# PAEDB: A SYNTHETIC PRIMARY-AMBIENT DATASET GENERATION PIPELINE FOR AUTOMATIC UPMIXING USING DEEP NEURAL NETWORKS

*Nicholas N. Tong* *

Dept. of Music Engineering
University of Miami, Coral Gables, FL
Skywalker Sound, Nicasio, CA
nicktong@skysound.com

*Tom Collins*

Dept. of Music Engineering
University of Miami, Coral Gables, FL
tom.collins@miami.edu

## ABSTRACT

Automatic blind upmixing aims to convert audio from a smaller channel format (e.g. mono or stereo) into a multichannel format using estimates of direct and diffuse spatial statistics within the signal. Current approaches rely on primary-ambient extraction (PAE) algorithms, which lack real-world context through limited processing windows. Deep learning music source separation (MSS) models have been applied in voice-primary-ambient extraction (VPA) upmixing systems for handling direct components, but still rely on DSP methods of surround channel generation. This work further investigates utilizing source separation within VPA upmixing, focusing specifically on the task of stereo decorrelation and ambience extraction for 5.1 surround. We also release PAEDB (Primary–Ambient Extraction Dataset), a high-quality music dataset derived from MUSDB18-HQ and MoisesDB, comprising 1,809 primary–ambient stem pairs totaling over 550 hours of audio. The performance of selected DNNs trained on PAEDB is then evaluated using signal metrics and a listening study. Our findings indicate that DNNs can effectively model the behavior of PAE algorithms, establishing PAEDB as a strong foundation for ML upmixing systems and underscoring the need for higher-quality multichannel data to advance beyond conventional methods.

## 1. INTRODUCTION

The evolution from mono to stereo and now multichannel formats reflects a continuous pursuit of increasingly immersive and realistic listening experiences [1]. However, a significant gap remains in making high-quality surround content as accessible and widespread as stereo audio. In typical production workflows, multichannel mixes are already delivered as part of the final assets, and isolated stems are only required when engineers need to re-pan, revise, or creatively reinterpret the spatial layout. When source stems are unavailable (e.g., when working from a stereo master or legacy content), a blind upmix becomes necessary to derive the spatial components needed for multichannel reproduction.

Current algorithmic approaches include surround matrix decoding with direct-diffuse steering and repanning [2, 3, 4], among other DSP techniques that decompose the stereo signals into separate components based on interchannel correlation [5, 6, 7, 8]. These fall under the common strategy of primary-ambient extraction (PAE), which focuses on decomposing stereo audio into primary and ambient signal components. These processed components are then distributed across the surround channels of the final multichannel output. While effective, PAE algorithms rely on frame-wise assumptions about the signal without considering long-term contextual modeling. This often leads to audible artifacts in the output, especially when dealing with diffuse content.

Deep learning offers a promising alternative by reframing upmixing as a data-driven generative audio task. Deep neural networks (DNN) in music source separation (MSS) have achieved high accuracy in isolating musical stems while preserving temporal and spectral coherence [9, 10, 11, 12, 13, 14]. These capabilities suggest strong potential for learning decorrelated signals directly from stereo audio.

Direct application of DNNs for PAE in the context of multichannel upmixing remains limited, with existing approaches primarily employing basic feature extraction pipelines and shallow network designs [15, 16, 17]. Recent upmixing approaches use MSS models to separate instruments and dialogue, while still relying on traditional DSP to generate surround channels [18, 19].

This paper investigates how state-of-the-art MSS architectures can be adapted for ambient signal extraction to enable a more data-driven approach to the surround channel generation task in blind stereo-to-surround upmixing. The main contributions of this work include: (1) the creation of a custom synthetic dataset for the PAE task, (2) the adaptation and training of deep neural networks on the dataset for ambient signal extraction, and (3) the development and evaluation of an upmixing system demonstrating both the effectiveness of these models and the value of the proposed dataset.

## 2. BACKGROUND

### 2.1. Traditional Approaches to Upmixing

Primary-ambient extraction (PAE) operates on the assumption that a stereo signal is comprised of primary and ambient components. The primary components of a signal consist of correlated signals of prominent sound sources shared between channels. In contrast, the ambient component is comprised of decorrelated signals that are less common between channels. A stereo mixture $X_{LR}$ can be expressed as the sum of components $P_{LR}$ and $A_{LR}$.

$$\begin{aligned} X_L &= P_L + A_L, \\ X_R &= P_R + A_R, \end{aligned} \tag{1}$$

* This work was completed in partial fulfillment of a master's thesis at the University of Miami's Frost School of Music, and during an internship with the Applied Audio Research Team at Skywalker Sound. Thank you to Nicolas Tsingos for guidance and careful review of this work.

In upmixing, PAE can be used to perceptually expand the spatial field of lower-format audio when decorrelated copies of the ambient signals are routed to the surround and height speakers of a multichannel system (e.g., the Ls and Rs of a 5.1 configuration). This concept is leveraged in traditional state-of-the-art upmixers such as the hybrid scale-factor and variable matrix algorithm proposed in [3], where the steered primary signal is stabilized in the front sound stage while ambient material is panned toward the rear surrounds. An enhanced approach uses multiband matrix-decoding for improved separation of direct signals to higher channel formats of arbitrary size, and duplicates the decorrelated ambient signals across output channels for increased envelopment [4].

