Audio Engineering Society

Convention Paper

Presented at the AES 2026 Convention
2026 July 29–31, Detroit, MI, USA

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Implementing Two Upmixing Concepts With One Conditional GAN

Christian Uhle[1,2], Philipp Weyers[1], and Johannes Zimmermann[1]

[1]*Fraunhofer IIS, Erlangen, Germany*
[2]*International Audio Laboratories Erlangen, Erlangen, Germany - A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS*

Correspondence should be addressed to Christian Uhle (christian.uhle@iis.fraunhofer.de)

ABSTRACT

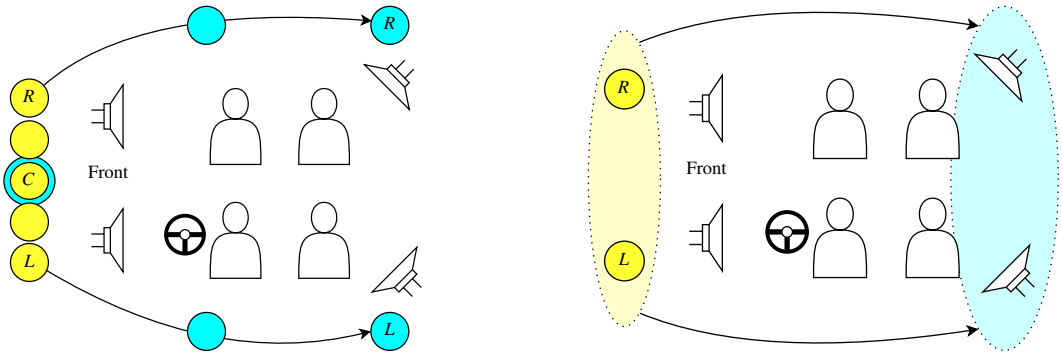
Signal processing using artificial intelligence (AI) has gained increasing interest because it outperforms existing solutions in many fields. A significant challenge for deep neural networks lies in meeting strict requirements regarding latency, computational load and memory, which is vital in automotive audio. This paper presents CUpGAN (Conditional Upmix GAN), a computationally efficient method for extracting upmix signals with low latency, leveraging signal separation for two upmixing concepts using one conditional generative adversarial network (CGAN). One upmix approach utilize spatial positions of direct sources within the stereo image, allowing for the distribution of sources around the listener. The second approach separates direct and diffuse signals to create an ambience signal for rear surround loudspeakers.

By employing phase-aware loss functions, integrating residual connections in the generator, and training with coherent input and target signals, we achieve high sound quality in the generated signals. The residual signals, i.e., difference between input and generated signals can be used as complementary signals for both upmixing concepts. The proposed method reduces memory as most parameters are shared between both applications, allowing seamless switching between upmixing approaches without the need for parameter loading; instead, a small number of parameters are computed by control network and used to modulate the latent variables. The GAN generator is trained on synthetically generated data, enabling control of separation characteristics that surpass traditional methods. We present an evaluation using listening tests and computational metrics, demonstrating the advantages of our approach compared to classical signal processing methods.

1 Introduction

Reproduction of audio signals in surround or immersive multi-channel formats finds wide adoption in cinema, broadcast, home entertainment, and automotive audio. Most musical recordings are available in two-channel stereo, and their reproduction in setups with additional loudspeakers requires the creation of additional signals, commonly referred to as upmixing [1]. Figure 1 illustrates two upmixing concepts used in automotive audio: panorama upmix (PU) and ambience upmix (AU).

PU distributes sound sources around the listener, stretches the perceived source positions, and thereby expands the stereo panorama. AU reproduces an ambient signal using rear or lateral loudspeakers [2], while



(a) Panorama Upmix: Direct sources (circles) are stretched around listeners.

(b) Ambience Upmix: Ambience (ellipse) is reproduced by rear loudspeakers.

Fig. 1: Schematic illustration of upmixing concepts with input signal (yellow) and after upmixing (blue), where C, L, and R denote positions left, center, and right, respectivly.

front loudspeakers reproduce the unprocessed input or a signal with reduced ambience. These two concepts can be combined such that sound sources are re-panned while ambience is reproduced by dedicated loudspeakers.

