NBU: NEURAL BINAURAL UPMIXING OF STEREO CONTENT

Philipp Grundhuber and Michael Lovedee-Turner

Fraunhofer Institute for Integrated Circuits (IIS)
Erlangen, Germany
philipp.grundhuber@iis.fraunhofer.de

Emanuël A. P. Habets

International Audio Laboratories Erlangen
Erlangen, Germany

ABSTRACT

While immersive music productions have become popular in recent years, music content produced during the last decades has been predominantly mixed for stereo. This paper presents a data-driven approach to automatic binaural upmixing of stereo music. The network architecture HDemucs, previously utilized for both source separation and binauralization, is leveraged for an end-to-end approach to binaural upmixing. We employ two distinct datasets, demonstrating that while custom-designed training data enhances the accuracy of spatial positioning, the use of professionally mixed music yields superior spatialization. The trained networks show a capacity to process multiple simultaneous sources individually and add valid binaural cues, effectively positioning sources with an average azimuthal error of less than $11.3\,^{\circ}$. A listening test with binaural experts shows it outperforms digital signal processing-based approaches to binauralization of stereo content in terms of spaciousness while preserving audio quality.

1. INTRODUCTION

In the past decade, the realm of immersive audio has gained significant attention from both academia and industry [1]. While the majority of consumed music is still stereo [2], an increasing number of streaming providers offer immersive content [3, 4]. The shift towards immersive content is also visible in the music industry as shown by recent advances in upmixing pop-cultural classics, which were originally produced in stereo [5]. This work explores a data-driven approach for automatic upmixing of music from stereo to immersive binaural. The scope of this work focuses on the reproduction of immersive content on headphones, one of the most accessible mediums for music consumption [6], which makes them an interesting platform for immersive playback and upmixing.

Traditionally, upmixing methods aim to increase the number of channels present in a given audio excerpt and can typically be split into two main categories, namely, direct ambient extraction and source separation [7]. Direct ambient extraction methods aim to decompose a given signal into direct and ambient components using techniques such as Wiener filtering [8], principal component analysis [9], or using Deep Neural Networks (DNNs) to estimate time-frequency masks [10, 11]. The extracted ambient component is then typically positioned in the rear or surround speakers [12]. In contrast, the second category of methods aims to generate additional output signals by identifying direct sources and re-panning them in a higher-order loudspeaker configuration [7,13–15]. Many upmixing algorithms combine both approaches, while decomposition and analysis is usually done in the time-frequency domain [7, 14], allowing for the separation and re-panning of multiple direct sources [15].

More recently, a new category has emerged focusing on end-to-end upmixing approaches. For example, Yang et al. [16] proposed an upmixing approach using variational autoencoders and neural style transfer. This approach is based on the disentanglement of the spatial attributes of a stereo mix from its musical content, thereby enabling the conversion of audio from two to five channels by adjusting the spatial image [16].

Focusing on source separation techniques, DNN-based approaches have become prominent in the field of Blind Source Separation (BSS), particularly for Music Source Separation (MSS). Typically MSS aims to isolate submixes from the original stereo track into four distinct categories: 'Drums', 'Vocals', 'Bass', and 'Other' [17]. The primary network architectures employed are based on Convolutional Neural Network (CNN) [18–21] or Recurrent Neural Network (RNN) [22]. CNN-based architectures typically incorporate two separate branches for time-domain and frequency-domain representations of the signal [19–21]. Notably, the bi-U-Net Hybrid Transformer Demucs [19], which combines Wave-U-Net [18] and HDemucs [20], delivers the best performance in Signal-to-Distortion Ratio (SDR) on the MUSDB18 dataset [23], surpassed only for extraction of vocal components by the RNN-based Band Split RNN (BSRNN) [22]. BSRNN uses complex-valued spectrograms and multiple dual-path RNNs, each acting on individual frequency bands [22]. While these models are computationally expensive, simpler models like the KUIELab-MDX-Net [21] exhibit marginally lower performance. The use of larger datasets has been shown to improve the performance of MSS, where the best-performing networks are typically trained with more than 800 songs [19].

Reproducing immersive mixes on headphones involves binaural rendering typically done by convolution of the composite monophonic signal of a given virtual scene with Head-Related Transfer Functions (HRTFs) representing a given position [24]. Recent work has explored binaural rendering using DNNs. Richard et al. [25] introduced WarpNet, a temporal convolutional network derived from WaveNet [26] for neural binaural rendering of speech with a sample rate of 16 kHz, synthesizing two-channel audio from mono input and listener position. Leng et al. [27] proposed BinauralGrad, a two-stage diffusion-based generative network that conditions on positional information to convert mono audio to binaural, outperforming WarpNet. LLuis et al. [28] developed Points2Sound, a model leveraging the HDemucs architecture [20], combined with three-dimensional point clouds to generate binaural audio from mono sources, demonstrating effective spatial audio synthesis for immersive applications.

While BSS and binauralization are well-researched areas, immersive mixing of music is subject to artistic intent. Nevertheless, studies such as [29] have found consistent instrument positions across award-winning immersive mixes, with rhythmic and bass elements centered, lead vocals in the front, and harmonic instruments placed wider. These consistent positions across samples motivate us to explore a combined approach to binaural upmixing in favor of the two separate processes of source separation and subsequent blind binauralization.