Variations to decompose direct and diffuse signals have been developed, including methods based on simple ambience panning, frequency-domain interchannel coherence [5], equal-level and equal-ratio of cross- and auto-correlation [6], principal component analysis (PCA) [7, 8], time-domain least-mean-square (LMS) filtering [20], and frequency-domain normalized least-mean-square (NLMS) adaptive filtering [19].

Center Channel Derivation (CCD), as proposed in [21], is a method of computing the front sound stage from a stereo signal. It extends the stereo signal model in Eq. 1 with the assumption that the primary-ambient signal components are encoded within the $LCR$ channels, where channel $C$ contains only $\vec{P}$, with $L$ and $R$ channels containing the mixture of $\vec{P}$ and $\vec{A}$ components.

$$X_L = \sqrt{0.5}\,\vec{C} + \vec{L} \quad \begin{cases} L = G_L\,\vec{P} + \vec{A_L} \\ C = G_C\,\vec{P} \\ R = G_R\,\vec{P} + \vec{A_R} \end{cases} \qquad (2)$$
$$X_R = \sqrt{0.5}\,\vec{C} + \vec{R}$$

where $G_s$ is the gain applied to channel $s \in \{L, C, R\}$.

The CCD algorithm leverages the psychoacoustic phenomenon of the "phantom center channel", where sounds located in the center of a stereo image are perceived to be originating in-between the physical stereo speakers. With an adapted three-channel primary-ambient signal model, the CCD algorithm applies vector-based signal decomposition in the frequency domain to geometrically estimate a dedicated center channel. This mitigates acoustic crosstalk inherent in the phantom center effect, where identical sounds from stereo speakers can create phase cancellations and comb-filtering [22]. It also expands the listening "sweet spot" in surround playback by reliably anchoring center content when the listener is not perfectly aligned with the left and right speakers [4].

### 2.2. Deep Learning Approaches to Upmixing

Primary-ambient decomposition has been identified as a promising direction for ML-based upmixing research [23]. Aligned with this directive, several deep learning approaches to upmixing have been proposed to estimate the underlying direct–diffuse signal characteristics targeted by traditional methods [24, 17, 15, 16].

One proposed method aims to extract ambient components from mono recordings by training a network to classify audio as "dry" or "wet" [24], where recordings with a negligible amount of ambience or reverberation are treated as "dry" signals. Another technique operates on the STFT-frame level, using a three-layer fully connected neural network to perform classification between reverberant and non-reverberant frequency frames [15]. Subband-based upmixing has also been explored, where a DNN is trained to generate the Quadrature Mirror Filter (QMF) subband signals which are then used to generate the center and surround channels using QMF synthesis [17]. Another similar approach optimizes two jointly trained networks for rendering primary and ambient signals using spectral and interchannel level differences to generate five-channel upmixes by applying the learned spectral weights of frequency-subbands to the input stereo signal [16].

Beyond PAE-based methods, latent-space approaches have been investigated using variational autoencoders (VAEs) trained on synthetically generated 5.1 mixes created with Vector-Based Amplitude Panning (VBAP) to spatially distribute stem sources. The VAE learns the latent representation of these spatial locations to upmix either blindly or via style-transfer from a 5.1 reference mix [25].

### 2.3. Source Separation and VPA Upmixing

Source separation is the process of decomposing a mixed audio signal into its individual components, often referred to as "stems". In most music separation tasks, these stem sources typically include primary musical components such as vocals, bass, drums, and other instrumental sources [14]. This further extends toward Cinematic Audio Source Separation (CASS), which targets dialogue, music and special effects [26, 27]. Source separation has broad applications from rearranging and remixing the original source material, speech enhancement [28], vocal and instrumental track removal [9], and audio upmixing [18, 19].

Historically, this is achieved using techniques like independent component analysis with blind source separation [28, 29], non-negative matrix factorization [30], and Wiener filtering [31]. The rise of deep learning has significantly advanced separation accuracy by enabling more complex temporal and spectral modeling for better separation performance, especially in challenging scenarios where overlapping frequencies may occur between stems. Architectural and processing advancements include autoencoder- and U-Net structures [32], hybrid time–frequency domain processing [12], frequency band-splitting and multi-band mask estimation [26, 33], and long-term context modeling with transformers and attention mechanisms [10, 12, 34].

Although not required for upmixing, disentangling audio mixtures with source separation can produce cleaner components for improved spatial-statistics estimation and allow for creative stem-level repanning. Building on this idea, Voice-primary-ambience extraction (VPA) upmixing utilizes MSS as a means of primary signal extraction by distributing the separated output to the front channels, while an ambience extraction algorithm is used for the surround channels. The VPA method is introduced in [18] using OpenUnmix [14] as a means of speech enhancement by isolating vocals to the center channel, with residual stems assigned to the left and right channels, and using the Equal-Levels Ambience Extraction (ELAE) algorithm [6] to generate the surround channels. Another approach [19] builds upon this by using four-stem Demucs [35] to separate and remix music stems across all the channels, then employs a NLMS adaptive filter with an Adaptive Panning (ADP) algorithm for ambience extraction.

