A DEEP LEARNING APPROACH TO THE PREDICTION OF TIME-FREQUENCY SPATIAL PARAMETERS FOR USE IN STEREO UPMIXING

Daniel Turner *

Creative Labs
Staines-on-Thames, UK
daniel_turner@cle.creative.com

Damian T. Murphy

AudioLab
Department of Physics, Engineering, and Technology
University of York
York, UK
damian.murphy@york.ac.uk

ABSTRACT

This paper presents a deep learning approach to parametric time-frequency parameter prediction for use within stereo upmixing algorithms. The approach presented uses a Multi-Channel U-Net with Residual connections (MuCh-Res-U-Net) trained on a novel dataset of stereo and parametric time-frequency spatial audio data to predict time-frequency spatial parameters from a stereo input signal for positions on a 50-point Lebedev quadrature sampled sphere. An example upmix pipeline is then proposed which utilises the predicted time-frequency spatial parameters to both extract and remap stereo signal components to target spherical harmonic components to facilitate the generation of a full spherical representation of the upmixed sound field.

1. INTRODUCTION

Audio upmixing can be described as the process of generating additional channels of audio data when the original signal contains fewer channels than the target reproduction system. Many of the upmix algorithms in the literature provide channel-based upmixing as they aim to generate additional signals to directly drive additional loudspeakers in a known configuration [1, 2, 3, 4, 5, 6, 7], such as 5.1, or by using methods such as (VBAP) [8] to upmix to arbitrary 2D or 3D configurations [3]. Within this context, upmix algorithms can be more simply defined as generating a higher number of output channels from a smaller number of input channels. For example, the program material may consist of a stereo recording where the target system is a 5.1 configuration and as such requires five full-range signals and one band-limited low-frequency signal. This paper proposes a scene-based approach to upmixing using stereo to B-format upmixing as an example application based on a deep learning approach to the prediction of time-frequency spatial parameters which can be utilised as a part of a novel upmixing algorithm. Scene-based audio usually refers to methods which spatially encode a sound field into a number of specified channels, which collectively describe the spatial characteristics of the sound field and can later be decoded to a chosen loudspeaker configuration.

2. RELEVANT BACKGROUND

Upmixing can be broadly classified into one of two categories. The first is upmixing as decoding, where an algorithm upmixes or decodes multi-channel content that has been previously encoded [9]. For instance, Dolby Pro-Logic encoding/decoding can encode 4-channel, 5-channel, and 7-channel surround sound into a two-channel matrix encoded signal that can itself be decoded to retrieve an approximation of the original multi-channel signals [10]. These algorithms are effective as the encoded input signal often contains signal cues such as relative channel phase, which can be used to aid the upmix process. The second, blind upmixing, is where additional channels are generated based solely on analysis of the input signal. As the vast majority of stereo content has not been downmixed from existing multi-channel content, the method proposed in this paper can be considered as belonging to the latter category.

Many stereo upmixing methods decompose a stereo signal into direct signal components and diffuse signal components, sometimes also referred to as the primary and ambient components respectively, by first transforming the signal into the time-frequency domain using techniques such as Short-Time-Fourier-Transform (STFT) [1, 4, 5, 11, 12]. Decomposition in the time-frequency domain enables more effective separation of temporally overlapping sources. Direct components are defined as those signal components that are highly correlated with existing channels and diffuse components are those signal components that have low correlation with the existing channels [13]. For a detailed review of existing direct and ambient decomposition methods see [14].

Several machine learning approaches to upmxing have been presented in recent years, although they have predominately focused on channel-based methods. Ibrahim and Allam [13] approach the task of direct-diffuse composition as a classification problem, training a feed-forward Neural Network (NN) to classify each complex valued time-frequency tile as either direct or diffuse. When used as part of an upmixing system to upmix from stereo to a quad array, 10 out of the 11 listeners preferred the NN method above traditional methods such as those proposed in [6] and [1], as well as achieving the highest signal to distortion ratio which was tested on each of the extracted direct and ambient components.

Park et. al. [7] proposed a deep neural network (DNN) to upmix from stereo to 5.1 within the MPEG-H 3D framework [15]. A DNN was trained using log-spectral magnitudes of quadrature mirror filter subbands to predict the center and surround channels from the input stereo signals. The input signals are then mapped in the subband space to the center and surround channels where they are transformed back into audio signals via quadrature mirror filter synthesis. The approach is based on the assumption that the center

[*] This work was undertaken while the author was a PhD student at The University of York and was supported by an EPSRC iCASE studentship in collaboration with BBC Research and Development

channel is some combination of the left and right channels and the surround channels are derived as some amount of the difference between two channels.

The method proposed in [2], utilises two DNNs, with one trained to perform direct-diffuse decomposition and the other to render the diffuse component. Both networks are trained to jointly minimise the Mean Squared Error (MSE) between the magnitude spectra of the original and the upmixed/decoded five-channel signal as well as minimising the loss for the Inter-Channel Level Difference (ICLD). The network predicts spectral weights which are then multiplied with each frequency bin in the stereo signal and serve as a mask to separate the direct and diffuse components. In all cases, the current methods are concerned with deriving signals to directly drive additional loudspeakers for use within channel-based upmixing.

