



# Audio Engineering Society Convention Paper 10304

Presented at the AES 160th Convention  
2026 May 28–30, Copenhagen, Denmark

*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.*

## Center Extraction GAN

Christian Uhle<sup>1,2</sup>, Julian Klapp<sup>1</sup>, Pablo Panter<sup>1</sup>, and Andreas Walther<sup>1</sup>

<sup>1</sup>Fraunhofer IIS, Erlangen, Germany

<sup>2</sup>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](mailto:christian.uhle@iis.fraunhofer.de))

### ABSTRACT

This paper presents a method for extracting a center signal from two-channel stereo signals for upmixing and reproduction with additional center loudspeakers. It uses a generative adversarial network with a generator trained with multiple reconstruction losses and adversarial losses obtained from a discriminator. The processing is of moderate computational complexity, causal and can be configured for latencies down to one audio frame of 46 ms length. It is described how training data are created using only publicly available signals and how the generation of target data enables to control the attenuation of diffuse signals and direct signals panned off-center. An evaluation with listening test and computational metrics SI-SDR and F2 measure is presented. It shows an advantage compared to methods based on classical signal processing in terms of computational metrics for source separation and listeners preference.

### 1 Introduction

Reproduction of audio signals in surround or immersive multi-channel formats finds wide adoption in cinema, broadcast, automotive audio, and home entertainment. Since most musical recordings are available in two-channel stereophony, format conversion for reproduction with more loudspeakers than there are signal channels is required, a process referred to as upmixing [1].

Multi-channel loudspeaker setups often feature a center loudspeaker placed midway in the front to improve stereo image reproduction, source localization, and clarity [2, 3]. This is particularly important when loudspeakers are not positioned symmetrically to the listener, where the perceived locations of the reproduced sources are shifted towards the most proximate loudspeaker [4], predominantly due to the precedence ef-

fect [5]. There is also evidence that intelligibility is improved and listening effort is reduced when dialog is reproduced using a center loudspeaker [6].

Center extraction involves generating a signal for the center loudspeaker, typically from two-channel stereo input signals, with applications in upmixing [7, 8, 9] or dialog enhancement [10, 11, 12]. An early approach is to compute a center signal as weighted sum of both input channels [13], with the disadvantage that all signals panned off-center are reproduced by the center loudspeaker, which results in loss of stereo width. The algorithm proposed in [7] uses a least-mean square adaptive filtering algorithm applied to the time-domain signal. Processing in time-frequency domain is especially effective, taking advantage of source sparsity [14], assuming that only single sources or multiple sources with similar spatial cues contribute to time-frequency bins. A method based on multi-channel

Wiener filtering, leading to a linear combination of all input channels with complex-valued weights, has been proposed in [9]. Other approaches use spectral weighting with real-valued weights, often based on spatial cues quantified with measures of similarity or correlation between spectral coefficients computed from both input channels [8, 15, 16, 10]. Spectral weighting with real-valued weights computed with functions of ratios of downmix and input signal are used in [17, 18], which enables control of attenuation separately for direct and diffuse signal components.

To the best of our knowledge, the application of deep learning for center extraction has not been investigated much. An approach for generating surround signals including a center signal based on style transfer has been proposed in [19].

We propose to compute center signals with generative adversarial networks (GANs) using a generator network with low computational complexity that can be configured to low latency. Training data are created by mixing anechoic audio signals and applying artificial reverberation. This implicitly enables to control the attenuation of panned direct sounds and of reverberation separately, and to attenuate reverberation as function of inter-channel level difference (ICLD) of its exciting source signal. Our approach is evaluated by means of a listening test and computational evaluation metrics. The paper is organized as follows: Sec. 2 describes the proposed implementation of center extraction with a neural network, in Sec. 3 the proposed method is evaluated, and conclusions are given in Sec. 4.

## 2 Method

### 2.1 Definition of input and target signals

To the best of our knowledge, no strict definition of desired characteristics for output signals of center extraction exists. In the following, we outline requirements that guided our work.

**Input signal model:** We assume input signals to be mixtures of direct signals which are correlated in both channels and diffuse signals with decorrelated channel signals.