## 3. METHODOLOGY

Our proposed upmixing approach further adapts the PAE-based VPA upmixing framework by combining MSS and the CCD algorithm in [21] for deriving the front sound stage. However, our approach diverges from prior methods by replacing the traditional ambience extraction block with a trained DNN optimized on a custom dataset of primary and ambient music audio (later referred

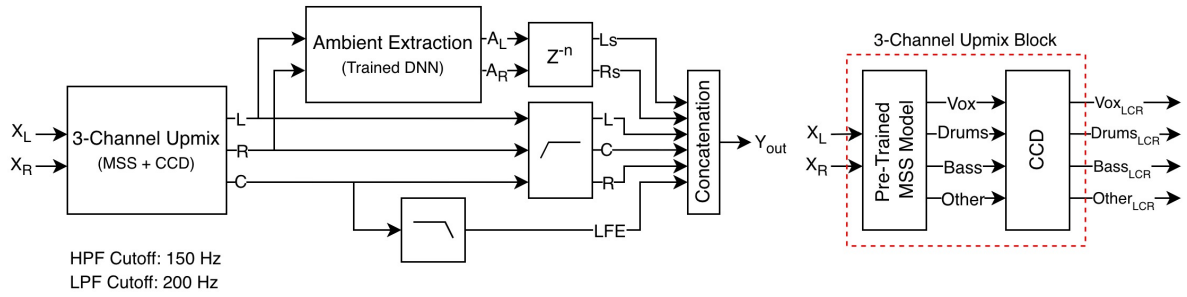


**Figure 1:** Block diagram of the proposed upmixing schema. The stereo input $X_{LR}$ passes through the 3-channel upmix, ambient extraction, and filtering blocks to produce $Y_{\text{out}}$. Details of the 3-channel upmix block are shown on the right.

to as PAEDB in Sec. 3.2). We evaluate the models' performance against prior approaches [18] using a user listening study (Sec. 4.2) and signal reconstruction metrics (Sec. 4.1).

This work focuses exclusively on improving VPA upmixing by evaluating source separation for ambience extraction in a 5.1 surround configuration. It does not yet address height-channel derivation for immersive formats, which remains well supported by traditional upmixers.

### 3.1. Upmixing System

Fig. 1 details the processing pipeline of the proposed upmixing system, from the stereo input $X_{LR}$ to the final concatenation of the multichannel output $Y_{out}$.

1. **3-Channel Upmix Block** – This block converts $X_{LR}$ into three channels by first applying a pretrained MSS model to extract drums, bass, vocals, and other stems. Each stem is then processed with the CCD algorithm [21] to generate its $LCR$ components, which are routed to the corresponding output channels. This anchors center-panned content in $C$ (e.g., lead vocals, kick/snare, bass) while positioning panned material in $L$ and $R$ (e.g., background vocals, drum overheads).
2. **Ambient Extraction Block** – This block applies the custom-trained DNN to the per-stem processed $L$ and $R$ outputs of the Three-Channel Upmix block. An additional 12 ms delay is applied to the output to enhance perceived separation and to mitigate potential phase issues between the front and rear channels.
3. **Additional Filtering Blocks** – This block applies 5th-order Butterworth low- and high-pass filters to generate the $LFE$ channel and to remove low-frequency content overlap from the surround channels. This is achieved using a low-pass filter with a cutoff of 200 Hz for the $LFE$ channel, and a high-pass filter with a cutoff of 150 Hz for $LCR$ channels.

### 3.2. Dataset Procurement and Data Preparation

Although the approaches in Sec. 2.2 show that direct–diffuse statistics can be learned, the training data used across studies are inconsistent, with each adopting its own definition of the signal classes. Many curated datasets rely on subjective ad hoc labeling or coarse binary labels, even though both classes may still contain remnants of direct and diffuse energy. This fails to represent the statistical complexity of the PAE task, resulting in unclear class boundaries and limited comparability between works. This highlights a need for a high-quality, standardized benchmarking dataset to advance PAE-based upmixing research, analogous to that of MUSDB18 [36] and EARS [37]. It is worth noting that PAE ambience is broader than reverberation in the acoustic sense: diffuse energy arises not only from room acoustics but also from the inherent resonant characteristics of the sound source (e.g., the natural decay of an acoustic guitar body). Consequently, a convolution-based dataset construction approach using room impulse responses would be insufficient, as no large corpus of truly "dry" music recordings exists, and even anechoically recorded material would retain instrument-level ambient energy that falls within the scope of PAE.

For the reasons stated, we introduce the Primary-Ambient Extraction Dataset (PAEDB) which utilizes MUSDB18-HQ [38] and MoisesDB [39], two publicly available music datasets for a combined total of 390 individual songs sampled at 44.1 kHz. Note that both datasets are available for non-commercial research use only. Although this work focuses on music, the dataset generation pipeline is general and can be applied to other audio domains.

Fig. 2 illustrates the preprocessing pipeline. Each PAEDB entry contains the original mixture along with its corresponding primary and ambient sources. From Eq. 2, we note that most ambient energy resides in the $L$ and $R$ channels of the 3-channel signal model, whereas the extracted $C$ channel contains only primary energy. Leveraging this property, we first apply the CCD algorithm in [21] and discard the $C$ channel to suppress primary components. Ambient stems are then estimated from the resulting $L$ and $R$ channels using the ELAE method [6]. Informal listening tests confirm that incorporating CCD as a preprocessing step to PAE extraction reduces primary bleed into the ambient output.