There are, however, some limitations to the current approaches for stereo upmixing, particularly around the directional estimation of components. Stereo signals traditionally only account for a source's lateral position, providing insufficient information for traditional methods to discern its elevation or whether it is positioned in front or behind the microphone capture array. It is the practice of stereo signals being reproduced over frontally placed loudspeakers that introduces a conceptual *front* and *back*. Upmixers aim to enhance this representation by generating ambience around the listener that seeks to simulate the reflections and reverberation of the recorded or synthesized environment [16]. They in effect create a frontally focused sound field with additional surrounding ambience, which is generally adequate for traditional screen-based media where the action will be coming from the front and therefore the attention of the audience will be directed towards the front. This approach, however, introduces challenges for stereo signals recorded in real environments as sources may be located at varying positions on both the median and horizontal planes.

Consider an example where a spaced stereo microphone pair is placed in the center of 4 loudspeakers positioned at azimuth, $\theta = 45^\circ, 135^\circ, 225^\circ, 315^\circ$. A sound is played from each speaker sequentially, starting with the speaker at 45° and continuing in an anti-clockwise direction. Traditional methods of panning estimation would yield near identical values for the sources at 45° and 135° as well as identical values for those positioned at 225° and 315°. The identical value pairs are a result of traditional stereo localisation estimation methods being limited to the lateral position, usually based on either the Time Difference of Arrival (TDOA) between the two microphone signals or the inter-channel amplitude difference. Subsequently, were these signals to be upmixed using systems such as those proposed in [1, 7, 17, 18] and reproduced over a 5.1 configuration the perception of source movement around the array would not be congruent with that observed during the recording. Instead, the direct components of the two source positions at 45° and 135° would be reproduced at the front left of the array, and the two sources at 225° and 315° reproduced at the front right, whilst the surround speakers would predominately contain the decorrelated diffuse component.

The work presented in this paper aims to develop a deep learning approach where, given appropriate input features containing time, amplitude, and phase information, an NN can be trained to approximate a mapping function that predicts spatial features for a 360° space from the information contained within and derived from a stereo signal. These spatial features can then be used to facilitate upmixing methods that move away from frontally biased systems to ones that aim to reproduce a sound field that approximates the spatial characteristics that would have been present at the time of recording.

3. METHODS

3.1. Model

The MuCh-Res-U-Net architecture proposed in this paper combines the multi-channel U-Net approach detailed in [19] with a similar Residual-U-Net backbone to that used in [20] to form a 9-level Res-U-Net architecture with a multi-channel output equal to the number of predicted time-frequency parametric spatial features. Each encoding and decoding block consists of two sequentially stacked convolutional blocks which contain a batch normalisation layer, a Leaky ReLU activation layer, and a convolutional layer.

The original U-Net architecture, developed for image segmentation tasks [21], has been shown to be effective when applied to a number of audio-related tasks including source separation [19, 22, 23], voice conversion and cloning [24, 25], denoising [26], and audio synthesis [27, 28]. Additionally, U-Net style architectures lend themselves to tasks where the input and output data are of similar dimensions due to the symmetry of the encoder and decoder paths.

3.2. U-Net

The original U-Net consists of an encoding path and a decoding path with skip connections that are passed from the encoding layer to the corresponding decoding layer, identical to those shown in Figure 1 which depicts a representation of the MuCh-Res-U-Net architecture proposed in this paper. The encoding path is similar to traditional convolutional neural networks (CNN) where the resolution of the feature maps decreases through consecutive layers while the number of feature maps/number of filters increases. The encoded data is then transformed within the latent space of the bottleneck encoding block before being decoded through the upsampling of the resulting feature maps in the decoding path. The skip connections allow for the propagation of information from the encoding layers to the decoding layers and serve to preserve and propagate localised features that may otherwise be lost due to the dimensionality reduction of the deeper encoding layers.

3.3. Residual Connections

The residual connections within the encoder and decoder blocks facilitate two main advantages; firstly, they reframe the modeling problem to one of modeling the residual between the input and targets, as opposed to the complete transform from input to target [20]. The residual block can be formulated as in [29]:

$$y_l = \mathcal{F}(x, \{W_i\}) + \mathcal{I}(x) \tag{1}$$

$$x_{l+1} = f(y_l) \tag{2}$$

where x_l and x_{l+1} are the input and output of the residual unit respectively, y_l is the output of layer l, $\mathcal{F}(\cdot)$ is the residual function, $f(\cdot)$ is the activation function, and where $\mathcal{I}(\cdot)$ is the identity mapping function where generally $\mathcal{I}(\cdot) = x_l$.

Secondly, they allow gradients to be back-propagated unimpeded to earlier initial layers due to the nature of derivatives of summation operations. This mitigates the *vanishing gradient problem*, which can cause gradients to approach zero for earlier layers due to sequential multiplications of small numbers[30].