**Attenuation of direct sources:** The desired center signal comprises all signal components with zero ICLD that are perceived as phantom center in stereo reproduction [13]. Direct sources with low non-zero ICLD are attenuated with gains decreasing monotonically

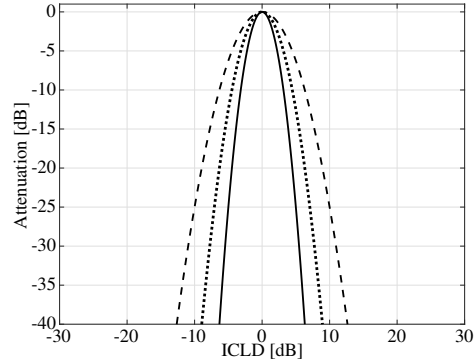


**Fig. 1:** Attenuation of source in target signal as function of its ICLD in the input as described in Subsection 2.5, parametrized to control lobe width with  $\alpha = 1$  (solid line),  $\alpha = 0.5$  (dotted line), and  $\alpha = 0.25$  (dashed line).

with absolute value of ICLD. Lateral sources (only present in one channel) should be inaudible in the center signal, because they are intended to be located in the left or right loudspeaker, and their residual components in the center signal reduce the perceived width of the stereo image.

**Attenuation of diffuse sounds:** Reproducing diffuse signals with a center loudspeaker reduces stereo width and listener envelopment, which motivates to attenuate diffuse signals in center extraction. On the other hand, attenuating reverberation can result in dereverberation artifacts. We also consider to attenuate reverberation as function of ICLD of the exciting source similarly to direct signals. This is investigated by training models using target data with and without reverberation, and comparing results with informal listening and numerical evaluation.

### 2.2 Generator

The proposed GAN uses a generator that processes complex-valued Short-Time Fourier Transform (STFT) coefficients of two-channel audio signals  $\mathbf{x}(m, k) = [X_0(m, k), X_1(m, k)]^T$ , with time index  $m$  and frequency bin index  $k$ . STFT coefficients are computed using DFT of length 2048, hop size 1024, and window size 2048, where sampling rate is 44100 Hz. Figure 2 shows a block diagram of the generator. STFT coefficients  $X$  (omitting time and frequency indices for brevity) of each audio channel are represented as triplets

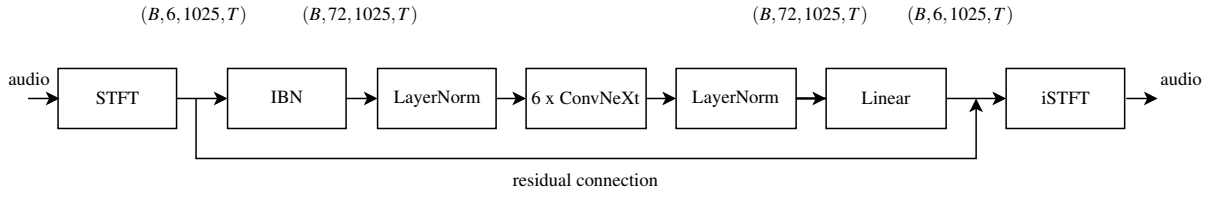


**Fig. 2:** Block diagram of generator. ConvNeXt is repeated 6 times. Tensor shapes with batch size  $B$  and number of frames  $T$ .

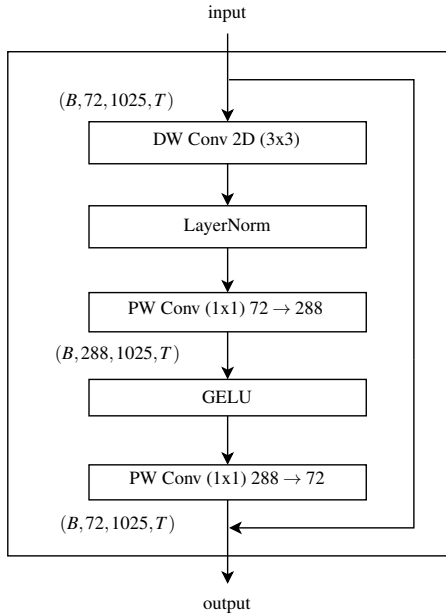


**Fig. 3:** ConvNeXt Blocks in generator, with pointwise (PW) and depthwise (DW) convolutions. Tensor shapes with batch size  $B$  and number of frames  $T$ .

$[\log(M + 10^{-7}), \text{real}(X/M), \text{imag}(X/M)]$ , with magnitudes  $M = |X|$ , yielding six channel inputs per time-frequency bin for two-channel audio signals. Magnitudes are compressed to reduce the large dynamic range of audio signals and to improve numerical optimization during training.

An inverted bottleneck (IBN) expands the input variable with six channels (for each time-frequency bin) to 72 dimensions using grouped convolution per audio channel with kernels of shape  $(3 \times 3)$ .

The latent representation is processed by layers of ConvNeXt [20] blocks shown in Figure 3. ConvNeXt blocks were proposed for image classification as

computationally efficient alternative to Vision Transformers (ViTs) [21] and later applied for generating audio signals from Mel spectrograms in [22]. In contrast to [22], we implement ConvNeXt blocks, shown in Figure 3, with 2D convolutions with kernels operating along time and frequency, reduce kernel size to  $3 \times 3$ .