After removing tracks containing silent or near-silent stems, a total of 385 valid song tracks remain, accumulating to 1,925 individual stereo WAV files and approximately 120 hours of audio. To expand the total dataset size, each individual file from the base 4-stem dataset (vocals, drums, bass, other, and mixture) is treated as its own unique entry. After generating the corresponding primary and ambient components for each original audio file, the original target stem is repurposed as the new mixture file. This process effectively increases the trainable dataset size by approximately a factor of 5, corresponding to the five file types (vocals, drums, bass, other, and mixture) derived from each song. After additional validation, the final PAEDB consists of 1,809 unique training elements (`primary.wav`, `ambient.wav`, and `mixture.wav`), resulting in 5,427 individual files and approximately 550 hours of audio, totaling around 150 GB of 44.1 kHz stereo WAV files. Dataset generation scripts are made publicly available for reproducibility [1].

[1] `https://github.com/nick7ong/paedb`

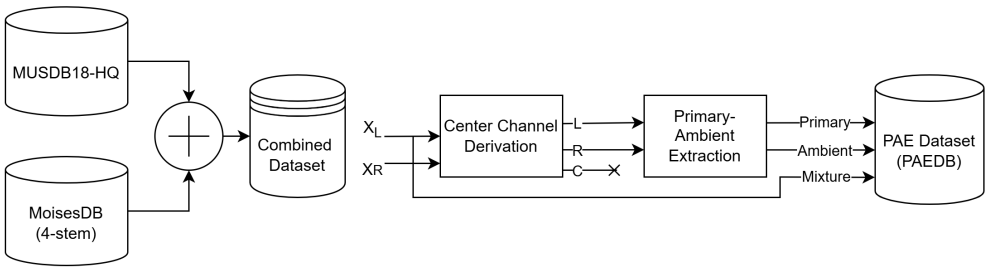


**Figure 2:** Diagram illustrating the preparation and generation process of the Primary-Ambient Extraction Dataset (PAEDB).

### 3.3. Model Selection

To evaluate the DNN architectures for the PAE task, we train two distinct models on the PAEDB dataset. The first selected model is the Mel-RoFormer [10] which employs a Mel-spectrogram transform in a band-split RNN and utilizes hierarchical Transformers with Rotary Position Embeddings to model both inner- and inter-band sequences for multi-band mask estimation. As of this paper, the Mel-RoFormer achieves state-of-the-art SDR performance on the MUSDB18-HQ benchmark [10, 38].

The second model is a smaller multi-layer perceptron (MLP) architecture inspired by the design proposed in [15], consisting of three fully connected linear layers, and operates by flattening the real and imaginary components of the input STFT representation into a one-dimensional feature vector. The network is also comprised of ReLU activations, reducing the feature dimensionality through hidden layers of sizes 15 and 10 before outputting a single probability through a Sigmoid activation function. This output is used as an estimated STFT mask that is applied to the original input audio to yield the separated primary and ambient components. Since the original architecture was explicitly designed for PAE and relatively simple to adapt, we selected it as an additional supporting method of prior approaches.

### 3.4. Training Setup

Our training procedure follows a supervised learning strategy over the prepared PAEDB audio dataset. All training runs were conducted using 2 NVIDIA A10 GPUs, each with 23 GB of vRAM. Both models were trained from scratch using randomly initialized weights. The model and training parameters used were the same as respectively reported in [10, 15] (see Sec. 6.1 for discussion).

**Mel-RoFormer** – The optimizer used is AdamW, with a learning rate of 5e-5, a reduction factor of 0.95, a patience of 2 epochs, and an EMA momentum of 0.999 for stabilizing parameter updates. We train this model for 4,000 steps per iteration over 100 epochs total with a batch size of 3 per GPU (an effective batch size of 6 across both GPUs). Optimization combines L1 loss for waveform reconstruction in the time domain with a weighted multi-resolution STFT (MRSTFT) loss for spectrogram reconstruction in the frequency domain.

**3-Layer MLP** – This model is trained using the AdamW optimizer with a learning rate of 1e-3, reduction factor of 0.95, and a patience of 5 epochs. Because of its small size, the MLP was trained using a single GPU with a batch size of 16, and 4,000 steps per iteration over 200 epochs. Optimization uses L1 loss applied to the final Sigmoid function output, with binary labels assigned 0 and 1 for primary and ambient STFT frames respectively.

## 4. EVALUATION

To evaluate the DNN-based upmixing system and the quality of the PAEDB dataset, we conduct an objective signal-based assessment on a collected test set of full-length music tracks, followed by a subjective listening study using a curated subset of stimuli.

### 4.1. Objective Metric Evaluation

Objective evaluation is performed on an audio test set consisting of 12 full-length stereo music tracks sourced from YouTube, each ranging from 2 to 6 min in duration and sampled at 44.1 kHz.

The audio excerpts are acquired using a custom Python tool [2] that extracts the highest-quality audio stream from a given URL (128 kb/s, `.m4a` format), which is then converted to 44.1 kHz WAV format using `ffmpeg`. All audio content is used strictly for academic research purposes and is not intended for redistribution or commercial use.