This paper presents a data-driven binaural upmixing method, which leverages recent advances in BSS and binauralization to develop an end-to-end approach using the HDemucs architecture – referred to as Neural Binaural Upmixer (NBU). Audio examples processed by NBU are made available and can be accessed on the Audiolabs website[1].

The structure of the paper is as follows: Section 2 presents the problem formulation, Section 3 covers the network architecture and training methods, Section 4 details the data used in this study, Section 5 presents and discusses the results, and Section 6 concludes the paper.

2. PROBLEM FORMULATION

In traditional stereo, individual recordings, also known as stems, are placed horizontally between two speakers by amplitude panning [1]. In contrast, binaural audio leverages spatial cues characteristic of the human auditory system to position these stems at a given point in space within a three-dimensional sound field [1]. Therefore, the aim of a NBU is to position the composite sound sources of a given stereo mix within a three-dimensional sound field based on positional data in the stereo mix while maintaining the timbral characteristics and fidelity of the original mix.
In this work, a stereo signal $s_{\mathrm{L,R}} \in \mathbb{R}^{2\times N}$ is defined as a mixture of M amplitude-weighted mono signals $s \in \mathbb{R}^{1\times N}$ of length N samples as,

$$s_{\mathrm{L,R}} = \sum_{m=1}^{M} s_m \mathrm{A}(\theta_m), \tag{1}$$

where $\mathrm{A}(\theta_m)$ represents an amplitude panning matrix that weights the amount of signal s_m that is distributed to left and right audio channels depending on a given panning angle θ_m. Expanding (1) to the binaural use case gives us,

$$\tilde{s}_c = \sum_{m=1}^{M} s_m \circledast \mathrm{H}_c(\theta_m, \varphi_m, \mathrm{d}_m) \text{ for } c \in \{\mathrm{L}, \mathrm{R}\}, \tag{2}$$

where $\circledast$ denotes the convolution operation and $\mathrm{H}_c(\theta_m, \varphi_m, \mathrm{d}_m)$ is the Head-Related Impulse Response (HRIR) corresponding to the position of source m for the left (L) and right (R) ears with elevation angle φ_m and the distance d_m.

The proposed NBU represents some model F, in this study a DNN, that generates an output $\hat{s}_{\mathrm{L,R}} \in \mathbb{R}^{2\times N}$ from a short segment of a stereo mix $s_{\mathrm{L,R}}$, using only the implicit spatial information in the original stereo mix, that approaches a binauralized mix, i.e., $\hat{s}_{\mathrm{L,R}} = F(s_{\mathrm{L,R}})$ approaches $\tilde{s}_{\mathrm{L,R}} = [\tilde{s}_{\mathrm{L}}, \tilde{s}_{\mathrm{R}}]$.

[1] https://www.audiolabs-erlangen.de/resources/2024-DAFx-Neural-Binaural-Upmix

3. PROPOSED METHOD

3.1. Model Architecture

Based on its prior use and performance in both source separation [20] and binauralization [28], HDemucs was chosen as the base architecture. The original and modified HDemucs are trained and compared. Changes to architecture are motivated by the outcome of an ablation study by Pons et al. [30] to improve the upsampling output of a reduced Demucs architecture. Autoencoder architectures, as used in HDemucs, are known to produce tonal and filtering artifacts during upsampling [31]. Pons et al. found these can be alleviated by disabling the biases in all convolution layers, adopting a higher sampling frequency to counteract high-frequency attenuation, and deactivating the Gaussian Error Linear Unit (GELU) activation function in the most external layers [30]. We were able to reproduce these findings for the full HDemucs architecture by feeding a sine signal into an untrained and randomly initialized network, allowing for the spectral influence of the network architecture on the inferred output to be assessed. In Figure 2, it can be seen that for the original architecture spectral replica of the sine wave at $10\,\mathrm{kHz}$, $11\,\mathrm{kHz}$, $12\,\mathrm{kHz}$, $13\,\mathrm{kHz}$, and $23\,\mathrm{kHz}$ exist, whereas the modified architecture shows these spectral lines reduced by up to $10\,\mathrm{dB}$ while the original sine wave is preserved. The checkerboard patterns are typical for autoencoders using transposed convolution layers [30].

The modified HDemucs architecture designed for this study is defined by the following changes:

- Based on the recommendations in [30], the GELU activation functions in the most external layers and the biases in each convolutional layer were deactivated, which is shown in Figure 1. The sample rate is set to $48\,\mathrm{kHz}$.
- The input chunk size is lowered to $16\,384$ (from $441\,000$), indirectly increasing the network's capacity by using less audio data as input frame, which allows for contextual processing. During preliminary testing, it was found that using even smaller chunk sizes produced substantially reduced subjective audio quality.
- A Long Short-Term Memory (LSTM) is added to the fourth layer shown in Figure 1, which was found to improve spatialization during preliminary subjective evaluation.
- The number of estimated outputs was set to one.
- Normalization was deactivated, which was based on the preliminary objective evaluation that revealed an unstable signal level at the output due to the use of smaller chunk sizes.