Layer normalization [23] is applied before the first and after the last ConvNeXt block to improve numerical optimization by normalizing the latent towards zero mean and unit variance, where statistics are estimated given all layer inputs. This has shown 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.

### 2.3 Discriminator

The generator is trained together with three instances of complex-valued multi-resolution spectrogram discriminators (MRSDs) [24] with window lengths of [2048, 1024, 512] samples. It extends the approach of using discriminators operating on time-frequency representations at multiple resolutions proposed in [25] to process real and imaginary parts of complex-valued STFT coefficients instead of magnitudes. The discriminator is implemented with 2D convolution layers separately operating on STFT coefficients divided into five frequency bands.

We also experimented with combinations of MRSD with additional multi-scale waveform discriminator (MSD) [26] and multi-period waveform discriminator (MPD) [27], but could not observe significant improvements in informal listening.

### 2.4 Training loss

The generator is trained with multiple loss functions comprising reconstruction losses computed from target signal and generator output, and adversarial losses

obtained from discriminator instances.

**Reconstruction loss** is the weighted sum of well-known multi-resolution STFT loss [28], multi-resolution loss of complex-valued STFT coefficients, and mean squared error (MSE) loss of time signal high-pass filtered at 50 Hz cut-off frequency with low Q factor. The time signal loss is a simple and effective means to adjust the spectral balance of the generated signal. Complex-valued multi-resolution STFT loss is computed as sum of MSE of real part and MSE loss of imaginary part of target and prediction. Multi-resolution STFT loss use DFT of length [2048, 1024, 512], hop size [240, 120, 50], and window size [600, 1200, 240], whereas complex-valued multi-resolution STFT loss operate at half of these lengths to improve temporal resolution.

**Adversarial loss** is the weighted sum of least-square adversarial loss [29] and feature matching loss [30, 26], computed as L1 distance between discriminator feature maps obtained for target signals and generator output signals.

## 2.5 Training data

The training data comprises 2400 pairs (input signal  $\mathbf{x}[n]$ , target signal  $\mathbf{y}[n]$ ) of 6 sec length each with two channels sampled at 44100 Hz. The input signal is processed by the model to compute an estimate of the target signal.

**Input signals** were created by mixing four single-channel source signals from MUSDB18 (vocals, drums, bass, other accompaniment) in high quality [31] and one source signal containing speech from various audio books with intensity stereo as follows:

1. Add five source signals to create direct signal  $\mathbf{d}[n] = \sum_{i=0}^4 [g_{i,0}s_i[n+o], g_{i,1}s_i[n+o]]$  with random gains  $[g_{i,0}, g_{i,1}]$ , uniformly distributed in the range  $[-6, 6]$  dB, and single-channel signals  $s_i[n+o]$  with random time offset  $o$ .
2. Create ambience signal containing diffuse signal components  $\mathbf{a}[n] = \mathbf{d}[n] * \mathbf{r}[n]$ , with two-channel impulse responses (IRs) of reverberation [32]  $\mathbf{r}[n] = [r_0[n], r_1[n]]$ ,
3. Create input signal  $\mathbf{x}[n] = \mathbf{d}[n] + \mathbf{a}[n]$ .

Each source signal was convolved with one from a subset of 1271 IRs from the training split of ReverbFX

[32] discarding IRs with very large reverberation time and heavy modulation. The use of [31, 32] as main resource for data generation enables an approximate reproduction of our work. We selected the first channel of source signals when more channels were available to create mixes of mono signals, which is important for controlling the ICLD of direct sounds and the attenuation of panned signals in the target signal.

**Target signals** for center extraction are computed as  $\mathbf{c}[n] = \sum_{i=0}^4 h_i [g_{i,0}s_i[n], g_{i,1}s_i[n]]$  where each signal  $s_i[n]$  is scaled by attenuation  $h_i = f(\text{ICLD})$  such that signals  $s_i[n]$  panned off-center are monotonically attenuated the larger their ICLD is while signals with ICLD of one are multiplied by one. We define the attenuation as  $h = \alpha 10 \log_{10}(-|10^{0.05 \cdot g_{i,0}} / 10^{0.05 \cdot g_{i,1}}|)^2$ , shown in Figure 1, where  $\alpha$  controls the width of the lobe.