Both concepts utilize the separation of upmix signals based on spatial cues and require low latency and memory requirements, and easy adaptation when users switch between different modes of sound reproduction. This paper proposes a DNN-based method for separating upmix signals that meets these requirements. Its main contributions is the separation based on spatial cues of the localization of direct sounds and the diffuseness of ambient sounds, utilizing a CGAN. The generator network enables seamless switching between both separations. The method operates with memory requirements and computational loads comparable to those of a model designed for computing a single upmix signal. Additional upmix signals can be generated by subtracting the produced upmix signals from the input, resulting in a cost-efficient signal generation for four distinct loudspeaker signals. The training data are created by mixing anechoic audio signals with artificial reverberation, allowing for flexible adjustment of separation characteristics. Our approach is evaluated through listening tests and computational evaluation metrics and achieves higher sound quality compared to the methods we evaluated.

The paper is organized as follows: Sec. 2 describes the problem, Sec. 3 presents prior work, Sec. 4 explains the proposed implementation, in Sec. 5 the method is evaluated, and conclusions are provided in Sec. 6.

2 Problem description

The input signal $x[n]$ has two channels and is a mixture of direct signals $d[n]$ (correlated in both channels) and diffuse signals $a[n]$, such as reverberation and other ambient sounds.

We generate a four-channel signal for reproduction with two frontal and two lateral loudspeakers, resembling an entry level automotive set-up, based on one of the two described upmixing concepts. The first two-channel signal is computed using a DNN, while the second two-channel signal is obtained as difference signal between input and model output. The presented two-to-four-channel upmix can be extended to more channels with modifications of model architecture and training data, but at this stage of research we restrict our work to four output channels.

For direct-ambient upmix, the output signal $y_0[n]$ is the ambient signal $\hat{a}[n]$, and the direct signal is computed as $\hat{d}[n] = x[n] - \hat{a}[n]$. For the panorama upmix, the output signal $y_1[n]$ is the center signal $\hat{c}[n]$.

This center signal differs from commonly used single-channel signals designed for feeding the center loudspeaker placed midway in front, as described in [3], as follows: we compute a center signal with two channels that can be downmixed to a single loudspeaker channel or fed to left and right front loudspeaker.

Two-channels center signals have the additional advantage over single channel signals when computing the difference signal $\hat{o}[n] = x[n] - \hat{c}[n]$, which contains sources panned off-center.

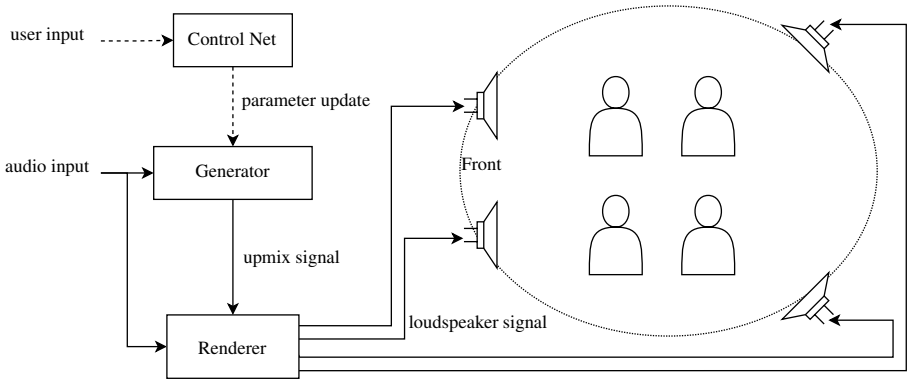


Fig. 2: Upmix processing and rendering for a simplified loudspeaker set-up. The generator processes audio input to obtain upmix signals. It is switchable between two upmixing concepts using a small parameter vector computed by a control network given a binary user input. The renderer module applies post-processing to input and upmix signal to compute loudspeaker signals.

3 Prior Work

Upmixing can be categorized into two types: (i) blind upmixing, which is addressed in this paper, and (ii) decoding that utilizes side information derived from multi-channel content during encoding, which is not discussed here.

3.1 Extraction of direct signals for upmixing

Direct signals are localized in the stereo image, where level and time differences between the loudspeaker signals determine the perceived direction of the phantom source [4]. Signals panned to a specific positional range can be extracted and reproduced by dedicated loudspeakers. The algorithm proposed in [2] employs a least-mean-square adaptive filtering approach applied to the time-domain signal. Processing in the time-frequency domain is particularly effective, leveraging source sparsity [5], which assumes that only one source (or multiple sources with similar spatial cues) contributes significantly to a given time-frequency bin compared to other sources. In [6], the authors applied spectral weighting with real-valued weights based on measures of similarity or correlation between the input channel signals. A multichannel Wiener filter (MWF) using a linear combination of input channels with complex-valued weights has been proposed in [7].