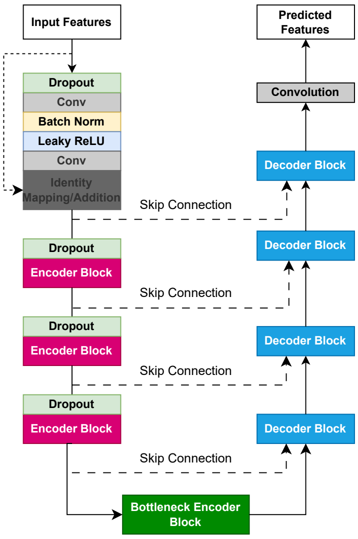


Figure 1: *Proposed MuCh-Res-U-Net architecture*

3.4. Input Features

The input feature vector used to train the model consisted of the short-time log-magnitude spectrum of each stereo channel and the Generalised Cross-correlation with phase transform (GCC-PHAT) [31] to provide the model with both spectral and phase information. The short-time log-magnitude spectrum was derived by first transforming each stereo channel into a time-frequency representation using the STFT, which was calculated using a non-symmetric Hann window with a length of 1024 and a hopsize of 512. This corresponds to a window length of 46 ms at a sampling frequency of 22.05 kHz with 23 ms between the onset of successive frames. The frequency resolution is approximately 21.5 Hz with the lowest detectable frequency being ≈100 Hz. A logarithmic function is then applied to retrieve the log-magnitude spectrum in dB.

The GCC-PHAT was chosen as the phase feature as it is widely used for estimating time difference of arrival (TDOA) and is commonly used for Sound Event Localisation and Detection (SELD) based machine listening tasks [32] and is calculated by first transforming the channels into the frequency domain and combining them through a generalised cross-correlation as defined in [33]:

$$\Psi G[\omega_k] = X_1^*[\omega_k] X_2[\omega_k] \tag{3}$$

where ΨG is the generalised cross-correlation, X_n is the frequency domain representation of the given channel, ω_k is the frequency index in radians, and * represents the complex conjugate of a complex number.

The phase transform (PHAT) is then applied such that the magnitudes are normalised and any effects due to amplitude are eliminated:

$$\Psi P[\omega_k] = \frac{\Psi G[\omega_k]}{|\Psi G[\omega_k]|} \tag{4}$$

The inverse Fast Fourier Transform (iFFT) is then applied which results in a frequency-weighted time-domain cross correlation and is obtained by:

$$\psi P[n] = \mathcal{F}^{-1}\left\{ \frac{\Psi G[\omega_k]}{|\Psi G[\omega_k]|} \right\} \tag{5}$$

where $\mathcal{F}^{-1}$ is the iFFT which results in the feature that will be used as input into the proposed network. The delay between the signals can be estimated by reading the histogram such that:

$$\tau = \text{arg max } \psi P[n] \tag{6}$$

It should be noted that when being used within machine learning applications it is common for the GCC-PHAT to be captured for each time frame resulting in a 2D feature map.

3.5. Target Features

The desired target features were the directional of arrival (DOA) of each time-frequency component in spherical coordinates measured in radians and a diffuseness index. These are common features often used in traditional upmixing systems to extract, reposition, and render the direct and diffuse components of a given signal. As the aim of the network is to predict these features within a 360° space, target features were extracted from the synthesised Ambisonic scenes using Directional Audio Coding (DirAC) analysis [34] using the same STFT parameters as detailed for the input feature extraction.

Directional analysis utilising B-format signals is performed as per [34] and [35], using an energetic analysis of the sound field based on the STFT domain representations of the sound pressure $P(m, \omega_k)$ and particle velocity $\vec{U}(m.\omega_k)$ at the recording position, where m, ω_k are time and frequency indices respectively. The W channel signal is regarded as proportional to the sound pressure, while the three orthogonal pressure gradient signals X, Y, and Z capture signal properties considered to be proportional to sound velocity. This gives the relationship [35]:

$$P(m, \omega_k) = W(m, \omega_k) \tag{7}$$

$$\vec{U}(m, \omega_k) = -\frac{1}{\sqrt{2}Z_0} \vec{X}'(m, \omega_k) \tag{8}$$

where $\vec{X}'(m, \omega_k) = [X(m, \omega_k), Y(m, \omega_k), Z(m, \omega_k)]^T$ is the vector of B-format pressure gradient signals and Z_0 is the characteristic impedance of air. The 3-dimensional instantaneous intensity vector is an estimate of the direction of the net flow of energy and is calculated for each time and frequency index as:

$$\vec{I}(m, \omega_k) = \Re\{\mathrm{E}\{P^*(m, \omega_k)\vec{U}(m, \omega_k)\}\} \tag{9}$$

where $\mathrm{E}\{\cdot\}$ represents a short time averaging operation.

As the intensity vector is said to point in the direction of the net flow of energy, the direction of incidence is defined to be the opposite direction of the intensity vector and points towards the source [34]. This can simply be defined as:

$$\vec{D}(m, \omega_k) = -\frac{\vec{I}(m, \omega_k)}{||\vec{I}(m, \omega_k)||} \tag{10}$$

The resulting matrix $\vec{D}$ contains time-averaged directional of arrival (DOA) estimates for each time-frequency tile. The desired azimuth and elevation angles in radians can be derived from this as follows [36]:

$$\theta = arctan\left(\frac{I_3}{I_1}\right) \tag{11}$$

$$\phi = arccos\left(\frac{I_2}{||\vec{I}||}\right) \tag{12}$$

where I_1, I_2, and I_3 are the first-order channel matrices contained within $\vec{I}$.

