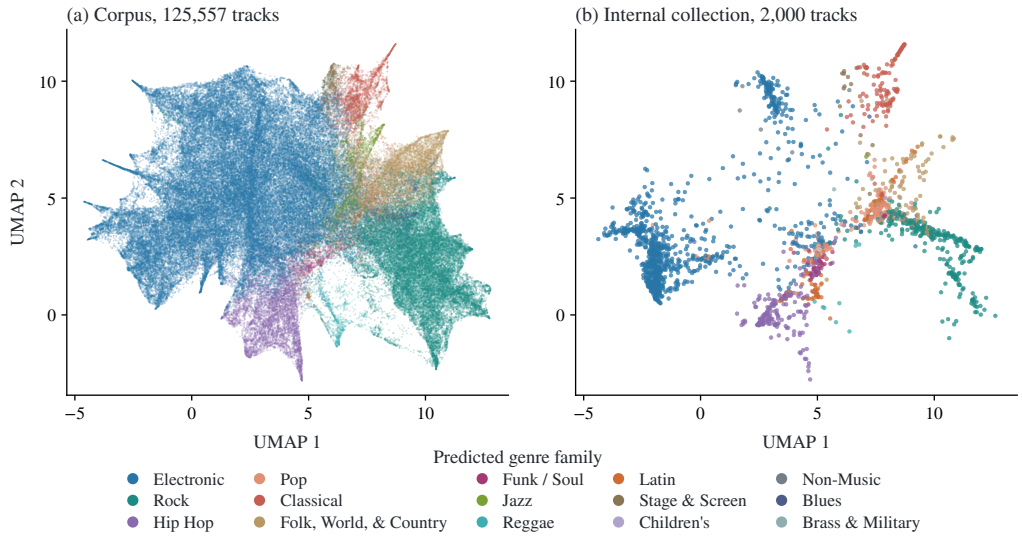


Figure 5: Discogs-EffNet style embedding of (a) the corpus and (b) the internal collection, one point per track, in one UMAP projection fitted on the corpus. Colour is the top-level family of the style with the highest mean activation, a model prediction rather than a calibrated probability, so it need not agree with the manual labels of Table 4 or the source-site labels of Table 12.

C NSFC Architecture and Training

Reference. Channels are ordered L, R, C, LFE, Ls, Rs, Lrs, Rrs, Ltf, Rtf, Ltr, Rtr. The mono reference is $m = w^\top x$ with $w = (0.707, 0.707, 1, 0, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.354, 0.354)$, not normalised by the sum of the weights.

Encoder input. The STFT uses a 1280-sample Hann window and a hop of 640 samples, giving 641 frequency bins and 75 frames per second at 48 kHz. For the thirteen spectra Z_i of $[m; x]$ with powers $P_i = |Z_i|^2$, the normalised spatial covariance is

$$C_{ij} = \frac{Z_i Z_j^*}{\sqrt{P_i P_j} + \epsilon} \; (i \neq j), \qquad C_{ii} = \log(P_i + \epsilon), \tag{4}$$

with $\epsilon = 10^{-8}$ and $C_{ij} = 0$ when $\sqrt{P_i P_j} < 10^{-10}$. The 91 upper-triangular entries, including the 13 real diagonal entries whose imaginary parts are zero, are split into real and imaginary parts and concatenated with the reference spectrum $\log(1 + |M|)\, e^{\mathrm{i}\angle M}$, giving 184 input channels per time–frequency point. The raw covariance used for speech arrays (Xu et al., 2024) is dominated by absolute energy; the normalisation separates the relations between channels from how loud the music is, which matters for the sparse height and LFE channels.

Encoder and decoder. The encoder has six stages, each a two-dimensional convolution that strides along frequency followed by two parallel residual blocks with group normalisation; channel widths grow from 64 to 256, and the frequency axis is reduced from 641 bins to 16, which are flattened into a 4096-dimensional vector per frame. The decoder mirrors the encoder with transposed convolutions and ends in a 5×5 convolution that outputs the $12 \times 3 \times 9$ complex filter taps of Eq. equation 2 at every time–frequency point. The mask of Eq. equation 3 is a 1×1 convolution without bias on the last decoder features, initialised at zero, with two output maps per channel whose $0.25 \tanh$ gives the real and the imaginary part of Δ_c.