Of particular interest are center loudspeakers placed midway in front to enhance stereo image reproduction [8, 9], especially when loudspeakers are not symmetrically positioned relative to the listener. Center extraction based on the sum and difference of spectral coefficients is proposed in [10]. Spectral weighting using real-valued weights computed from functions of the ratios of downmix and input signals is employed in [11, 12], allowing for separate control of attenuation for direct and diffuse signal components. A GAN-based approach has been proposed in [13]. Center extraction is also utilized for dialog enhancement in [14, 15, 16].

3.2 Direct-ambient decomposition

Ambient signals are diffuse and non-coherent between channels, making them well-suited for generating surround signals. They comprise reverberation, environmental sounds and ensembles of sound sources captured with multiple microphones.
Various methods for ambience extraction and direct-ambient decomposition (DAD) have been proposed, utilizing spectral weighting based on inter-channel coherence (ICC) [17, 18] and estimates of the ambience power [19]. Predicting the direct signal component using adaptive filtering is described in [20], where ambience is computed as the difference between input and direct signal estimate. Other approaches employ MWF [7] and principal component analysis (PCA) [19, 2, 21]. An extensive comparison of MWF and PCA approaches, along with a generic linear estimation framework, is presented in [22]. Distortions of

spatial cues introduced by the MWF have been analyzed in [23], where a parametric MWF is proposed to keep these distortions below psychoacoustic thresholds.
For single-channel input signals, methods based on non-negative matrix factorization [24] and shallow neural networks [25] have been proposed. For ambience extraction of multichannel signals, the authors of [26] used a two-channel downmix and Wiener filtering to estimate the power spectral density of the ambient signal. In [27], pairwise correlations were proposed.

3.3 Signal extraction for upmixing with AI

Generating a center signal and surround signals based on style transfer has been proposed in [28]. In [29], DAD is implemented using a feed-forward network with three hidden layers that processes spectral coefficients for two-channel inputs to predict a real-valued spectral mask. In contrast to our work, training data consist of either purely direct or purely diffuse signals, which may limit performance. A method to predict surround and center signals is presented in [30], where training data are obtained from multichannel signals. The prediction of direction of arrival and a diffuseness index for each time-frequency bin, which can be utilized for upmixing stereo recordings, is detailed in [31]. The model was trained with synthesized Ambisonic sound scenes. This approach has drawbacks, as multichannel formats may contain phantom center sources in the front channels and arbitrary mixtures of direct and diffuse signals in the surround channels.
In contrast to previous works, we propose a generative adversarial network (GAN) to improve sound quality, along with a conditional method to reduce memory usage, and generate training data with appropriate target signals.

3.4 Conditioned source separation

Conditioning in machine learning refers to modulating the computation of a model by auxiliary information and is used to combine input information with context. An effective way is to fuse input and context in the intermediary network layers using conditional normalization [32].
Feature-wise linear modulation (FiLM) [33] is a generalization of conditional normalization. It was originally proposed for visual reasoning and applies an affine transformation to the c-th latent variable $x_{i,c}$ at layer index i,

$$FiLM(x_{i,c}|\gamma_{i,c}(z),\beta_{i,c}(z)) = \gamma_{i,c}(z)x_{i,c} + \beta_{i,c}(z), \quad (1)$$

where $\gamma_{i,c}$ and $\beta_{i,c}$ are computed using a conditioning network given the integer conditioning variable z. FiLM has been employed for conditioned source separation [34], where it was suggested to apply same parameters γ_i, β_i to all features within each layer i.
The authors of [35] extend this approach with gated point-wise convolutional modulations; however, we found the simple approach [34] to be effective for our purpose and preferable for its simplicity.

4 Method

4.1 Generator

Figure 3 shows a block diagram of the generator used in the CGAN. It processes complex-valued short-time Fourier transform (STFT) coefficients of the two-channel input $\mathbf{x}(m,k) = [X_0(m,k), X_1(m,k)]^T$, computed using the discrete Fourier transform (DFT) with length 1024, hop size 512, and window size 1024 from signals with a sampling rate of 44100 Hz.
STFT coefficients X (omitting time m and frequency indices k for brevity) of each audio channel are represented as triplets $[\log(M + 10^{-7}), \mathrm{real}(X/M), \mathrm{imag}(X/M)]$, with magnitudes $M = |X|$, resulting in six channel inputs per time-frequency bin for two-channel audio signals. This representation enables to compress spectral magnitudes to reduce the large dynamic range of audio signals and thereby to improve numerical optimization during training.
It is first expanded to 21 dimensions using an inverted bottleneck (IBN) with grouped convolution for each audio channel, utilizing kernels of shape (3×3).
The latent representation is then processed by four layers of conditional ConvNext blocks [36], as shown in Figure 4. ConvNext blocks were initially applied for image classification as a computationally efficient alternative to Vision Transformers (ViTs) [37] and were later used for generating audio signals from Mel spectrograms in [38] using 1D convolutions. In contrast, we employ 2D convolutions with kernels operating along time and frequency, with kernel size 3×3.
ConvNext blocks are extended with FiLM layers