3.2. Objective

The original HDemucs was trained to minimize the L1-norm between the predicted and target waveform, which enforces a strict adherence to the time-domain target and its absolute phase. However, this objective did not produce the desired acoustic quality for the NBU. Instead, the commonly used spectra-based loss functions *Spectral Convergence* and *Spectral Log-Magnitude* are used [32, 33].

Spectral convergence (ℓ_{SC}) quantifies the difference between the magnitudes of the spectra from the Short-Time Fourier Trans-

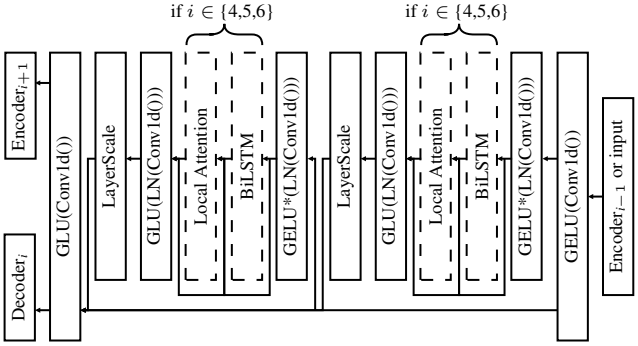


Figure 1: *Representation of the compressed residual branches that are added to each encoder layer reproduced from [20] with i marking the individual layers of encoder and decoder. For the 4th, 5th and 6th layer, a BiLSTM and a local attention layer are added. GELU layers marked with $*$ of encoder and decoder are replaced with Identity layers for $i = 1$.*

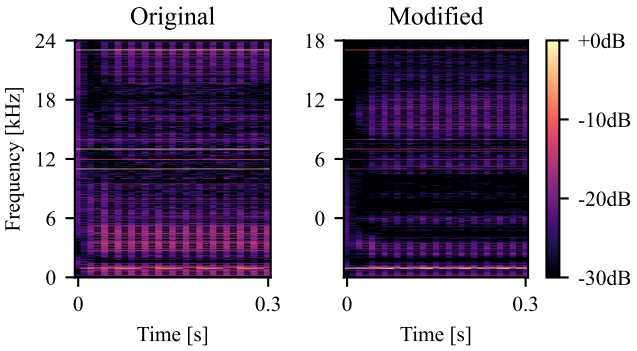


Figure 2: *Spectrum of a* 1 kHz *sine wave after random initialization for original and modified architecture.*

form (STFT) of a given target and predicted signal [32] as,

$$\ell_{\text{SC}}(\hat{s}, \tilde{s}) = \frac{\| \, |\text{STFT}(\tilde{s})| - |\text{STFT}(\hat{s})| \, \|_{\text{F}}}{\| \, |\text{STFT}(\tilde{s})| \, \|_{\text{F}}}. \tag{3}$$

Spectral log-magnitude (ℓ_{SM}) quantifies the difference between log-magnitudes of the spectra from the STFT of a given target and predicted signal [32] as,

$$\ell_{\text{SM}}(\hat{s}, \tilde{s}) = \frac{1}{N_S} \|\log(|\text{STFT}(\tilde{s})|) - \log(|\text{STFT}(\hat{s})|)\|_1, \tag{4}$$

where $||\cdot||_{\text{F}}$ is the Frobenius norm, $||\cdot||_1$ is the L1-norm, and N_S is the number of STFT frames [32]. ℓ_{SC} and ℓ_{SM} are averaged over channels and batches.

As phase is an essential aspect of binaural audio, a phase loss component is also introduced. Leng et al. [27] proposed a phase loss function that measures the phase discrepancy between a given target and predicted signal across frequency bins from the real $\Re$ and imaginary $\Im$ part of their complex frequency domain representations such that

$$\ell_{\text{P}}(\hat{s}, \tilde{s}) = \left| \arctan\left(\frac{\Im(\hat{s})}{\Re(\hat{s})}\right) - \arctan\left(\frac{\Im(\tilde{s})}{\Re(\tilde{s})}\right) \right|. \tag{5}$$

To ensure stable estimation, the phase loss function only evaluates frequency bins in the STFT that have a magnitude greater than 0.1 [27]. The phase differences are averaged over bins, channels, and batches.

The final loss for training aggregates spectral convergence, spectral log-magnitude, and phase loss as

$$\ell(\hat{s}, \tilde{s}) = \ell_{\text{SC}}(\hat{s}, \tilde{s}) + \ell_{\text{SM}}(\hat{s}, \tilde{s}) + \ell_{\text{P}}(\hat{s}, \tilde{s}). \tag{6}$$

As the loss terms lie in a similar range, they were weighted equally. To minimize the influence of the STFT processing, the random-resolution approach from [32] is utilized. During each loss calculation, this method randomly chooses the frame size, window type, and hop size from a default set of values as introduced in [32].

3.3. Training

Three networks were trained on eight NVIDIA A100 GPUs with $\ell(\hat{s}, \tilde{s})$ using the Adam optimizer with a learning rate of 0.001 and an exponential decay of $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1^{-8}$ and weight decay of $2 \cdot 10^{-6}$. In all cases, a maximum runtime of 24 hours and early stopping with a patience of 20 epochs was used. Two different datasets, outlined in Section 4.1, were used. NBU$_S$, utilizing the Studio dataset. NBU$_C$ and NBU$_{C+}$ were trained on the Cambridge MT dataset. For NBU$_{C+}$, silence was added to the training data as described in Section 5.1.