Quantiser. Residual vector quantisation with eight stages, each with a codebook of 1024 entries in an eight-dimensional L2-normalised space entered and left through learned projections, trained with codebook and commitment losses (weights 1.0 and 0.25) and a straight-through estimator as in

DAC (Kumar et al., 2023); codes unused within a batch are re-initialised from encoder outputs. The rate is $8 \times \log_2 1024 \times 75 = 6000$ bit/s.

Objective and optimisation. The loss is the sum of waveform L1 and L2 terms, a multi-resolution STFT loss with spectral convergence and log-magnitude terms at FFT sizes 512, 1024 and 2048 (weight 0.5), and the quantisation losses. Training uses Adam ($\beta_1 = 0.5$, $\beta_2 = 0.9$), a learning rate of 10^{-4} with exponential decay, 4-s segments, a global batch of 64 on eight GPUs and seed 1234. The checkpoint is selected on the validation split within the first 100k updates; the validation split is traversed in full (183 tracks, 10,701 segments).

Why the mask is bounded. An unbounded version of the mask, with a bias term, was trained first. After 10k updates its bias alone reproduced a static gain per channel, the encoder latent RMS had fallen by 6.9 dB relative to the codec without mask, and only 22% of the codebook entries remained in use, against all of them without the mask: the model had learned to shrink the latents and let the mask restore the level, bypassing the spatial stream. The bound $0.25 \tanh(\cdot)$ and the removal of the bias close this path.

NSFC-Flow postfilter. The postfilter follows FlowDec (Welker et al., 2025). It operates on $\Phi(Y) = \beta |Y|^{\alpha} e^{\mathrm{i}\angle Y}$ with $\alpha = 0.3$ and $\beta = 0.33$, using the STFT above and a shared per-example peak normalisation derived from the NSFC output. Training draws a flow time $u \sim \mathcal{U}(0, 1)$, starts from $z_0 = \Phi(\hat{Y}) + \sigma(f)\,\varepsilon$ with complex Gaussian ε, and regresses the direct vector to $\Phi(Y)$ with a complex squared error. The noise scale $\sigma(f)$ is the median, over 64 training segments, of the per-frequency residual RMS of the NSFC, smoothed along frequency and rescaled to a global anchor; one curve is shared by all channels and by the real and imaginary parts. The network is a convolutional U-Net without attention (widths 256, 416 and 544; time-conditioned residual blocks; 22.6M parameters) that takes the current state and the NSFC spectrum of all twelve channels as 48 real input channels and outputs 24. It is trained for 60k updates with Adam at 10^{-4} and an exponential moving average of the weights (decay 0.999), which is used for evaluation. Inference integrates with the midpoint rule at four network evaluations from a single draw.

D NSFC Evaluation Details

Test population. All systems decode the same 363 test tracks, cut into 21,427 non-overlapping 4-s segments at 48 kHz. The reference codec is the official 44.1-kHz DAC checkpoint with nine codebooks, with resampling to and from 48 kHz. Postfilter outputs use a single draw with a fixed seed per track; no best-of-K selection is made.

Metrics. SI-SDR and SNR are computed over the twelve channels. SSR, SRR and the relative-transfer-function (RTF) angle error follow SPAUQ (Watcharasupat & Lerch, 2024) on all twelve channels; the eleven-channel variants without LFE are reported only as a sensitivity analysis. LSD and the multi-resolution mel distance (MR-Mel) are computed per channel and averaged over the non-LFE channels. FAD uses CLAP music embeddings (Wu et al., 2023) through the fadtk toolkit (Gui et al., 2024) on a mono downmix; it is a set-level distance and its confidence intervals are obtained by resampling tracks and recomputing the distance.

Statistics. Point estimates weight valid segments equally. Differences are tested one-sided with a paired cluster bootstrap over tracks, centred on the null (99,999 resamples, seed 1234, $\alpha = 0.05$); 95% confidence intervals are two-sided percentile intervals from 1,000 track-level resamples. Three families were fixed before the test set was opened: F1, the codec with mask against the codec without mask at 100k and at 160k updates, on SI-SDR and SSR (four tests, Holm-corrected); F2, the complete system against per-channel Opus, which serves as a high-rate reference point rather than a competitor (it codes each channel independently, is not a multichannel spatial codec, and runs at fourteen times our rate), on SI-SDR and SSR (two tests, Holm-corrected; the reference is not reached); F3, NSFC-Flow against NSFC on SSR (one test). All other metrics and comparisons are reported with confidence intervals only.

Table 13: All test-set systems and metrics. Rates in kbps; for uncoded rows only the spatial stream is counted. Parameters in millions, of the trained spatial layer and postfilter; the pretrained DAC is not counted, and Opus has no learned parameters. RTF: angle error in radians. FAD in units of 10^{-3}. Bold: best within each reference block; per-channel Opus is not compared.