The diffuseness index is estimated in the STFT domain as [11]:

$$\psi(m,\omega_k) = 1 - \frac{\sqrt{2}||\Re\{\mathrm{E}\{P^*(m,\omega_k)\vec{U}(m,\omega_k)\}\}||}{|\mathrm{E}\{P^*(m,\omega_k)\}|^2 + ||\mathrm{E}\{\vec{U}(m,\omega_k)\}||^2} \quad (13)$$

where a value of $\psi = 0$ indicates the net flow of energy from a given time-frequency tile corresponds to the total energy within that time-frequency tile. A value of $\psi = 1$ indicates there is no net transfer of acoustic energy within that time-frequency tile and thus indicates a completely diffuse sound field.

Lastly, the short-time averaged energy vector can be derived as in [37]:

$$\vec{E}(m,\omega_k) = |\mathrm{E}\{P^*(m,\omega_k)\}|^2 + ||\mathrm{E}\{\vec{U}(m,\omega_k)\}||^2 \quad (14)$$

3.6. Training and Optimisation

The model contained 154,777,346 trainable parameters and was trained using mini-batch gradient descent with a batch size of 6 and using Adam with decoupled weight decay regularisation [38]. Gradient accumulation [39] was used to increase the effective batch size to 45 and negate issues associated with smaller batch sizes, such as larger inter-batch variance. A learning rate schedule was adopted that consisted of a linear warm-up over 1000 steps to a maximum learning rate of $\eta = 8 \times 10^{-4}$. The learning rate remained static for 2 × warm-up steps before following a scheme of Cosine Annealing with warm restarts [40] with 10 epochs for the initial restart with the number of epochs between subsequent restarts increasing each time by a factor of 2. The training period was 100 epochs, which took approximately 20 hours and 40 minutes.

An adaptive gradient clipping method [41] was utilised to allow a clipping threshold to be set based on the history of gradient norms observed in the training run. This helps minimise the risk of exploding gradients caused by the often non-smooth nature of NN loss landscapes [42] and allows for an appropriate selection of the clipping threshold parameter without including it in a hyper-parameter search. It was set to clip to the 10th percentile of the derived threshold as this would help to ensure any outliers would not have a disproportionate impact on the clipping threshold. The Mean Squared Error (MSE) between the estimated time-frequency parameter values and the ground truth parameter values was used as the loss function and can be defined as:

$$loss(\hat{y}_i, y_i) = \sum_{i=0}^{I} \frac{1}{K} \sum_{k=0}^{K} (\hat{y}_{ik} - y_{ik})^2 \quad (15)$$

where $\hat{y}_{ik}$ is the prediction for the k_{th} time-frequency tile in the i_{th} target feature map. The losses from each feature are summed to get the final loss. The model was trained for a total of 100 epochs.

At the feature extraction stage, before the time-frequency transform, a noise injection layer randomly adds Gaussian noise to the time-domain signals based on a given probability. Noise injection has been shown as an effective regularisation method as it serves as a type of data augmentation to prevent the network overfitting through continuous sampling of the noise inherent in smaller datasets [43, 44]. The amount of noise added is scaled according to each example to achieve an SNR of 20 dB, a value reached through iterative testing.

3.7. Dataset

An investigation into ML-driven upmixing of stereo signals requires a dataset that contains the relevant input-output pairs with which to train and evaluate the model. In the case of this work, a dataset containing equivalent stereo and Ambisonic signals was desired. Equivalent stereo and Ambisonic sound scenes were synthesised using a dataset of stereo and multi-channel IRs for a 50-point Lebedev quadrature sampled sphere, collected by the authors [45], which were convolved with audio files from the NIGENS dataset using a procedure that builds on that proposed in [46]. The IR dataset consists of two-channel stereo IRs for 9 stereo configurations, the 32 channels captured from an Eigenmike, and spherical harmonic components up to 4th order derived from the Eigenmike signals using the Eigenunits plugin [47]. Details of the microphone configurations and capture methodology are presented in [45].

For this initial investigation, the network was trained on a single stereo configuration to limit the complexity of the problem space. The *AB_Omni_40* set was chosen based on the results of an initial set of experiments conducted on 60 training examples to ascertain which stereo configuration had the potential to converge the fastest. It is acknowledged that 60 examples is too small a dataset on which to base any definitive conclusions of training potential, however, as the work was practically limited by available compute power it was decided this would be adequate in deciding on a configuration with which to conduct this initial investigation. Additionally, the omnidirectional signals allow for either channel to be taken as an approximation for the omnidirectional pressure component which will simplify the upmixing pipelines as detailed in Section 4. All scenes generated were 7 seconds in length and a total of 6000 unique scenes were generated. This resulted in a dataset comprising of 11 hours and 36 minutes of sound material with a split of 4500 (75%) examples for training and 750 (12.5%) for each of the validation and test sets.