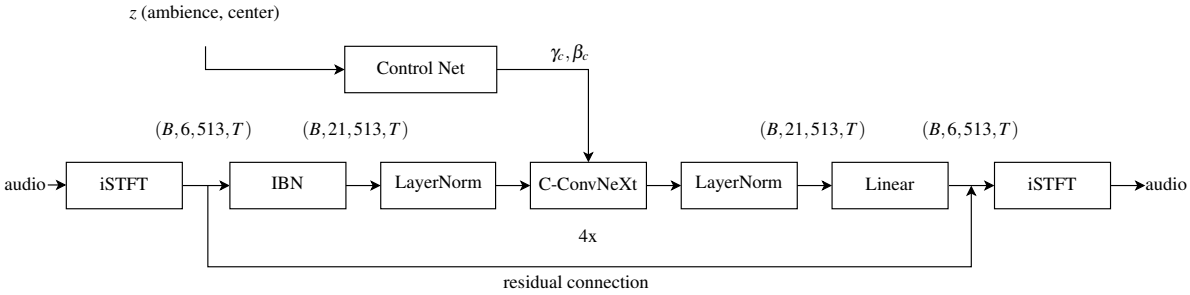


Fig. 3: Block diagram of generator. Conditioned ConvNeXt is repeated 4 times. Tensor shapes with batch size B and number of frames T.

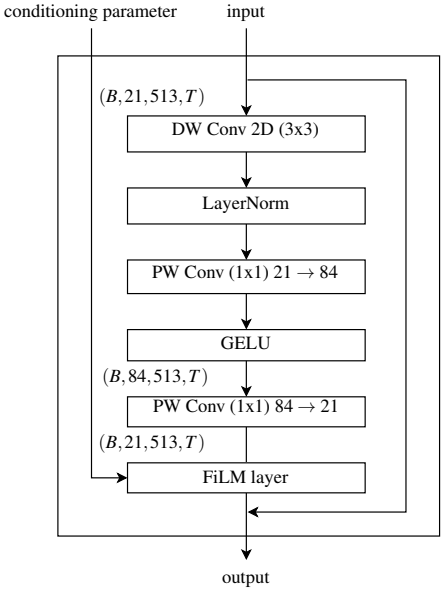


Fig. 4: ConvNeXt Blocks in generator, with pointwise (PW) and depthwise (DW) convolutions. Tensor shapes with batch size B and number of frames T.

to apply (1). Conditioning parameters γ_c, β_c are computed with a control net with four linear layers with $[2,4,8,16]$ neurons given the binary index z for switching between ambience extraction and center extraction.

Layer normalization [39] is applied before the first and after the last ConvNeXt block to reduce sensitivity to weight initialization and speed up training. The linear layer implements the bottleneck to obtain tensor of same shape as X with six channels.

4.2 Discriminator

The generator is trained alongside three instances of complex-valued multi-resolution spectrogram discriminators (MRSDs) [40] with window lengths of $[2048, 1024, 512]$ samples. This approach builds upon the method introduced in [41], which employs discriminators that operate on time-frequency representations at multiple resolutions, focusing on the real and imaginary components of complex-valued STFT coefficients rather than just their magnitudes. The discriminator uses 2D convolutional layers that independently process STFT coefficients in five frequency bands.

We also explored combinations of MRSD with instances of multi-scale waveform discriminator (MSD) [42] and multi-period waveform discriminator (MPD) [43], but did not notice significant improvements.

4.3 Training loss

The generator is trained using multiple loss functions as in [13], which include reconstruction losses, as well as adversarial losses derived from the discriminator instances.

Reconstruction loss is the weighted sum of multi-resolution STFT loss [44], multi-resolution loss of complex-valued STFT coefficients, and mean squared error (MSE) loss of time signal that has been high-pass filtered at 50 Hz cut-off frequency with low Q factor.

The time signal loss provides a straightforward and effective way to adjust the spectral balance of the generated signal. Complex-valued multi-resolution STFT loss is calculated as sum of MSE of real part and MSE of imaginary part.

Multi-resolution STFT loss utilizes DFT lengths of $[2048, 1024, 512]$, hop size $[240, 120, 50]$, and window

size $[600, 1200, 240]$, while complex-valued multi-resolution STFT loss operate at half of these lengths to enhance temporal resolution.