Signal metrics are used to assess the separation and reconstruction quality of the ambience produced by the trained Mel-RoFormer and MLP, using the ELAE-generated signals as targets. These reference signals serve to evaluate how accurately each network reproduces the target ambient signal generated for the surround channels (see Sec. 6.2 for discussion). Following common MSS benchmarks, we compute Signal-to-Distortion Ratio (**SDR**), Scale-Invariant SDR (**SI-SDR**), and Log-Spectral Distance (**LSD**).

### 4.2. Subjective User Listening Study

To assess the perceptual quality of the trained models in the context of stereo-to-5.1 upmixing, a user listening study was conducted with 12 participants in an acoustically treated environment equipped with a calibrated 5.1 speaker array to ensure consistent audio playback. Four upmixing systems were evaluated:

- **All-Channel Stereo (Baseline):** The original stereo mix duplicated across the left ($L$), right ($R$), back left ($Ls$), and back right ($Rs$) speakers. This serves as a baseline system, similarly used in [19, 25].
- **VPA-ELAE (Reference):** VPA upmixing using ELAE to derive the surround channels, as described in [18, 19].
- **VPA-MLP:** VPA upmixing using the optimized MLP model trained on the PAEDB dataset to extract ambience for the surround channels.
- **VPA-Mel-RoFormer:** VPA upmixing using the optimized Mel-RoFormer [10], trained on the PAEDB dataset to extract ambience for the surround channels.

[2] `https://github.com/nick7ong/YTAudioScraper`

A subset of six 1 min excerpts are selected from the full test set. The selection aims to reflect a diverse range of genres, energy levels, and ambient textures, with attention to varying production styles. Each participant listened to 24 stimuli in total, consisting of the six audio excerpts processed by each of the four upmixing systems. To mitigate bias and order effects, we randomized both the stimulus sequence and the system presentation order. Stimuli included *Sound & Color* [40], *Let You Break My Heart Again* [41], *Snow Song* [42], *Take Five* [43], *Money* [44], and *Prison Break Scene* [45], spanning rock, orchestral, folk, jazz, and cinematic genres.

Excerpts were evaluated on a 7-point Likert scale across three perceptual metrics: **Overall Envelopment and Immersion** (listener engagement and overall preference), **Spatial & Temporal Quality** (localization accuracy, perceived depth, and temporal coherence), and **Spectral Quality** (timbral fidelity, spectral balance, and artifact presence).

The study and its adherence to ethical guidelines were approved by the Institutional Review Board (IRB) at the University of Miami.

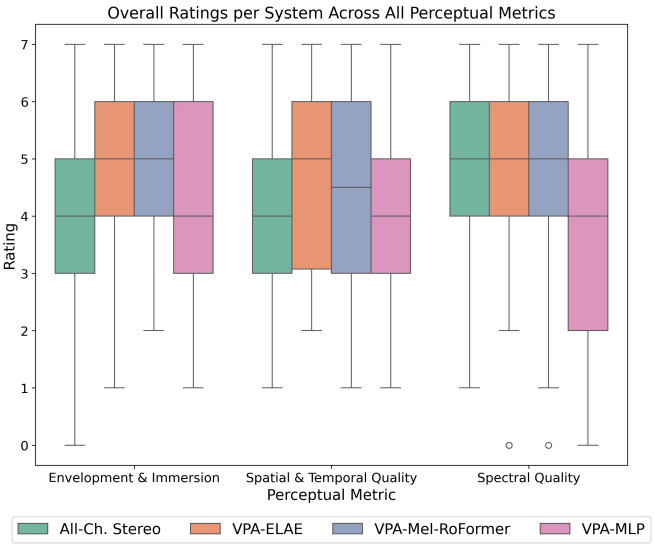


**Figure 3:** Boxplot of user perceptual ratings across systems.

## 5. RESULTS

### 5.1. Objective Metric Results

As outlined in Sec. 4.1, we employed three objective signal metrics (SDR, SI-SDR, and LSD) to quantitatively evaluate the reconstruction performance of the trained neural networks and their ability to generate ambient signals. These metrics were computed across the test set outlined in Sec. 4.2 and using ELAE-generated ambient signals as ground truth references (see Sec. 6.2).

Table 3 presents the averaged results (mean $\pm$ standard deviation) for the Mel-RoFormer and MLP models evaluated separately from the VPA systems. The Mel-RoFormer is shown to consistently outperform the MLP across all signal metrics. Notably, the Mel-RoFormer model's higher SI-SDR and lower LSD scores indicates better signal reconstruction and spectral consistency respectively in the extracted ambient signals. In contrast, the MLP exhibited greater spectral deviation and lower SDR ratios, aligning with the observations of audible artifacts resulting in lower perceived Spectral Quality as reported in Fig. 3.

### 5.2. Subjective Listening Study Results

The collected user ratings across all combinations of systems, stimuli, and perceptual metrics, resulted in 264 total observations. A non-parametric Friedman test was used to assess the main effects of the four systems across the perceptual metrics for each stimuli.