4. EXPERIMENTAL SETUP

4.1. Datasets

In this study, two audio datasets were utilized: Cambridge MT [34] and Studio Mix. The Cambridge MT dataset consists of more than 500 professionally recorded songs for which all single-track stems are available, including more than 15 000 mono and stereo single source tracks with a sample rate of 44.1 kHz. The audio tracks are primarily supplied raw, without additional effects.

The Studio Mix dataset consists of 30 hours of studio-produced music mixed for a 9.1.4 speaker setup of diverse genres (Pop: 24.5%, Electronic: 16.2%, Jazz: 13.5%, Classical: 12.3%, Hip-Hop: 12.2%, Country: 12.1%, Rock: 9.2%). This dataset provides professionally mixed and mastered immersive mixes.

Two datasets are created from these datasets under the assumption that the sole difference between the stereo and the immersive binaural mix is the spatial processing techniques employed. Therefore, both stereo and binaural mixes are derived from a common immersive mix, employing amplitude panning and a generic binauralizer for the stereo and binaural versions, respectively. The following section explains the data processing in detail.

4.2. Data Processing

The Cambridge MT dataset is employed to create a synthetic dataset featuring randomly chosen sources, positions, and augmentations. This dataset serves as a benchmark for comparing against fully mixed songs in the Studio Mix dataset. The comparison aims to assess the DNN's utilization of musical information, like common rhythm, pitch, and phrases, which are only present in the Studio Mix dataset. The results are discussed in Section 5. To construct the Cambridge MT dataset, the stereo and binaural excerpts are generated through the following steps:

Table 1: *Positions of the fixed weights for the soft-panning downmix, derived from loudspeaker positions of the ITU standard 7.1 configuration [37] and weighted by [38] with* $g = \frac{1}{\sqrt{2}}$. *Weights are placed on a sphere with unit radius.*

Speaker	C	FL	FR	SL	SR	SBL	SBR
Azimuth	0	-30	30	-90	90	-135	135
Weight L	0.5	g	$1-g$	1	0	g	$1-g$
Weight R	0.5	$1-g$	g	0	1	$1-g$	g

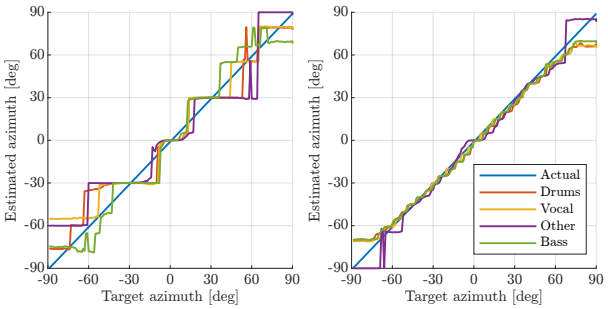
1. **Excerpt Selection and Augmentation:** A 10-second non-empty excerpt was randomly selected from the available stems. This segment undergoes augmentation with a 50% probability, where two effects, such as Reverb, Chorus, Compression, Delay, Phaser, and Distortion, were applied using Pedalboard [35]. All parameters were randomized to lie within the default value and zero.
2. **Combination of Excerpts:** Multiple excerpts were concatenated to create continuous audio tracks of one-minute length.
3. **Spatial Positioning:** The tracks were assigned random positions with $-180° \leq \theta \leq 180°$ and $0° \leq \varphi \leq 90°$. Steps 1-3 are repeated 5 times to yield 5 augmented audio sources, each associated with a distinct spatial position.
4. **Binauralization:** The audio was binauralized based on the positional metadata using a reference proprietary binaural rendering solution consisting of HRTF convolution and reverb. We propose that the binaural rendering solution can be considered a black box in this use case, and as the process is time-invariant, and the same HRTFs are always used, any good quality binauralizer using appropriately pre-processed HRTFs (see [36]) should be sufficient. This binaural mix serves as the target output for the DNN.
5. **Downmix:** A stereo version was created from the mix using the positional metadata and a soft-panning downmix algorithm, which linearly interpolates weighting coefficients for each source from a set of fixed weights, as shown in Table 1. The final output is the summation of all downmixed signals, normalized to ± 1 to prevent clipping. This stereo downmix is the input vector for the DNN.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The performance of the NBU is assessed both objectively, through comparison of spectral features and source position estimation, and subjectively with a listening test (see Table: 2 for an overview of all evaluated conditions). The mono sources used for predicted azimuth estimation, as well as the music excerpts used for spectral comparison and subjective tests, are defined as:

Mono Sources: Four 20 s excerpts of the same song from the MUSDB18HQ dataset containing drums, vocals, bass, and others, mixed to mono.

Music: In a preliminary subjective listening session, five 10 s excerpts of songs from a library of 80 immersive studio mixes from different genres were selected for their extensive utilization of spatial effects like spatial panning and reverberation. It is important to note that these songs were not included in training dataset. Binaural reference and stereo downmix are rendered according to Section 4.2, and the files are subsequently normalized to K-weighted -20 LUFS.