4. STEREO TO B-FORMAT UPMIXING

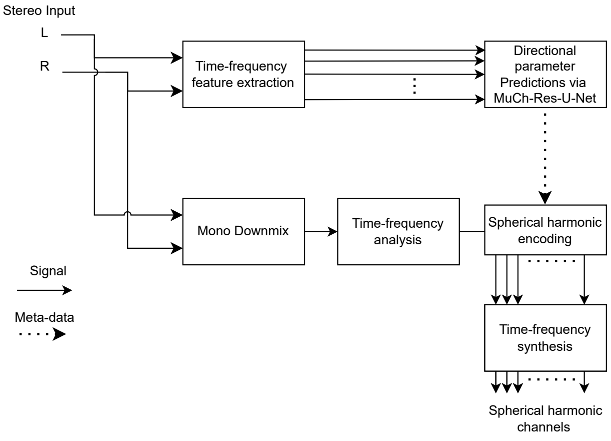

Stereo Input
L
R
Time-frequency feature extraction
Directional parameter Predictions via MuCh-Res-U-Net
Mono Downmix
Time-frequency analysis
Spherical harmonic encoding
Time-frequency synthesis
Spherical harmonic channels
Signal
Meta-data

Figure 2: *Block diagram of proposed stereo to B-format upmixer utilising directional parameters predicted by MuCh-Res-U-Net*

The spatial parameters predicted by the proposed network can be applied to a number of different upmixing scenarios. As an illustrative example, a pipeline (shown in Figure 2) is presented that utilises the network within a stereo upmixing algorithm. The example algorithm takes a stereo signal, captured with a spaced stereo microphone pair, and upmixes the signals to first order spherical harmonic components, which will be referred to by their B-format channel labeling.

First, a mono signal must be derived to represent the W channel, which can be approximated by an omnidirectional pressure signal. Due to the low inter-channel amplitude differences, the mono signal can be approximated using one of the two stereo signals The directional spatial parameters predicted by the model are then used to extract and weight the frequency components according to the target spherical harmonic coefficients [48]:

$$\beta^{\sigma}_{mi}(m,\omega_k) = W(m,\omega_k)Y^{\sigma}_{mi}(\hat{\theta}(m,\omega_k),\hat{\phi}(m,\omega_k)) \qquad (16)$$

Where:

- $\beta^{\sigma}_{mi}(m,\omega_k)$ is the time-frequency representation of the Ambisonic channel representing the spherical harmonic Y^{σ}_{mi},
- $W(m,\omega_k)$ is the time-frequency representation of the W channel from which the frequency components are being extracted and remapped. This approach is similar to that proposed in [49] where DirAC for telecommunications only transmits the metadata and W channel, discarding the other B-format channels after DirAC analysis.
- $\hat{\theta}(m,\omega_k)$ and $\hat{\phi}(m,\omega_k)$ are the predicted time-frequency directional parameters for azimuth and elevation respectively.

Finally, the resulting time-frequency Ambisonic channels can then be returned into the time-domain using the inverse STFT.

5. EVALUATION OF UPMIX PIPELINE

The IRs used to synthesise the training data were also used to spatialise a 3s pink noise burst, followed by 0.5s of silence, at all sampled locations on the horizontal and all elevation locations directly frontal to the receiver, which due to the Lebedev sampling scheme were located at azimuth positions 0°, 18°, or 45°. The directional performance of the upmix algorithm is evaluated based on the spherical distance, as defined in [50], between the DOA estimations (DOA-Est) for the upmixed B-format signals and the ground-truth B-format signal and will be referred to as the Total DOA error. It can be calculated as follows:

$$\begin{aligned}\Delta DOA^{3D} &= \arccos(\sin(\hat{\phi})\sin(\phi) \\ &\quad + \cos(\hat{\phi})\cos(\phi)\cos(|\theta-\hat{\theta}|)) \qquad (17)\end{aligned}$$

where ΔDOA^{3D} is the Total DOA error as spherical distance in degrees $^{\circ}$ and $\hat{\theta},\hat{\phi}$ are the DOA-Est from the upmixed B-format signals and θ,ϕ are the DOA-Est for the ground-truth B-format signals.

When referring to the DOA error for a single direction, either θ or ϕ, the 2D angular distance used, which can be defined as:

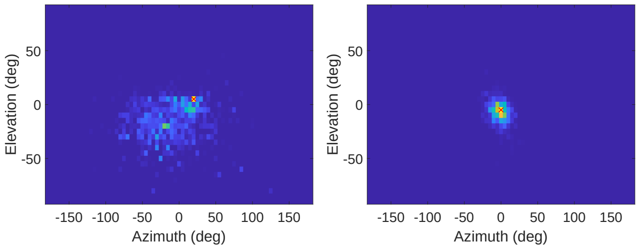
$$\Delta DOA^{2D\theta} = |\theta-\hat{\theta}| \qquad (18)$$

where $\Delta DOA^{2D\theta}$ is the error in the azimuthal direction and where the error in the elevation direction, $\Delta DOA^{2D\phi}$, is calculated by:

$$\Delta DOA^{2D\phi} = |\phi-\hat{\phi}| \qquad (19)$$

The DOA-Est are derived from the unsmoothed intensity vector, using the MATLAB library presented in [51]. The acoustic intensity measurements are sampled across time using a window length of 100 samples with an overlap of 50% and are used to compute histograms of their estimated DOAs, weighted by the magnitude of the vectors. DOA-Est are made on a vector of spherical grid points with a resolution of 5°. The grid locations associated with the greatest number of DOA estimates are assumed to represent the directions of the dominant sound sources and are determined based on Von-Mises peak-finding, presented in [52], which facilitates DOA estimates for a specified number of sources over the length of the given signal.