Table 1 presents the overall ratings computed as the aggregated scores of the perceptual metrics averaged over each stimulus, with systems ranked by combined mean listener ratings. Participants preferred the VPA-ELAE system the most with a mean score of 14.63, followed closely by the VPA–Mel-RoFormer system at 14.45. All-Channel Stereo and VPA-MLP rated comparatively lower, with VPA-MLP obtaining the lowest preference score.

Table 2 summarizes the statistical analysis conducted separately for each metric between systems, which revealed a statistically significant effect of System for Overall Envelopment & Immersion ($\chi^2 = 10.816$, $p = 0.0128$), Spatial & Temporal Quality ($\chi^2 = 9.816$, $p = 0.0202$), and Spectral Quality ($\chi^2 = 10.500$, $p = 0.0148$). All reported p-values are evaluated at a significance threshold of $p < 0.05$. The chi-square results show that listener ratings differed significantly across systems, indicating that listeners reliably perceived the systems as distinct, and that system presentation had a meaningful impact on preference as well as the perceived spatial and spectral quality.

Fig. 3 illustrates the distribution of listener ratings across the subjective metrics for each processing system. For both the Overall Envelopment & Immersion and Spatial & Temporal Quality metrics, VPA-Mel-RoFormer and VPA-ELAE exhibited higher medians and upper quartiles compared to All-Channel Stereo, despite having lower mean ratings. This suggests that a subset of listeners consistently rated these systems highly, indicating a preference for those specific conditions.

In contrast, Spectral Quality medians and interquartile ranges for all systems, with exception of VPA-MLP, were nearly identical which indicates comparable performance and greater agreement among listeners on the perceived audio quality and coherence. Participant feedback attributed this to audible artifacts introduced by the binary gating mechanism of the MLP output over the processed audio frames. Interestingly, some listeners reported that these same artifacts occasionally enhanced immersion when coinciding with sound-effect onsets in the cinematic sample or with drum hits in music samples, suggesting that their impact on perception can be context-dependent.

### 5.3. Computational Benchmark

Tables 4 and 5 summarize the benchmarked runtime and memory footprint of the ambience extraction methods in Sec. 4.2 averaged across 10 repeated trials, with algorithms listed in order of fastest runtime. Each method was tested using a randomly generated 1 s stereo noise waveform sampled at 44.1 kHz. CPU and GPU memory usage was captured using `tracemalloc.get_traced_memory()` and `torch.cuda.max_memory_allocated()`, respectively. All experiments are conducted on an Intel Xeon Gold 6338 CPU with 100 GB of RAM and an NVIDIA A100 GPU with 20 GB of vRAM.

For the CCD-ELAE variants, we use an internal STFT window size of 1024 samples with an overlap of 512 samples. The

**Table 1:** Overall combined mean ratings of perceptual metrics across all systems.

| System | Total Mean Rating | Participant Ranking |
|---|---|---|
| **VPA-ELAE** | **14.63** | **1** |
| VPA-Mel-RoFormer | 14.45 | 2 |
| All-Channel Stereo | 12.91 | 3 |
| VPA-MLP | 12.32 | 4 |

**Table 2:** Friedman test across all perceptual metrics.

| Perceptual Metric | Chi-square ($\chi^2$) | p-value |
|---|---|---|
| Overall Envelopment & Immersion | 10.816 | **0.0128** |
| Spatial & Temporal Quality | 9.816 | **0.0202** |
| Spectral Quality | 10.500 | **0.0148** |

**Table 3:** Average objective signal metrics across deep learning-based systems.

| System | SDR (dB) ↑ | SI-SDR (dB) ↑ | LSD (dB) ↓ |
|---|---|---|---|
| **Mel-RoFormer** | $\mathbf{2.22 \pm 2.21}$ | $\mathbf{3.44 \pm 1.61}$ | $\mathbf{106.93 \pm 10.80}$ |
| MLP | $0.89 \pm 0.80$ | $-5.29 \pm 4.74$ | $446.20 \pm 520.56$ |

**Table 4:** CPU Benchmark Summary

| Algorithm | Runtime (s) | Memory (MB) |
|---|---|---|
| MLP | $0.347 \pm 0.174$ | $0.006 \pm 0.000$ |
| CCD-ELAE (PyTorch) | $0.582 \pm 0.317$ | $0.003 \pm 0.003$ |
| CCD-ELAE (NumPy) | $1.087 \pm 0.004$ | $13.42 \pm 0.010$ |
| Mel-RoFormer | $5.161 \pm 1.553$ | $0.318 \pm 0.003$ |

**Table 5:** GPU Benchmark Summary

| Algorithm | Runtime (s) | Memory (MB) |
|---|---|---|
| MLP | $0.004 \pm 0.0003$ | $38.444 \pm 0.000$ |
| CCD-ELAE (PyTorch) | $0.058 \pm 0.122$ | $6.5324 \pm 0.223$ |
| Mel-RoFormer | $0.064 \pm 0.001$ | $213.95 \pm 0.000$ |
| CCD-ELAE (NumPy) | – | – |

device-specific results for the PyTorch CCD-ELAE implementation correspond to the device onto which the tensors are mapped to using `waveform.to(device)` (i.e., `"cpu"` or `"cuda"`).