(a) NBU$_S$ (b) NBU$_{C+}$

Figure 3: *Single source azimuth estimation of the output of all NBUs for each source, position and network. The blue line indicates the target position.*

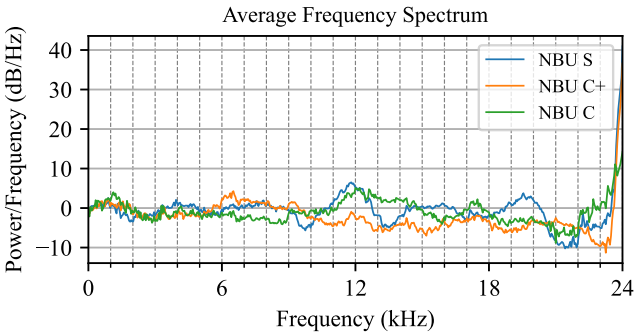


Figure 4: *Average difference in long-term average spectra over all music items for all trained networks compared against the binaural reference. Negative values indicate a loss of energy compared to the reference and vice-versa.*

5.1. Spectral Differences

During training, it was found that introducing silence to the training data reduces tonal artifacts while also decreasing overall output energy at higher frequencies. For the training of NBU$_{C+}$, randomly positioned silence segments of 3000 samples were inserted into sections of the chunk. To avoid signal discontinuities, half of a hamming window 100 samples in length was used to crossover the transition. This additional silence was found to be a good trade-off between tonal artifacts and loss of high-frequency audio content. In Figure 4 it can be seen that NBU$_{C+}$ exhibits a loss of energy at frequencies between 9 kHz and 18 kHz of up to 8 dB/Hz. In contrast, NBU$_C$ deviates only up to 5 dB/Hz, while for this network, the tonal artifacts are perceived loudly. As the strong tonal artifacts are perceived as far more disturbing than the loss of high-frequency audio content, NBU$_C$ is dropped in favor of NBU$_{C+}$ for further evaluation. NBU$_S$ was trained without added silence as this resulted in little tonal artifacts already while preserving high-frequency content well, although it deviates from the reference up to 6 dB/Hz in the region of 9 kHz to 14 kHz.

Table 2: *The six conditions of the listening test and their creation process.*

Condition	Description
Binaural Reference (Ref)	Studio-produced 9.1.4 Dolby Atmos mix binauralized with speaker positions according to the Dolby Atmos layout [39].
Binauralized Downmix (BD)	Studio-produced 9.1.4 Dolby Atmos mix downmixed using the soft panning downmix with speaker positions using the Dolby Atmos layout [39]. The two-channel output is then binauralized, with the left and right channel positioned at $\theta = \pm 45^\circ$ and $\varphi = 0^\circ$.
Halo Upmix	The downmix is processed by Halo Upmix, a commercially available upmixer designed to take a stereo, or surround, mix and create a higher channel count upmix of up to 7.1.4 while maintaining the spatial balance and image of the original mix and is used as a baseline. A preset designed for maximizing the spaciousness of musical content was used to define the upmix parameters. The upmixed output is binauralized with speaker positions according to the Dolby Atmos layout [40].
NBU$_S$	Modified HDemucs architecture trained on studio-produced 9.1.4 music.
NBU$_{C+}$	Modified HDemucs architecture trained on Cambridge MT dataset with added silence in the dataloader.

5.2. Localization Analysis

The NBU's spatial rendering accuracy is quantified by estimating the perceived azimuth of individual sources, utilizing the probabilistic model proposed by May et al. in [41]. The framework is based on supervised learning of azimuth-dependent binaural feature maps using Interaural Time Differences (ITDs) and Interaural Level Differences (ILDs) estimated from a binaural signal. It approximates the human auditory system using a gammatone filter bank, half-wave rectification, and low-pass filtering [41]. The localization model is restricted to the frontal horizontal plane; consequently, the angular range of the test sources was limited to $-90^\circ \leq \theta \leq 90^\circ$, and evaluation is performed for up to four simultaneous sources.

Single Source: The four mono sources are individually placed within the defined acoustic scene, which is rendered using the soft-panning downmix and subsequently inferred by the networks. In Figure 3, it can be seen that NBU$_S$ exhibits step-wise positioning of sound sources, while NBU$_{C+}$ closely follows the target position. This can be attributed to the discrete number of loudspeaker positions in the Studio Mix dataset, limiting the network's ability to place sources in between these positions. In contrast, NBU$_{C+}$ was trained with sources placed in the entire azimuth range, which allows for higher spatial rendering resolution, as shown by the smaller mean absolute angle difference (Table 3).

Multiple Sources: Two to four simultaneously playing sources are placed within the acoustic scene at random positions, maintaining a minimum distance of 10° between all sources to allow for source identification in the estimation analysis. Each source is used up to once per scene. One hundred scenes are rendered for two, three, and four sources, and the distance between the actual and estimated position for each source is calculated. In Table 4, it can be seen that the mean error of NBU$_S$ remains constant while the mean error of NBU$_{C+}$ doubles from four to one source. This correlates with the number of sources present in the training data, as for NBU$_S$ fully mixed songs are used and for NBU$_{C+}$ often only one to two sources are playing at a given time.

The analysis suggests that, at least for up to four sources in the frontal hemisphere, both NBU are capable of independently positioning each source within, on average, a positional error of 11.3° or less from the expected source position.