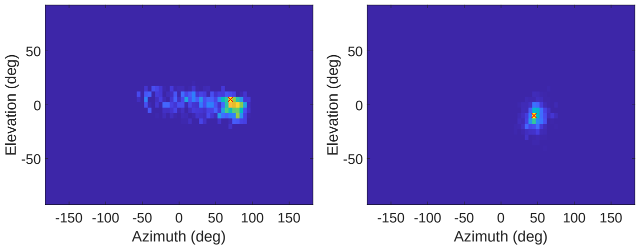
Figure 3 shows the DOA-Est histograms for a pink noise burst spatialised to $\theta = \phi = 0°$ for both the upmixed and ground truth B-format signals. As well as a Total DOA error of 22.34°, there is also evidence of greater fluctuations and variability in the DOAs estimated for the upmixed B-format signal. This infers that some spatial instability exists between time frames within the predicted directional parameters, where the predicted values cause DOA estimates to fluctuate between time-frames to a greater extent than is present in the ground truth data.



(a) *DOA estimates for upmixed B-format signal*

(b) *DOA estimates for ground-truth B-format signals*

Figure 3: *Directional grid of DOA estimates resulting from the time-sampled intensity vectors of the (a) upmixed B-format signals and (b) ground truth B-format signals for a source spatialised to* $\theta = \phi = 0°$.



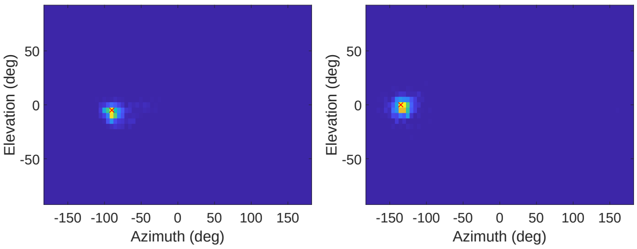
(a) *DOA estimates for upmixed B-format signal*

(b) *DOA estimates for ground-truth B-format signals*

Figure 4: *Directional grid of DOA estimates resulting from the time-sampled intensity vectors of the (a) upmixed B-format signals and (b) ground truth B-format signals for a source spatialised to* $\theta = 45°,\ \phi = 0°$.

Figure 4 shows the results for a pink noise burst spatialised at $\theta = 45°$, $\phi = 0°$, which resulted in a DOA error of 29.07°. There is also similar evidence of spatial instability with respect to the spatialisation derived from the predicted parameters. However, in this instance, the instability seems to be much more localised to the horizontal plane within ±20° elevation with the DOA-Est having a higher concentration around the dominant peak, which suggests a more stable spatial image.

For a source at $\theta = -135^\circ$, shown in Figure 5, the DOA error is 45° and the predicted parameters have been unable to produce an upmix where the source is positioned to the rear of the receiver but instead positioned it at the extent capable of traditional stereo directional estimates, at $\theta = -90^\circ$. The source appears, however, to be spatially stable, as evidenced by the high concentration of DOA-Est within a smaller number of grid locations. Interestingly, although being symmetric about the median plane, results for a pink noise burst located at $\theta = +135^\circ$ showed greater instability with DOA estimations spread through across a range of azimuth values from -90° to $+90^\circ$. For these positions, the predicted spatial parameters have failed to result in any DOA-Est to the rear of the receiver, which means the network was unable, in this instance, to predict directional parameters that result in the sources being remapped to the rear by the upmix process.



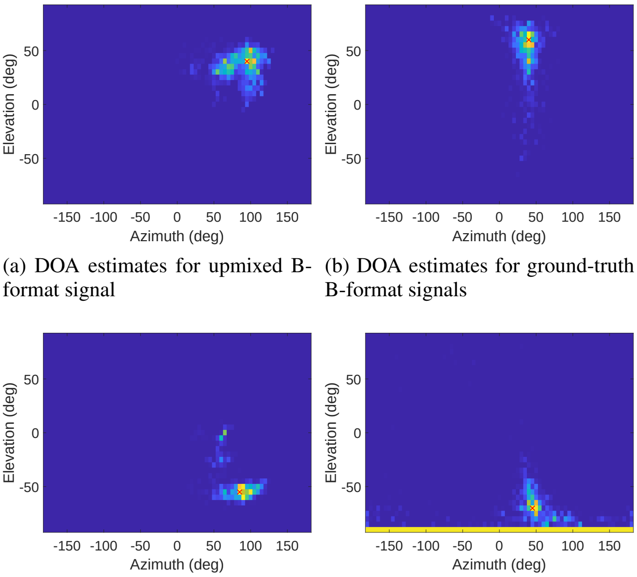
(a) *DOA estimates for upmixed B-format signal* (b) *DOA estimates for ground-truth B-format signals*

Figure 5: *Directional grid of DOA estimates resulting from the time-sampled intensity vectors of (a) the upmixed B-format signals and (b) the ground truth B-format signals for a source spatialised to* $\theta = -135^\circ,\ \phi = 0^\circ$.

Figures 6 shows the DOA-Est for sources located at $\theta = 45^\circ$, $\phi = \pm 65^\circ$. The results show that the predicted parameters are able to facilitate the upmix algorithm in positioning sources at both positive and negative elevations, although their position is underestimated in both the above and below cases. Whilst source positions of $\theta = 45^\circ, \phi = \pm 65$ resulted in Total DOA errors of 39.07° and 23.07 for $\phi = 65^\circ$ and $\phi = -65^\circ$, respectively, it should also be highlighted that a large contributor to the error values for the elevated source positions are due to larger errors in the azimuthal direction, with the results for elevation direction in isolation being within 25° of the DOA-Est resulting from the ground truth.