In terms of raw speed, the MLP achieves the lowest runtime on both CPU and GPU, with the PyTorch CCD-ELAE as a close second. While the Mel-RoFormer is the slowest on CPU due to its parameter count, all PyTorch-based methods achieve faster-than-real-time inference on GPU. The NumPy CCD-ELAE, being CPU-bound, is excluded from GPU benchmarking. The lower CPU memory figures reflect a limitation of `tracemalloc`, which tracks only Python-level heap allocations, unlike `torch.cuda.max_memory_allocated()` which reports full CUDA memory usage.

For comparability, benchmarking used a batch size of one to match the DSP baselines. In deployment, batched GPU inference and compilation tools such as ONNX or TensorRT could substantially reduce runtime and memory overhead beyond the figures reported here.

## 6. DISCUSSION

### 6.1. Optimization Metrics and Loss Functions

Although promising results were observed with the Mel-RoFormer's performance, signal metrics still remain on the lower-end compared to MSS benchmarks on traditional stem tasks. This could suggest a potential limitation in relying on conventional signal reconstruction metrics for evaluating ambient signal quality.

Ambient components are by nature diffuse and decorrelated, making the separation task arguably more difficult than conventional sources. Even when the reconstructed output perceptually resembles ambience and exhibits the expected statistical decorrelation from the input mixture, small deviations from the reference signal can still yield disproportionately low signal scores.

Similarly, using L1 and MRSTFT loss may not be fully ideal either. While they may help guide the model toward reconstructing the target waveform and spectral structure, there is still potential that the optimization function used in training fails to align with perceptual relevance. This may cause the network to converge on solutions that technically satisfy the loss function, but fail to produce perceptually accurate results. This may also contribute to prior works reporting lower SDR values when training with these loss functions [16, 17].

### 6.2. Data Scarcity and Synthetic Training Targets

While datasets for speech dereverberation exist [37], there still lacks a comparable corpus for ambience extraction in broader music and film contexts. To address this, we rely on the CCD and ELAE algorithms in [21, 6] to synthetically generate training targets. Although prior works have similarly synthesized training data [25], this strategy imposes a potential performance ceiling during training since a model trained solely to imitate a DSP algorithm cannot initially exceed it.

However, one could argue that pretraining a DNN to mimic a DSP algorithm provides a useful foundation for later finetuning on higher-quality datasets when they become available. Such datasets

could include professionally produced multichannel mixes where true ambient stems are available from the original recording session, human-curated primary-ambient pairs validated by experienced mix engineers, or consensus targets derived from multiple PAE algorithms to reduce single-method bias. Although the networks are initially constrained by the quality of the DSP-generated targets, this could be interpreted as a form of guiding the model's internal parameters toward the desired feature distribution rather than a fixed performance ceiling. By first learning from algorithmic outputs, the model develops a strong inductive prior that can then be further finetuned on higher-quality or task-specific data without requiring any redesign of the DSP method's underlying assumptions. Future work could further enrich the training feature distribution by incorporating targets from multiple PAE algorithms or by convolving dry recordings with room impulse responses in addition to the generation pipeline proposed in this paper.

## 7. CONCLUSION

This study presents a ML-based upmixing approach that reframes ambient signal extraction as a source separation problem by further integrating DNNs into the upmixing pipeline. Our approach extends the VPA upmixing framework by leveraging a pretrained MSS model and CCD algorithm to improve distribution of primary components while using the custom-trained models optimized for decorrelated ambience extraction. To address inconsistencies in prior training data, we introduced PAEDB, a large-scale audio dataset built from publicly available music data using a reproducible CCD-ELAE processing pipeline for deriving clean primary and ambient stem sources.

The evaluation results demonstrate that the Mel-RoFormer pretrained on PAEDB can effectively learn the core algorithmic behavior of the CCD-ELAE algorithm while remaining fully differentiable and adaptable. In this regard, the PAEDB dataset provides a valuable foundation for pretraining neural networks as potential submodules within a fully differentiable upmixing system. This opens the door to subsequent finetuning on real multichannel material and positions neural-based upmixing approaches to ultimately surpass the capabilities of their traditional DSP counterparts.

Future work should prioritize collecting real professionally produced 5.1 material to eliminate reliance on synthetic targets, and explore perceptually motivated loss functions derived from interchannel decorrelation or mid/side processing. Generative approaches such as diffusion-based and variational latent-space models also present promising directions for more realistic stereo-to-surround upmixing.

## 8. REFERENCES


[1] S. Agrawal, S. Bech, K. De Moor, and S. Forchhammer, "Influence of changes in audio spatialization on immersion in audiovisual experiences," *Journal of the Audio Engineering Society*, vol. 70, no. 10, pp. 810–823, 2022.

[2] R. Dressler, "Dolby surround pro logic decoder principles of operation," *Dolby White paper*, 2000.

[3] M. F. Davis, C. Q. Robinson, and M. S. Vinton, "Signal models and upmixing techniques for generating multichannel audio," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. Audio Engineering Society, 2010.

[4] M. Vinton, D. McGrath, C. Robinson, and P. Brown, "Next generation surround decoding and upmixing for consumer and professional applications," in *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology–Cinema, Television and the Internet*. Audio Engineering Society, 2015.

[5] C. Avendano and J.-M. Jot, "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2002, pp. II–1957.