Reference	System	Params	Rate	SI-SDR↑	SSR↑	SRR↑	RTF↓	LSD↓	MR-Mel↓	FAD↓	SNR↑
Uncoded	w/o mask, 100k	89.13	6	9.00	12.57	**19.62**	0.323	5.56	6.73	0.17	9.54
	w/o mask, 160k	89.13	6	9.39	13.11	19.24	0.317	**5.23**	6.41	0.14	9.88
	NSFC	89.14	6	**9.58**	13.39	18.95	0.313	5.46	**6.28**	0.15	**10.03**
	NSFC-Flow	111.74	6	8.34	**13.61**	15.63	**0.277**	7.25	7.34	**0.07**	8.79
DAC	NSFC	89.14	13.75	**6.35**	10.85	**10.28**	0.326	**13.47**	**12.85**	**33.94**	**7.20**
	NSFC-Flow	111.74	13.75	5.67	**10.97**	9.56	**0.305**	14.08	13.37	34.45	6.55
Per-channel Opus		–	192	10.91	19.16	10.73	0.400	29.08	10.64	42.84	11.22

Table 14: Paired differences (left minus right), computed before rounding, so they can differ by 0.01 from differences of the rounded values in Tables 2 and 13; 95% confidence intervals below each estimate. FAD in units of 10^{-3}. †Holm-corrected one-sided $p < 10^{-4}$ ($p < 10^{-4}$ uncorrected for the single test in F3); ‡test not supported ($p = 1$).

Comparison	SI-SDR (dB)	SSR (dB)	SRR (dB)	LSD (dB)	FAD
NSFC − w/o mask, 100k	$+0.58^{\dagger}$ [0.55, 0.61]	$+0.82^{\dagger}$ [0.77, 0.89]	-0.67 [−0.76, −0.57]	-0.10 [−0.10, −0.09]	-0.023 [−0.031, −0.018]
NSFC − w/o mask, 160k	$+0.18^{\dagger}$ [0.17, 0.20]	$+0.28^{\dagger}$ [0.25, 0.30]	-0.29 [−0.32, −0.25]	$+0.23$ [0.21, 0.25]	$+0.005$ [0.002, 0.009]
NSFC-Flow − NSFC	-1.24 [−1.30, −1.18]	$+0.22^{\dagger}$ [0.17, 0.26]	-3.32 [−3.43, −3.20]	$+1.79$ [1.76, 1.83]	-0.074 [−0.101, −0.039]
NSFC-Flow − NSFC, DAC	-0.68 [−0.72, −0.64]	$+0.13$ [0.10, 0.16]	-0.73 [−0.78, −0.68]	$+0.61$ [0.58, 0.63]	$+0.51$ [0.21, 0.80]
NSFC + DAC − per-ch. Opus	$-4.57^{\ddagger}$ [−4.80, −4.34]	$-8.31^{\ddagger}$ [−8.81, −7.82]	-0.45 [−0.64, −0.28]	-15.61 [−16.27, −15.04]	-8.91 [−13.23, −4.67]

Training budget. Updates are counted within the checkpoint-selection window. The codec with mask reaches the SI-SDR and SSR of the codec without mask trained for 160k updates with 100k updates, that is 62.5% of the updates; this does not imply a proportional saving in GPU time or total training cost. The postfilter adds 60k updates on the frozen NSFC and is not a comparison at equal total budget.

Rates. Nominal rates follow the codec definitions: 6 kbps for the spatial stream (eight 10-bit codes at 75 Hz) and 7.75 kbps for the DAC-coded reference (nine 10-bit codes at 86.13 Hz), hence 13.75 kbps for the complete system; per-channel Opus runs at 12×16 kbps. Container overhead is not included.

Per-channel Opus. Each channel is coded as an independent mono stream with libopus 1.4 (48 kHz input, 20-ms frames, `application=audio`, hard constant bit rate of 16 kbps); the encoder selects the coded bandwidth, and the twelve decoded streams are trimmed of their pre-skip and re-assembled in the original channel order without any gain or delay correction. This construction replaced a first one, in which the twelve channels formed a single multichannel Opus stream (channel mapping family 255, FFmpeg 6.1.1 with libopus 1.4) at a constant 72 kbps; on real development audio its decoder rejected packets as invalid, as it also did at 72 kbps with variable rate and at a constant 96 kbps, and the outputs of that construction were discarded without being analysed. The cause was not isolated. The rate was fixed before the test set was evaluated as the single operating point and was not tuned on any metric. Opus defines no operating point below 6 kbps per stream (Valin et al., 2012), so twelve independent streams cannot come closer than 72 kbps to the rate of the complete system, over five times its 13.75 kbps; an equal-rate comparison with per-channel Opus does not exist.