6. DISCUSSION

From these preliminary results, it appears that whilst the model has begun to learn a mapping function for lateral position, it has been unable to approximate a mapping function for front/rear source mapping. Two possible reasons could be hypothesised as to why this is. Firstly, is that the input features do not contain the required information to adequately differentiate between front and rear source positions, and different, or additional, input features are required. Secondly, is that a more suitable training strategy is required for optimisation of the loss function, which would both investigate whether the model being better optimised results in more accurate estimations of frontal azimuth positions and whether better optimisation would result in the model learning a more accurate mapping function to differentiate between front and rear source positions. The results also suggest that the model has begun to learn an approximate mapping function for source elevation, which results in upmixed sources being correctly positioned at either positive or negative elevation values.



(a) DOA estimates for upmixed B-format signal (b) DOA estimates for ground-truth B-format signals

(c) DOA estimates for upmixed B-format signal (d) DOA estimates for ground-truth B-format signals

Figure 6: Directional grid of DOA estimates resulting from the time-sampled intensity vectors of (a) and (c) the upmixed B-format signals and (b) and (d) the ground truth B-format signals for a source spatialised to (a), (b) $\theta = 45^\circ,\ \phi = 65^\circ$, and (c), (d) $\theta = 45^\circ,\ \phi = -65^\circ$.

7. FUTURE WORK

With an initial model developed and preliminary evaluation undertaken, future work should address more extensive evaluation of the algorithm including comparisons with existing upmixing methods, subjective listening tests, and testing using audio which has been spatialised with unseen IRs. Furthermore, model optimisation utilising different/multiple stereo configurations contained within the wider dataset collected by the authors should also be investigated alongside. Finally, perceptually weighted loss functions based on the measured diffuseness of the time-frequency components may help to optimise the prediction of directional parameters.

8. CONCLUSIONS

This paper detailed the development, investigation, and evaluation of a deep learning approach to predicting time-frequency spatial parameters for use within stereo upmixing using input feature vectors extracted from stereo signals. Relevant background was presented with respect to the current approaches to stereo upmixing that exist within the literature. The methodology for spatial parameter prediction was then detailed including the neural network architecture, dataset collection, input feature extraction, and model optimisation. Following this, an example upmix pipeline was presented that utilised the predicted time-frequency spatial parameters to facilitate a stereo to B-format upmix. Evaluation of the upmixed signals showed that whilst the current model could not predict directional parameters that resulted in the spatial remapping of time-frequency components such that objects were evaluated as being placed to the rear of the spatial scene, the predicted parameters were able to map to both positive and negative elevation values. These results provide evidence that there exists information within stereo signals that can be used to derive height information whilst further work is required to optimise a model to improve its performance when applied to front/rear remapping of time-frequency components.

9. ACKNOWLEDGMENTS

This research was funded by an EPSRC iCASE PhD Studentship in collaboration with BBC Research and Development.

10. REFERENCES

[1] Carlos Avendano and Jean Marc Jot, "A frequency-domain approach to multichannel upmix," *Journal of the Audio Engineering Society*, vol. 52, no. 7-8, pp. 740–749, 2004.

[2] Jeonghwan Choi and Joon-Hyuk 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, jan 2021.

[3] Sebastian Kraft and Udo Zölzer, "Low-complexity stereo signal decomposition and source separation for application in stereo to 3D upmixing," in *140th Audio Engineering Society International Convention 2016*, 2016.

[4] C. Avendano and J.M Jot, "Frequency domain techniques for stereo to multichannel upmix," in *22nd International Audio Engineering Society Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

[5] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003.

[6] M.M Goodwin and J.M Jot, "Primary-ambient signal decomposition and vector-based localisation for spatial audio coding and enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[7] Su Yeon Park, Chan Jun Chun, and Hong Kook Kim, "Subband-based upmixing of stereo to 5.1-channel audio signals using deep neural networks," *2016 International Conference on Information and Communication Technology Convergence, ICTC 2016*, pp. 377–380, 2016.

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

[9] Mark S. Vinton, Mark F. Davis, and Charles Q. Robinson, "Signal models and upmixing techniques for generating multichannel audio," *Proceedings of the AES International Conference*, pp. 1–12, 2011.

[10] Mark Vinton, David McGrath, Charles Robinson, and Phillip Brown, "Next generation surround decoding and upmixing for consumer and professional applications," *Proceedings of the AES International Conference*, vol. 2015-Janua, pp. 1–9, 2015.

[11] C. Faller, "Multiple-loudspeaker playback of stereo signals," *Journal of the Audio Engineering Society*, vol. 11, no. 54, pp. 1051–1064, 2006.

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

[13] Karim M. Ibrahim and Mahmoud Allam, "Primary-ambient source separation for upmixing to surround sound systems," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pp. 431–435, 2018.

[14] Jianjun He, *Spatial audio reproduction using primary ambient extraction*, Ph.D. thesis, Nanyang Technological University, 2016.

[15] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "Mpeg-h 3d audio - the new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, pp. 770–779, 2015.

[16] John Usher, "Design criteria for high quality upmixers," in *28th AES International Conference on The Future of Audio Technology–Surround and Beyond*, 2006, pp. 1–13.