[6] J. Merimaa, M. M. Goodwin, and J.-M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Audio Engineering Society Convention 123*. Audio Engineering Society, 2007.

[7] Y.-H. Baek, S.-W. Jeon, S.-p. Lee, and Y.-C. Park, "Efficient primary-ambient decomposition algorithm for audio upmix," *Journal of Broadcast Engineering*, vol. 17, no. 6, pp. 924–932, 2012.

[8] K. M. Ibrahim and M. Allam, "Primary-ambient extraction in audio signals using adaptive weighting and principal component analysis," in *Proc. Sound and Music Computing Conference (SMC)*, 2016, pp. 227–232.

[9] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020.

[10] W.-T. Lu, J.-C. Wang, Q. Kong, and Y.-N. Hung, "Music source separation with band-split rope transformer," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 481–485.

[11] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[14] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix-a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[15] K. M. Ibrahim and M. Allam, "Primary-ambient source separation for upmixing to surround sound systems," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 431–435.

[16] J. Choi and J.-H. Chang, "Exploiting deep neural networks for two-to-five channel surround decoder," *Journal of the Audio Engineering Society*, vol. 68, no. 12, pp. 938–949, 2021.

[17] S. Y. Park, C. J. Chun, and H. K. Kim, "Subband-based upmixing of stereo to 5.1-channel audio signals using deep neural networks," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2016, pp. 377–380.

[18] R. T. Paez Amaro, C. Tejeda Ocampo, E. Souza Blanes, S. Bharitkar, and L. Madrid Herrera, "Deep learning based voice extraction and primary-ambience decomposition for stereo to surround upmixing," in *Audio Engineering Society Convention 154*. Audio Engineering Society, 2023.

[19] M. Negru, B. Moroşanu, A. Neacşu, D. Drăghicescu, and C. Negrescu, "Automatic audio upmixing based on source separation and ambient extraction algorithms," in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2023, pp. 12–17.

[20] R. Irwan and R. M. Aarts, "Two-to-five channel sound processing," *Journal of the Audio Engineering Society*, vol. 50, no. 11, pp. 914–926, 2002.

[21] E. Vickers, "Frequency-domain two-to three-channel upmix for center channel derivation and speech enhancement," in *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[22] ——, "Fixing the phantom center: diffusing acoustical crosstalk," in *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[23] J.-M. Jot, G. Wichern, T. Paez, and S. Bharitkar, "Towards leveraging machine learning for multichannel audio upmix," in *Proceedings of the AES International Conference on Artificial Intelligence and Machine Learning for Audio*, London, UK, Sep. 2025, late Breaking Demo Paper.

[24] C. Uhle and C. Paul, "A supervised learning approach to ambience extraction from mono recordings for blind upmixing," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*. DAFx Helsinki, 2008.

[25] H. Yang, S. Wager, S. Russell, M. Luo, M. Kim, and W. 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 (ICASSP)*. IEEE, 2022, pp. 426–430.

[26] K. N. Watcharasupat, C.-W. Wu, Y. Ding, I. Orife, A. J. Hipple, P. A. Williams, S. Kramer, A. Lerch, and W. Wolcott, "A generalized bandsplit neural network for cinematic audio source separation," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 73–81, 2024.

[27] D. Petermann, G. Wichern, Z.-Q. Wang, and J. Le Roux, "The cocktail fork problem: Three-stem audio separation for real-world soundtracks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022.

[28] E. Visser and T.-W. Lee, "Speech enhancement using blind source separation and two-channel energy based speaker detection," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[29] A. R. López, N. Ono, U. Remes, K. Palomäki, and M. Kurimo, "Designing multichannel source separation based on single-channel source separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 469–473.

[30] Y. Zhang and Y. Fang, "A nmf algorithm for blind separation of uncorrelated signals," in *2007 International Conference on Wavelet Analysis and Pattern Recognition*, vol. 3. IEEE, 2007, pp. 999–1003.

[31] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE, 2013, pp. 1–4.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[33] Y. Luo and J. Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1893–1901, 2023.

[34] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[35] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[36] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[37] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *ISCA Interspeech*, 2024.

[38] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-hq - an uncompressed version of musdb18," Aug. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373

[39] I. Pereira, F. Araújo, F. Korzeniowski, and R. Vogl, "Moisesdb: A dataset for source separation beyond 4-stems," 2023.

[40] Alabama Shakes, "Sound & Color," https://www.youtube.com/watch?v=bBh71Sl1i2Y, 2015, accessed: 2025.

[41] Laufey & Philharmonia Orchestra, "Let you break my heart again," https://www.youtube.com/watch?v=NLphEFOyoqM, 2021, accessed: 2025.

[42] Adrianne Lenker, "Snow song," https://www.youtube.com/watch?v=fs1sRHUQOgc, 2018, accessed: 2025.

[43] The Dave Brubeck Quartet, "Take five," https://www.youtube.com/watch?v=ryA6eHZNnXY, 1959, accessed: 2025.

[44] Pink Floyd, "Money," https://www.youtube.com/watch?v=2aW7HweAf3o, 1973, accessed: 2025.

[45] Guardians of the Galaxy, "Prison break scene," https://www.youtube.com/watch?v=gOLy6JArNYo, 2014, accessed: 2025.