Adversarial loss is the weighted sum of least-square adversarial loss [45] and feature matching loss [46, 42], computed as the L1 distance between discriminator feature maps obtained for target signals and generator output signals.

Generator (including the control net) and discriminator are trained in an end-to-end manner.

4.4 Training data

Training data are synthetic mixtures of single-channel signals panned with real-valued scalars and processed with convolution reverb. This procedure allows to compute appropriate target signals for both separation concepts, because the panning of direct sources is known and diffuse and direct signals are separately available during data creation.

Input signals were generated by mixing four source signals from MUSDB18 (vocals, drums, bass, other accompaniment) [47] and one source signal containing speech from various audiobooks.

First, we add source signals $s_i[n+o]$ to create the direct signals $\mathbf{d}[n] = \sum_{i=0}^{4} [g_{i,0} s_i[n+o], g_{i,1} s_i[n+o]]$ with random gains $[g_{i,0}, g_{i,1}]$, uniformly distributed in the range $[-6, 6]$, with random time offset o.

The ambience signals $\mathbf{a}[n]$ are generated by applying the same mixing as for direct signals $\mathbf{d}[n]$ and additionally convolving each source signal $s_i[n]$ with different two-channel impulse response (IR) of reverberation $\mathbf{r}[n]$ taken from the training split of ReverbFX [48].

The input signals are obtained by adding direct and ambience signal $\mathbf{x}[n] = \mathbf{d}[n] + \mathbf{a}[n]$. The target signals for ambience extraction are diffuse signal components $\mathbf{a}[n]$. Target signals for center extraction are computed by using the same mixing as for direct signals $\mathbf{d}[n]$ and additionally scaling source signals $s_i[n]$ with attenuation $h_i = f(\mathrm{ICLD})$, yielding $\mathbf{c}[n] = \sum_{i=0}^{4} h_i [g_{i,0} s_i[n]\, g_{i,1} s_i[n]]$. They do not contain reverberation, and the scaling for panned signals is $h = 0.5 \cdot 10 \log_{10}(-\left|10^{0.05 \cdot g_{i,0}} / 10^{0.05 \cdot g_{i,1}}\right|)^2$.

The resulting training data set consists of 2,400 pairs of input signals $\mathbf{x}[n]$ and target signal $\mathbf{y}[n]$ of 6 sec length each with two channels sampled at 44.1 kHz.

4.5 Optimizer settings for training

Models were trained for $400,000$ iterations using the Adam optimizer with an initial learning rate of 5×10^{-4}, which exponentially decays with a decay factor of 0.9999995, with gradient clipping at maximum norm of 5. Each training iteration utilizes batches of 16 signals of length 0.5 seconds each.

4.6 Computational load, latency, and memory requirements

Training time per model is approximately 32 hours on an A100 GPU. The generator model has $14,000$ parameters and requires 264×10^6 MACs per second of audio. The latency of the generator equals 23 ms, corresponding to the window size of the STFT. It does not introduce additional latency when integrated into processing pipelines using STFT, as all convolutional layers are implemented with causal padding.

5 Evaluation

5.1 Objective assessment

We assess the proposed model using SI-SDR [49] and the 2F measure [50], both of which are commonly employed in source separation evaluations. SI-SDR quantifies the distortion caused by interfering sources and artifacts, while the 2F measure assesses sound quality by predicting the mean score obtained in a MUSHRA listening test. These metrics have been used for evaluating center extraction in [13], but their application to ambience extraction is not common and, as results for SI-SDR indicate, not always useful.

The proposed condition is referred to as CUpGAN-C for center extraction and CUpGAN-A for ambience extraction. For comparison, we use two previous methods that do not utilize machine learning: the center extraction with generalized signal-to-downmix ratios (GSDR) [12] and the direct-ambient decomposition (DAD) from [23]. These conditions have been parametrized using one calibration item such that residuals of interfering signals are perceived with equal loudness when comparing the proposed and the previous method.

To investigate the effect of the conditioning mechanism, we train dedicated models for each task that have the same model complexity as CUpGAN without the control net, referred to as GAN-C and GAN-A.

	SI-SDR [dB] ↑	2F ↑
CUpGAN-C	9.6	45.1
GAN-C	10.0	44.2
GSDR [12]	7.1	33.8
CUpGAN-A	-11.6	28.3
GAN-A	-11.7	26.7
DAD [23]	-11.7	32.1

Table 1: Objective evaluation for proposed center extraction (CUpGAN-C) and ambience extraction (CUpGAN-A), unconditioned GAN-C and GAN-A, and GSDR and DAD. SI-SDR is not useful for evaluating ambience extraction.