5.3. Subjective Analysis: Listening Test

Multiple approaches to upmixing and binauralization of stereo content are described in Table 2. This includes Halo Upmix [42] that was used as a baseline. These approaches were compared in two separate listening tests to assess spatial attributes and overall audio quality using WebMUSHRA [43].

Table 3: *Mean and standard deviation of absolute difference to target azimuth in degrees for all networks per source type and for all mono sources.*

Source	Drums	Vocal	Other	Bass
NBU$_S$	9.2±6.6	9.6±7.7	12.3±8.7	8.6±5.5
NBU$_{C+}$	**5.0±4.9**	**4.2±5.0**	**4.8±4.4**	**4.4±4.7**

Table 4: *Mean and standard deviation of absolute difference to target azimuth for all networks for one to four simultaneous sources.*

Sources	One	Two	Three	Four
NBU$_S$	10.0±7.1	11.1±7.8	11.5±8.4	11.3±9.1
NBU$_{C+}$	**4.6±4.8**	**7.0±8.6**	**6.5±7.0**	**9.1±7.9**

The first listening test was taken by seven expert listeners, who either work in binaural and/or are expert binaural listeners. The listening test was structured in two phases. Phase I considered the spatial attributes 'spatial clarity' and 'spaciousness' as used in prior work [44, 45], the descriptions for each attribute are defined in Table 5. The subjects were asked to rate all conditions for the given attribute on a scale between 0 and 100, labeled less and more, respectively. Listener judgments were purely comparative, as no hidden reference was provided. Phase II of the listening test was a standard MUSHRA test considering overall audio quality. The results are depicted in Figure 5.

For all statistical tests, the significance level is set to $\alpha \leq 0.05$. Results of a Shapiro-Wilk test conclude that the data is not normally distributed. Therefore, the data is analyzed using the Kruskall-Wallis test, which shows statistically different groups exist in each attribute (spatial clarity: $p = 5.69 \times 10^{-17}$, spaciousness: $p = 8.86 \times 10^{-17}$ and audio quality: $p = 1.36 \times 10^{-16}$). A follow-up Dunn-Bonferroni test is performed to analyze the upmix approaches pairwise. For spatial clarity, the reference is rated significantly better than other approaches, and NBU$_{C+}$ significantly worse than NBU$_S$ ($p = 2.81 \times 10^{-4}$). No significant differences are found between NBU$_S$, BD, or Halo. The reference is rated most spacious with a statistical difference from all others, while no differences can be found between BD, Halo, and NBU$_{C+}$. NBU$_S$ is rated significantly more spacious than all

Table 5: *Attribute description for Phase I.*

Attribute	Description
Spatial Clarity	Impression of how clearly different elements in a scene can be distinguished from each other, and how well various properties of individual scene elements can be detected.
Spaciousness	Describes how much the sound appears to surround you.

approaches except the reference (BD: $p = 9.15 \times 10^{-3}$, Halo: $p = 9.79 \times 10^{-3}$, NBU_{C+}: $p = 4.47 \times 10^{-7}$, and Reference: $p = 8.95 \times 10^{-2}$). The reference is rated significantly better for audio quality, while no significant differences are observed between Halo and BD. NBU_{C+} is rated worst with statistical significance ($p = 1.55 \times 10^{-4}$).

A second test with naive binaural listeners was performed to give insight into the performance of NBU_S and the impact of neural upmixing for naive listeners. Based on feedback from the first test, the scales were adjusted to be in a range of -50 to 50 with labels provided as [-50 -30]: Much Worse, [-30 -10]: Worse, [-10 10]: Same, [10 30]: Better, [30 50]: Much Better. Of seventeen participants, five had to be excluded due to inconsistent identification of the hidden reference. In Figure 6, it can be seen that NBU_{C+} performed worst for all attributes, and there is only a slight trend to prefer the reference in spaciousness. The remaining conditions were rated comparably.

Similar to the first listening test, the data is analyzed using the Dunn-Bonferroni as data is not normally distributed. NBU_{C+} is rated worst ($p \leq 7.4 \times 10^{-13}$) in spatial clarity, while no significant differences are found between all other conditions. Regarding spaciousness the reference is rated significantly better than other approaches ($p \leq 1.7 \times 10^{-2}$) and NBU_{C+} is rated worst ($p \leq 8.9 \times 10^{-4}$), while no significant differences are found between BD, Halo, and NBU_S. The reference is rated significantly better than all other conditions except NBU_S ($p \leq 3.2 \times 10^{-2}$) with respect to audio quality. Furthermore, NBU_{C+} is rated as performing significantly worse ($p \leq 1.3 \times 10^{-7}$), while there are no significant differences between BD, Halo, and NBU_S. The absence of differences in spatial clarity between the reference and other methods motivated a detailed examination of individual outcomes. It was observed that participants who favored the reference also tended to prefer NBU_S over BD and Halo. The lack of head tracking, which is known to improve the externalization, especially for frontal and rear sources [46], may have been a contributing factor for naive binaural listeners perceiving little difference between the conditions and is a topic for future research.