Listening test. The listening test used 30 tracks: eight from FMA, four from ccMixter, five from MTG-Jamendo, three from Dogmazic and ten from AlbumDB. Excerpts of 10 s were played at

48 kHz with 0.5-s fades and a common integrated loudness of -18 LUFS applied equally to all channels, over 7.1.4 loudspeakers calibrated to 83 dB SPL. For each track the unrated reference was followed by per-channel Opus and two of the four NSFC systems; each of the six pairs of NSFC systems occurs on five tracks, so every NSFC system is heard on 15 tracks. Two listening passes collected the three quality attributes (spectral balance, timbral fullness, transients) and the three spatial attributes (sound-field balance, image localisation, envelopment) on a five-point scale. MOS values in Table 2 are raw means over 24 listeners; adjusting for track effects changes the domain means by at most 0.06 and leaves the ranking unchanged. The analysis is descriptive and was defined after data collection.

Comparison with per-channel Opus. Per-channel Opus, rated on all 30 tracks, received a quality MOS of 3.35 and a spatial MOS of 3.60; NSFC-Flow with the DAC-coded reference, at about one fourteenth of its rate ($13.75/192$), received 3.26 on both. Opus also retains higher SI-SDR and SSR (Table 14): at fourteen times our rate, the high-rate reference of F2 is not reached, as expected for a stream of 13.75 kbps.

E Disentanglement Test Details

Samples. Validation and training tracks are ordered by a fixed hash of their identifier and taken in that order; each contributes two complete 4-s windows near one third and two thirds of its duration, at least 1 s from either end. Windows were fixed before inference and were not replaced on the basis of model outputs. AlbumDB masters are resampled to 48 kHz before mirroring. Mirroring swaps the five left–right channel pairs and leaves C and LFE unchanged; the reference of the mirrored mix differs from the original by at most 2.4×10^{-7}. The validation split was used for checkpoint selection in Section 4, and the AlbumDB tracks come from a single album by one artist.

Controlled conditions. Let $b \in \mathbb{R}^{12}$ be one on the eleven non-LFE channels and zero on the LFE, and let $e_s \in \mathbb{R}^{12}$ be the standard basis vector of channel $s \in \{\mathrm{Ls}, \mathrm{Rs}, \mathrm{Ltf}, \mathrm{Rtf}\}$. The gain vector emphasising channel s is $a_s = (b + e_s)/6.622$, which satisfies $w^\top a_s = 1$ since $w^\top b = 6.122$ and $w_s = 0.5$; the mix is $a_s c$ for the track's reference c. The emphasised channel thus receives four times the energy of each other non-LFE channel, a target share of $4/14 = 0.286$ against $1/14 = 0.071$. Emphasis levels in dB are stated relative to the share $1/14$, so the target is $10 \log_{10} 4 \approx 6$ dB. Tracks enter only if both windows of the reference exceed -60 dBFS RMS. Table 15 gives the values plotted in Figure 4(a).

Table 15: Energy share of the emphasised channel under the four controlled conditions (validation tracks). The target is 0.286, 6 dB above the 1/14 share of each other non-LFE channel. Drop: codes of the same condition from another track minus codes of another condition, with 95% block-bootstrap confidence interval.

Emphasised channel	Own codes	Same condition, other track	Other condition	Drop
Rs	0.322	0.319	0.221	0.098 [0.094, 0.102]
Ls	0.254	0.252	0.210	0.043 [0.040, 0.046]
Ltf	0.047	0.047	0.036	0.011 [0.010, 0.012]
Rtf	0.059	0.059	0.034	0.025 [0.022, 0.029]

Read-out. Each code sequence is re-embedded through the codebooks and averaged over time. A standardised ridge classifier (squared loss, $\lambda = 1$, unpenalised intercept, closed-form solution, no hyperparameter search) is fitted once on 16 training tracks and evaluated on 32 validation tracks; accuracy is aggregated per track. The same classifier on channel-energy statistics of the mix calibrates the ceiling, and on the mono spectrum and on a constant input serves as negative controls. The mono spectrum and the constant predictor stay at 25% accuracy, and channel-energy statistics of the mix reach 100%.