Test items were generated with the same procedure as in the training phase, with modifications to create distinct data distributions between training and testing. We use the test split of MUSDB18 along with the test split of ReverbFX. Speech signals that were part of the training data were excluded from the evaluation; instead, we ensured that all four source signals were active. Table 1 shows the results for SI-SDR and 2F measure.
For center extraction, both GANs show better results compared to GSDR. The unconditioned and the conditioned models achieve similar scores.
For ambience extraction, we find that SI-SDR does not produce meaningful results. This is likely due to the fact that reference and prediction are uncorrelated, which results in large residual signals when computing SI-SDR. The previous method yields better 2F scores compared to the GAN.

5.2 Listening test

We evaluate center extraction and ambience extraction with CUpGAN in two A/B listening tests comparing to the previous methods GSDR and DAD, respectively. Input signals were eleven commercial recordings of five to 13 seconds length each of various gengres.
In both test, ten expert listeners were asked to compare the proposed method with the baseline using a comparison category rating scale [51] according to two aspects: the separation effect and overall sound quality of the signals. The tests were conducted using headphones, and listeners were free to choose the headphone model. Listeners could switch to unprocessed signals. We found it not helpful to apply loudness normalization, because signals with less separation and higher levels of residual signals will be attenuated more, thereby reducing the level of the desired signals.

Figures 5 and 6 show median and confidence interval of listener ratings per item and for all items. The proposed method is rated significantly better than the other condition in both tests. The improvement is larger for center extraction than for ambience extraction. Some listeners reported that it was not easy to compare the extracted ambience signals when sources in the mix have different attenuation of direct components between conditions.

6 Conclusion

We introduced CUpGAN for the extraction of upmix signals with low computational load and no additional latency, allowing for seamless switching between center and ambience extraction through a conditioning mechanism.
To our knowledge, this is the first implementation of center extraction and ambience extraction using conditional neural networks. A crucial aspect is the creation of training data, which involves pairs of input and reference signals.
The proposed approach results in difference signals that can be used to feed separate loudspeaker channels. The results of our evaluation indicate an improvement over methods based on classical signal processing in terms of listeners preference. Further work comprises the extension to more output channels with similar complexity and the improvement of the objective evaluation for ambience extraction.

References

[1] Rumsey, F., "Signal processing for 3-D audio," *J. Audio Eng. Soc*, 56(7/8), 2008.

[2] Irwan, R. and Aarts, R., "Two-to-Five Channel Sound Processing," *J. Audio Eng. Soc.*, 50, 2002.

[3] International Telecommunication Union, "Multichannel stereophonic sound system with and without accompanying picture." Recommendation ITU-R BS.775-2, 2006.

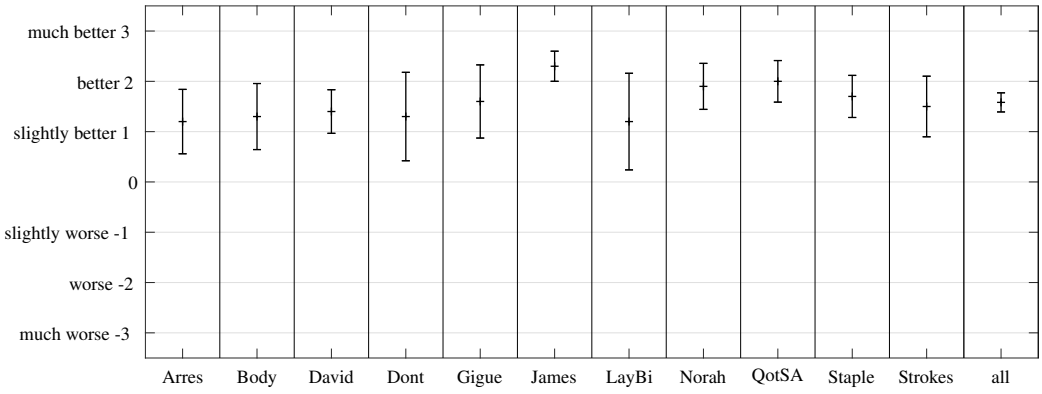


Fig. 5: Results of listening test comparing center extraction with CUpGAN to method GSDR [12]. Positive ratings indicate preference of CUpGAN.