The results of the binaural expert listening tests show significant improvement in spaciousness and a trend towards more spatial clarity with small quality deterioration for NBU_S relative to all other approaches. These results could not be reproduced with naive listeners, which could possibly be due to difficulties for naive listeners in rating spatial attributes. Furthermore, the absence of head tracking might amplify the perceived similarity across all methods. While the objective evaluation of Section 5.2 suggested NBU_{C+} to be more precise in spatial positioning than NBU_S, NBU_{C+} falls behind in all subjective evaluation, which is possibly due to the notable loss in high-frequency content, as discussed in Section 5.1.

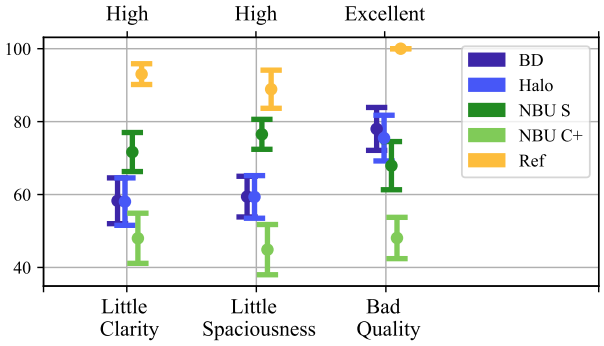


Figure 5: *Listening test results for binaural expert listeners.*

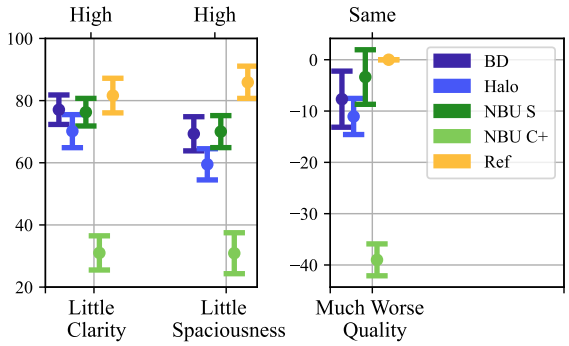


Figure 6: *Listening test results for naive listeners.*

6. CONCLUSION AND FUTURE WORK

We presented a data-driven approach to binaural upmixing of stereo content. The key contributions of this work include an end-to-end approach to binaural upmixing, a transfer of source separation to binaural upmix domain for the HDemucs architecture, and the objective and subjective analysis of neural networks trained with two different datasets. While training with a synthetically constructed dataset improved the network's accuracy in positioning sources in space, training with professionally mixed immersive music, for which instrument positioning is more consistent, yielded significantly higher subjective performance in two listening tests.

The resulting end-to-end upmixing approach can effectively transform stereo into immersive binaural as it shows the capability to position up to four simultaneous sources within, on average, a positional error of 11.3 ° or less from the expected source position. It also demonstrates a significant improvement in spatialization for headphone playback as perceived by expert binaural listeners, albeit with a slight compromise in audio quality, which shows an NBU could enhance the large body of legacy content. Notably, these enhancements were not perceptible to naive listeners, potentially due to the absence of head tracking and possible difficulties for untrained listeners to rate spatial attributes.

For future work, it would be worthwhile exploring real-time capable architectures to facilitate the integration of head-tracking, which could further enhance perceived spaciousness. Another potential area of exploration is to extend the application domain to non-musical content, such as movie soundtracks, thereby broadening its applicability.

7. REFERENCES

[1] Michael Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, Springer International Publishing, Cham, 2020.

[2] Bob Mehr, "Dolby Atmos Wants You to Listen Up. (And Down. And Sideways.)," https://www.nytimes.com/2023/06/21/arts/music/dolby-atmos.html, accessed November 22, 2023.

[3] Apple, "Apple Spatial Audio," https://www.apple.com/newsroom/2021/05/apple-music-announces-spatial-audio-and-lossless-audio/, accessed November 17, 2023.

[4] TIDAL, "Tidal," https://tidal.com/partners/dolbyatmos, accessed November 22, 2023.

[5] Brian Hiatt, "The Beatles in Spatial Audio," https://www.rollingstone.com/music/music-features/beatles-best-spatial-audio-albums-apple-music-abbey-road-giles-martin-1202832, accessed November 29, 2023.

[6] T Walton, "The Overall Listening Experience of Binaural Audio," *4th International Conference on Spatial Audio (ICSA)*, 2017.

[7] Carlos Avendano, "A Frequency-Domain Approach to Multichannel Upmix," *J. Audio Eng. Soc.*, vol. 52, no. 7, pp. 740–749, 2004.

[8] Andreas Walther and Christof Faller, "Direct-ambient decomposition and upmix of surround signals," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 277–280.

[9] Yong-Hyun Baek, Se-Woon Jeon, Seok-Pil Lee, and Young-Cheol Park, "Efficient Primary-Ambient Decomposition Algorithm for Audio Upmix," *Journal of Broadcast Engineering*, vol. 17, no. 6, pp. 924–932.

[10] Karim M. Ibrahim and Mahmoud Allam, "Primary-Ambient Source Separation for Upmixing to Surround Sound Systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 431–435.

[11] Jeonghwan Choi and Joon-Hyuk Chang, "Exploiting Deep Neural Networks for Two-to-Five Channel Surround Decoder," *J. Audio Eng. Soc.*, vol. 68, no. 12, pp. 938–949, 2021.