Oracle amplitude panning. For every non-LFE channel k the gain $g_k = \langle x_k, m\rangle / \langle m, m\rangle$ is fitted to the target segment, and the output is $g_k m$. The gain is unquantised and computed from the target itself.

Error measures. The normalised error of an output $\hat{x}$ against a target x is $\|\hat{x} - x\|_F^2 / \|x\|_F^2$, taken over the non-LFE channels or over all twelve channels as stated. The energy-distribution distance is the total variation distance $\frac{1}{2}\sum_k |\hat{p}_k - p_k|$ between the shares p_k of the eleven non-LFE channels in the total non-LFE energy of target and output.

Transfer. The 32 validation tracks form 16 disjoint pairs in a fixed order, used in both directions and with matching window indices. Each recipient is decoded with its own codes, with the donor's codes under the same condition, and with the donor's codes under another condition taken in a fixed cycle. We report the total variation distance between the non-LFE energy distributions of output and target, and the normalised error over all channels. An analytic transfer of the energy distribution recovers the target exactly in this single-source construction (256 of 256 rows), which verifies the construction. With the donor's codes under the same condition, the energy-distribution distance to the target differs from that with the recipient's own codes by -0.0001 [-0.0028, 0.0026].

Two-part mixes. AlbumDB raw stems are grouped into vocals and accompaniment according to their official names; the grouping was fixed before inference. The two templates place the groups with the same carrier-preserving gains, vocals on Ls and accompaniment on Rs, and the reverse; five disjoint pairs of tracks are used in both directions. Codes of the same template transfer without loss (error difference -0.0037 [-0.0045, -0.0024]), codes of the other template do not (energy-distribution distance $+0.019$ [0.011, 0.028]), and in both templates the output favours the side of the accompaniment, the louder group.

Mirroring. The cosine between output and target changes in channel energy exceeds 0.95 for 21 of the 32 validation tracks and 0.99 for 16 (Figure 4(b)), and the magnitude of the change matches the target (0.155 against 0.154). NSFC-Flow agrees 4.32 dB better in SSR with the correct target than with the mismatched one, $+0.23$ [0.02, 0.45] above NSFC. On real music, the untested L, R and C have energy shares within 1.1% relative error.

AlbumDB masters. AlbumDB (McKenzie et al., 2026) provides, to our knowledge, the largest public set of human 7.1.4 music masters: ten tracks from a single album by one artist, mixed in a centre-heavy style far from common 7.1.4 music practice. Its balance departs far from the corpus: a median 71.7% of the non-LFE energy lies in the centre channel and the median relative LFE level is 6.1%, against 0.26% and 0.07% in the corpus. On these masters, mirroring does not steer the output (cosine 0.11 [-0.21, 0.41]; SSR advantage 0.018 dB), and oracle amplitude panning reconstructs them better than NSFC (normalised error 0.364 against 0.570). Ten tracks of one album and one mixing style are too few to characterise behaviour on human 7.1.4 mixes in general; we report them as lying outside the domain of the corpus.

Content retrieval. With the condition fixed, the time-averaged codes of the first window of each track query the second windows of all 32 tracks; ties share the hit. The null distribution permutes gallery identities 19,999 times. With the spatial condition held fixed, so that the 32 validation tracks share identical spatial information, the codes of one window retrieve the same track from another window in 54.7% of the cases (32-way retrieval; chance 3.1%, 95% upper bound of the permutation null 10.2%; mono spectrum 65.6%). The stream thus also carries content, as expected from an encoder that sees the reference spectrum; for coding, this points to redundancy that a lower rate could remove.

Statistics. Real-music results are averaged per track; transfer results per disjoint pair. Confidence intervals are 95% percentile intervals from 10,000 block-bootstrap resamples (seed 20260913), with blocks defined by track or pair and shared across systems. The intervals are per endpoint and not corrected for multiplicity.

Numerical note. All results decode from integer codes alone. Because the frozen quantiser rounds with a straight-through estimator, re-embedding the codes differs from the training-time quantised

tensor by at most 4.8×10^{-7}, which the decoder amplifies to a waveform difference of at most 3.0×10^{-5}; the code indices are identical.

Window length. The decoder normalises over the whole input window, so the energy of the rear surround and height channels depends slightly on the window length; from 12-s to 48-s windows it changes by -1.36 to -0.20 dB, with the total level unchanged. The objective tests use 4-s windows and the listening test 10-s excerpts, so their spatial values are not numerically comparable.