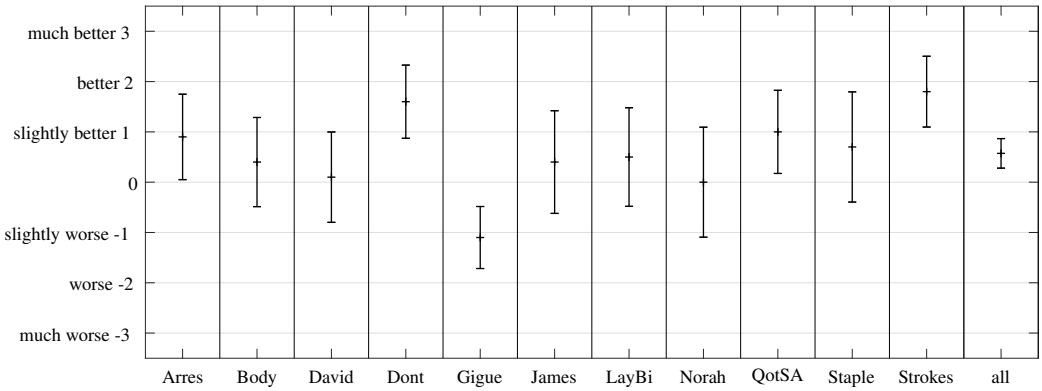


Fig. 6: Results of listening test comparing ambience extraction with CUpGAN to method DAD [23]. Positive ratings indicate preference of CUpGAN.

[4] Wittek, H., *Perceptual differences between wavefield synthesis and stereophony*, University of Surrey (United Kingdom), 2007.

[5] Jourjine, A., Rickard, S., and Yilmaz, O., "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures," in *Proc. Int. Conf. Acoust., Speech, and Signal Process., ICASSP*, 2000.

[6] Avendano, C. and Jot, J.-M., "A frequency-domain approach to multi-channel upmix," *J. Audio Eng. Soc.*, 52, 2004.

[7] Faller, C., "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, 54, 2006.

[8] Theile, G., "On the performance of two-channel and multi-channel stereophony," in *Proc. AES 88th Conv.*, 1990.

[9] Holman, T., "New factors in sound for cinema and television," *J. Audio Eng. Soc*, 39(7-8), 1991.

[10] Vickers, E., "Two-to-three Channel Upmix for Center Channel Derivation and Speech Enhancement," in *Proc. AES 127th Conv.*, 2009.

[11] Uhle, C., "Center Signal Scaling using Signal-to-Downmix Ratios," in *Proc. Int. Conf. Digital Audio Effects, DAFx*, 2013.

[12] Uhle, C. and Habets, E., "Subband Center Signal Scaling Using Power Ratios," in *Proc. AES 53rd Int. Conf. Semantic Audio*, 2014.

[13] Uhle, C., Klapp, J., Panter, P., and Walther, A., "Center Extraction GAN," in *Proc. AES 160th Conv.*, 2026.

[14] Härmä, A. and Park, M., "Extraction of Voice from the Center of the Stereo Image," in *Proc. AES 130th Conv.*, 2011.

[15] Craciun, A., Uhle, C., and Bäckström, T., "An Evaluation of Stereo Speech Enhancement Methods for Different Audio-Visual Scenarios," in *Proc. 23rd IEEE European Signal Process. Conf.*, 2015.

[16] Geiger, J., Grosche, P., and Parodi, Y., "Dialogue Enhancement of Stereo Sound," in *Proc. 23rd IEEE European Signal Process. Conf.*, 2015.

[17] Allen, J., Berkeley, D., and Blauert, J., "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, 62, 1977.

[18] Avendano, C. and Jot, J. M., "Ambience Extraction and Synthesis from Stereo Signals for Multi-channel Audio Upmix," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process., ICASSP*, 2002.

[19] Merimaa, J., Goodwin, M., and Jot, J.-M., "Correlation-Based Ambience Extraction from Stereo Recordings," in *Proc. AES 123rd Conv.*, 2007.

[20] Usher, J. and Benesty, J., "Enhancement of Spatial Sound Quality: A New Reverberation-Extraction Audio Upmixer," *IEEE Trans. Audio, Speech, and Language Process.*, 15, pp. 2141–2150, 2007.

[21] Goodwin, M. and Jot, J.-M., "Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process., ICASSP*, 2007.

[22] He, J., Tan, E.-L., and Gan, W.-S., "Linear Estimation Based Primary-Ambient Extraction for Stereo Audio Signals," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, 22(2), 2014.

[23] Uhle, C. and Habets, E., "Direct-Ambient Decomposition using parametric Wiener filtering wih spatial cue control," in *Proc. Int. Conf. on Acoust., Speech and Signal Process., ICASSP*, 2015.