[12] Alexander Poets and Stephan Preihs, "On the Evaluation of Perceived Spatial Immersion in the Application of Automatic Upmixing for 3D Surround Sound Systems," *DAGA Deutsche Jahrestagung für Akustik*, pp. 977–980, 2023.

[13] Taegyu Lee, Yonghyun Baek, Young-cheol Park, and Dae Hee Youn, "Stereo upmix-based binaural auralization for mobile devices," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 411–419, 2014.

[14] Sebastian Kraft and Udo Zölzer, "Stereo signal separation and upmixing by mid-side decomposition in the frequency-domain," *Proc. of the 18th Int. Conference on Digital Audio Effects*, pp. 1–6, 2015.

[15] Mark Vinton, David Mcgrath, Charles Robinson, and Phillip Brown, "Next generation surround decoding and upmixing for consumer and professional applications," *AES 57th International Conference*, pp. 1–9, 2015.

[16] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 426–430.

[17] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Guildford, UK, July 2–5, 2018, Proceedings 14*, 2018, pp. 293–305.

[18] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR, 2018*, Paris, France, 2018, pp. 334–340.

[19] Simon Rouard, Francisco Massa, and Alexandre Défossez, "Hybrid transformers for music source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[20] Alexandre Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Online, 2021.

[21] Minseok Kim, Woosung Choi, Jaehwa Chung, Daewon Lee, and Soonyoung Jung, "Kuielab-mdx-net: A two-stream neural network for music demixing," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, Online, 2021.

[22] Yi Luo and Jianwei Yu, "Music source separation with band-split rnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[23] Papers with Code, "Papers with Code - MUSDB18 Benchmark (Music Source Separation)," hrefhttps://paperswithcode.com/sota/music-source-separation-on-musdb18, accessed November 29, 2023.

[24] Agnieszka Roginska and Paul Geluso (eds), *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*, Routledge, Taylor & Francis Group, New York; London, 2018.

[25] Alexander Richard, Dejan Markovic, Israel D Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, and Yaser Sheikh, "Neural synthesis of binaural speech from mono audio," in *International Conference on Learning Representations*, Online, 2020.

[26] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, p. 125.

[27] Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiangyang Li, Tao Qin, et al., "Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23689–23700, 2022.

[28] Francesc Lluís, Vasileios Chatziioannou, and Alex Hofmann, "Points2sound: from mono to binaural audio using 3d point

cloud scenes," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, pp. 1–15, 2022.

[29] Afonso Lopes, José Ricardo Barboza, and Gilberto Bernardes, "Instrument position in immersive audio: An empirical review of award-winning practices," *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–8, 2023.

[30] Jordi Pons, Santiago Pascual, Giulio Cengarle, and Joan Serrà, "Upsampling artifacts in neural audio synthesis," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3005–3009.

[31] Jordi Pons, Joan Serra, Santiago Pascual, Giulio Cengarle, Daniel Arteaga, and Davide Scaini, "Upsampling layers for music source separation," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 311–315.

[32] Christian J Steinmetz, Jordi Pons, Santiago Pascual, and Joan Serrà, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 71–75.

[33] Ben Hayes, Charalampos Saitis, and György Fazekas, "Neural waveshaping synthesis," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, (ISMIR)*, Online, 2021, pp. 254–261.

[34] Cambridge MT, "The 'Mixing Secrets' Free Multitrack Download Library," accessed November 29, 2023, https://www.cambridge-mt.com/ms/mtk/.

[35] Peter Sobot, "Pedalboard," July 2021, https://doi.org/10.5281/zenodo.7817838.

[36] Cal Armstrong, Lewis Thresh, Damian Murphy, and Gavin Kearney, "A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database," *Applied Sciences*, vol. 8, no. 11, pp. 2029, 2018.

[37] "RECOMMENDATION ITU-R BS.775-4*," Tech. Rep., ITU-R, 2022.

[38] "ATSC Standard: Digital Audio Compression (AC-3, E-AC-3) doc. a/52:2018," Tech. Rep., Advanced Television Systems Committee, 2018.

[39] "9.1.4 overhead speaker setup," Tech. Rep., Dolby, Atmos, 2023.

[40] "7.1.4 overhead speaker setup," Tech. Rep., Dolby, Atmos, 2022.

[41] Tobias May, Steven Van De Par, and Armin Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 1–13, 2010.

[42] NUGEN Audio, "Halo upmix," 2022, https://nugenaudio.com/haloupmix/.

[43] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, pp. 8, 2018.

[44] Gregory Reardon, Andrea Genovese, Gabriel Zalles, Patrick Flanagan, and Agnieszka Roginska, "Evaluation of binaural renderers: multidimensional sound quality assessment," in *AES International Conference on Audio for Virtual and Augmented Reality*. Audio Eng. Soc., 2018.

[45] Alexander Lindau, "Spatial audio quality inventory (saqi): Test manual. v1.2," Tech. Rep., Technische Universität Berlin, 2015.

[46] Etienne Hendrickx, Peter Stitt, Jean-Christophe Messonnier, Jean-Marc Lyzwa, Brian FG Katz, and Catherine de Boishéraud, "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *J. Audio Eng. Soc.*, vol. 141, no. 3, pp. 2011–2023, 2017.