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

[18] S. Jeon, Y.C. Park, S.P Lee, and D.H Youn, "Robust Representation of Spatial Sound in Stereo-to-Multichannel Upmix," in *128th Audio Engineering Society Conference*, 2010.

[19] Venkatesh S. Kadandale, Juan F. Montesinos, Gloria Haro, and Emilia Gomez, "Multi-channel u-net for music source separation," *IEEE 22nd International Workshop on Multimedia Signal Processing, MMSP 2020*, 2020.

[20] Linlin Ou and Yuanping Chen, "Acoustic bandwidth extension by audio deep residual u-net," in *2022 5th International Conference on Information Communication and Signal Processing, ICICSP 2022*, 2022, pp. 549–554.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, 5 2015, pp. 234–241.

[22] Woosung Choi, Minseok Kim, Jaehwa Chung, Daewon Lee, and Soonyoung Jung, "Investigating u-nets with various intermediate blocks for spectrogram-based singing voice separation," in *21st International Society for Music Information Retrieval*, 2020.

[23] Daniel Stoller, Sebastian Ewert Spotify, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, 2018.

[24] Da Yi Wu, Yen Hao Chen, and Hung Yi Lee, "Vqvc+: One-shot voice conversion by vector quantization and u-net architecture," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 4691–4695.

[25] Rui Li, Dong Pu, Minnie Huang, and Bill Huang, "Unet-tts: Improving unseen speaker and style transfer in one-shot voice cloning," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8327–8331.

[26] Eloi Moliner and Vesa Välimäki, "A two-stage u-net for high-fidelity denoising of historical recordings," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 841–845.

[27] Pedro Morgado, Yi Li, and Nuno Vasconcelos, "Learning representations from audio-visual spatial alignment," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[28] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin, "Visually informed binaural audio generation without binaural audios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 15485–15494.

[29] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[31] C.H. Knapp and G. C Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, 1976.

[32] Jeonghwan Choi and Joon Hyuk Chang, "Convolutional neural network-based direction-of-arrival estimation using stereo microphones for drone," *2020 International Conference on Electronics, Information, and Communication, ICEIC 2020*, 2020.

[33] Nicholas Jillings, Alice Clifford, and Joshua D Reiss, "Performance optimization of gcc-phat for delay and polarity correction under real world conditions," in *134th Audio Engineering Society Convention*, 5 2013.

[34] Ville Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 46, no. 6, pp. 458–466, 2007.

[35] Archontis Politis, Tapani Pihlajamäki, and Ville Pulkki, "Parametric spatial audio effects," in *15th International Conference on Digital Audio Effects (DAFx-12)*, 9 2012.

[36] Marc C Green, *Environmental Sound Monitoring Using Machine Listening and Spatial Audio*, Ph.D. thesis, University of York, 2021.

[37] Ville Pulkki and Christof Faller, "Directional audio coding: Filterbank and stft-based design," in *120th Convention of the Audio Engineering Society*, 5 2006.

[38] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[39] Thomas Wolf, "Training neural nets on larger batches: Practical tips for 1-gpu multi-gpu & distributed setups," Available at https://medium.com/huggingface/training-larger-batches-practical-tips-on-1-gpu-multi-gpu-distributed-setups-ec88c3e51255, 2018.

[40] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *5th International Conference on Learning Representations (ICLR)*, 4 2017.

[41] Prem Seetharaman, Gordon Wichern, Bryan Pardo, and Jonathan Le Roux, "Autoclip : Adaptive gradient clipping for source separation networks," in *2020 IEEE International Workshop on Machine Learning for Signal Processing*, 2020.

[42] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie, "Why gradient clipping accelerates training: A theoretical justification for adaptivity," in *8th International Conference on Learning Representations (ICLR)*, 5 2020.

[43] Gwantae Kim, David K. Han, and Hanseok Ko, "Specmix : a mixed sample data augmentation method for training with time-frequency domain features," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2021, pp. 6–10.

[44] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech 2015*, 2015, pp. 3586–3589.

[45] Daniel Turner and Damian Murphy, "Dataset of stereo and multi-channel IRs for a 50-point Lebedev quadrature.," May 2023, Available at https://doi.org/10.5281/zenodo.7990195.

[46] Archontis Politis, Sharath Adavanne, and Tuomas Virtanen, "A Dataset of Reverberant Spatial Sound Scenes with Moving Sources for Sound Event Localization and Detection," 2020, Available at http://arxiv.org/abs/2006.01919.

[47] mH Acoustics, "em32 Eigenmike microphone array release notes (v17.0)," NA, Available at https://mhacoustics.com/sites/default/files/ReleaseNotes.pdf.

[48] Cal. Armstrong and Gavin Kearney, "Ambisonics understood," in *3D Audio*, Justin. Paterson and Hynukook Lee, Eds., pp. 99–129. Routledge, New York, NY, 2021.

[49] Jukka Ahonen, Ville Pulkki, and Tapio Lokki, "Teleconference application and b-format microphone array for directional audio coding," *Proceedings of the AES International Conference*, pp. 1–10, 2007.

[50] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen, "Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2018-July, 2018.

[51] Archontis Politis, "Spherical array processing," 2021, Available at https://github.com/polarch/Spherical-Array-Processing.

[52] Sakari Tervo, "Direction estimation based on sound intensity vectors," in *2009 17th European Signal Processing Conference*, 2009, pp. 700–704.