[24] Uhle, C., Walther, A., Hellmuth, O., and Herre, J., "Ambience Separation from Mono Recordings using Non-negative Matrix Factorization," in *Proc. AES 30th Int. Conf. on Intelligent Audio Environments*, 2007.

[25] Uhle, C. and Paul, C., "A Supervised Learning Approach to Ambience Extraction from Mono Recordings for Blind Upmixing," in *Proc. Int. Conf. Digital Audio Effects, DAFx*, 2008.

[26] Walther, A. and Faller, C., "Direct-Ambient Decomposition and Upmix of Surround sound signals," in *Proc. IEEE Workshop on Applications of Signal Process. to Audio and Acoust., WASPAA*, 2011.

[27] Thompson, J., Smith, B., Warner, A., and Jot, J.-M., "Direct-Diffuse Decomposition of Multichannel Signals Using a System of Pairwise Correlations," in *Proc. AES 133rd Conv.*, 2012.

[28] Yang, H., Wager, S., Russell, S., Luo, M., Kim, M., and Kim, W., "Upmixing Via Style Transfer: A Variational Autoencoder for Disentangling Spatial Images And Musical Content," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process., ICASSP*, 2022.

[29] Ibrahim, K. M. and Allam, M., "Primary-Ambient Source Separation for Upmixing to Surround Sound Systems," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process., ICASSP*, 2018.

[30] Park, S. Y., Chun, C. J., and Kim, H. K., "Subband-based upmixing of stereo to 5.1-channel audio signals using deep neural networks," in *IEEE Int. Conf. on Information and Communication Technology Convergence, ICTC*, IEEE, 2016.

[31] Turner, D. and Murphy, D. T., "A deep learning approach to the prediction of time-frequency spatial parameters for use in stereo upmixing," in *Proc. 27th Int. Conf. on Digital Audio Effects, DAFx*, pp. 428–435, DAFx, 2024.

[32] Sheth, I., Rahman, A. A., Havaei, M., and Kahou, S. E., "Pitfalls of conditional batch normalization for contextual multi-modal learning," *arXiv preprint arXiv:2211.15071*, 2022.

[33] Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A., "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. on Artificial Intelligence*, volume 32, 2018.

[34] Meseguer-Brocal, G. and Peeters, G., "Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations," *arXiv preprint arXiv:1907.01277*, 2019.

[35] Choi, W., Kim, M., Chung, J., and Jung, S., "LaSAFT: Latent source attentive frequency transformation for conditioned source separation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process., ICASSP*, IEEE, 2021.

[36] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S., "A convnet for the 2020s," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, CVPR*, 2022.

[37] Dosovitskiy, A., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[38] Siuzdak, H., "Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis," in *Int. Conf. on Representation Learning*, 2024.

[39] Ba, J. L., Kiros, J. R., and Hinton, G. E., "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[40] Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., and Kumar, K., "High-fidelity audio compression with improved RVQGAN," *Advances in Neural Information Processing Systems*, 36, 2023.

[41] Jang, W., Lim, D., Yoon, J. E. S., Kim, B., and Kim, J., "UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," *arXiv preprint arXiv:2106.07889*, 2021.

[42] Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., De Brebisson, A., Bengio, Y., and Courville, A. C., "MelGAN: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, 32, 2019.

[43] Kong, J., Kim, J., and Bae, J., "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, 33, 2020.

[44] Yamamoto, R., Song, E., and Kim, J.-M., "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process., ICASSP*, 2020.

[45] Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S., "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. on Computer Vision, ICCV*, 2017.

[46] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O., "Autoencoding beyond pixels using a learned similarity metric," in *Proc. Int. Conf. on Machine Learning, ICML*, 2016.

[47] Rafii, Z., Liutkus, A., Stoeter, F.-R., Mimilakis, S. I., and Bittner, R., "MUSDB18-HQ - an uncompressed version of MUSDB18," 2019, doi: 10.5281/zenodo.3338373.

[48] Richter, J., Svajda, T., and Gerkmann, T., "ReverbFX: A Dataset of Room Impulse Responses Derived from Reverb Effect Plugins for Singing Voice Dereverberation," in *ITG Conference on Speech Communication*, 2025.

[49] Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R., "SDR–half-baked or well done?" in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process., ICASSP*, 2019.

[50] Kastner, T. and Herre, J., "An Efficient Model for Estimating Subjective Quality of Separated Audio Source Signals," in *IEEE Workshop on Applications of Signal Process. to Audio and Acoust., WASPAA*, 2019.

[51] International Telecommunication Union, "Methods for subjective determination of transmission quality." Recommendation ITU-T P.800, 1996.