This procedure has two important implications. First, we only consider training data with intensity stereo (zero inter-channel time differences (ICTDs)). This is a disadvantage when processing recording with spaced microphones and stereophonic delay effects, and we will consider this in future work. Secondly, these target data contain direct signals without reverberation and consequently the center extraction will result in dereverberation in addition to attenuation of panned direct sources. We investigate an alternative approach by generating the reverberated target signals  $\mathbf{c}_r[n] = \mathbf{c}[n] + \mathbf{c}[n] * \mathbf{r}[n]$ . This results in target signals with artificial reverberation attenuated by  $h_i = f(\text{ICLD})$ , where  $h$  is computed from the ICLD of the signal that excites the reverberation.

## 2.6 Optimizer settings for training

Models were trained with optimizer Adam with initial learning rate  $5 \cdot 10^{-4}$  exponentially decaying with decay 0.9999995 for 400.000 iterations, gradient clipping with maximum norm of 5. Each iteration of training uses batches of 16 signals of length 0.25 seconds each.

## 2.7 Computational load, latency, and memory requirements

Training time per model is about 35 hours on an A100 GPU. The generator model has 258.000 parameters and 1186 Mega MAC per second audio signal length according to profiling with `torchinfo.summary()`. The latency of the generator equal 46 ms when using causal

padding in time direction in convolution layers, corresponding to the window size of the STFT. When convolution layers are implemented with symmetric padding, the latency increases by hop size  $\times$  (number of layers-1) to 153 ms. Lower latency is achievable when reducing frame length.

### 3 Evaluation

#### 3.1 Objective assessment

We evaluate by means of SI-SDR [33] and 2F measure [34]. Both metrics are widely used for source separation evaluations, which is related to center extraction since both potentially degrade sound quality with processing artifacts and residual signal components of interfering sources. SI-SDR quantifies the distortion caused by interfering sources and artifacts, whereas 2F measure assesses sound quality by predicting a mean score obtained in a MUSHRA listening test. We train and evaluate four models with computational evaluation metrics:

- Center Extraction GAN (CEG): trained with direct target signals  $\mathbf{c}[n]$ , with 186 ms latency
- CEG-R: trained with reverberated target signal  $\mathbf{c}_r[n]$ , with 186 ms latency
- CEG-LL: trained with direct target signals  $\mathbf{c}[n]$ , operating with low latency (46 ms),
- CEG-RLL: trained with reverberated target signal  $\mathbf{c}_r[n]$ , operating with low latency.

These are compared with two methods implemented in commercially available software, in the following referred to as S1 and S2, and the center extraction as described in [18] (GSDR). All conditions have been parametrized using a calibration item containing sources with equal level in both channels and other sources that are only present in one channel. We tried to tune these conditions such that signals only present in one channel are barely audible in the mixture, aiming at comparable separation for all conditions. This was not possible for S2, resulting in more leakage for S2.

Test items were generated with the same procedure as used for training with following modifications to have different data distributions in training and test. We use

| Metric  | SI-SDR ([dB] $\uparrow$ ) |                   | 2F ( $\uparrow$ ) |                   |
|---------|---------------------------|-------------------|-------------------|-------------------|
|         | $\mathbf{c}[n]$           | $\mathbf{c}_r[n]$ | $\mathbf{c}[n]$   | $\mathbf{c}_r[n]$ |
| CEG     | 10.6                      | 11.4              | 38.9              | 37.6              |
| CEG-R   | 11.4                      | 12.5              | 48.6              | 49.1              |
| CEG-LL  | 9.5                       | 10.0              | 30.9              | 30.3              |
| CEG-RLL | 11.1                      | 12.2              | 46.3              | 46.5              |
| S1      | 6.8                       | 7.2               | 30.7              | 30.0              |
| S2      | 9.0                       | 9.6               | 37.2              | 36.5              |
| GSDR    | 7.3                       | 7.8               | 32.9              | 32.6              |

**Table 1:** Objective evaluation using SI-SDR and 2F measure with references  $\mathbf{c}[n]$  and  $\mathbf{c}_r[n]$

the test split of MUSDB18 and test split of ReverbFX. Speech signals as used for training data were not used for evaluation, instead we ensured that all four source signals were active. Evaluation metrics are computed with reference signals according to both definitions for targets,  $\mathbf{c}[n]$  and  $\mathbf{c}_r[n]$ .

Table 1 shows results for SI-SDR and 2F measure. For all metrics, low latency models achieve lower scores in all comparisons with other proposed models. Models trained with targets with reverberation, CEG-R and CEG-RLL, perform better than CEG and CEG-LL. Comparing all conditions, best scores are achieved by model CEG-R for all scores and reference signals.

#### 3.2 Listening test

A multi-stimulus listening test compares conditions CEG, S1, S2, and GSDR, presented with an anchor signals, open input signal, and hidden input signal. We selected CEG as only condition of proposed models to reduce the number of conditions in the test. The anchor signal is the mean of both channels with lowpass filtering at 3.5 kHz cut-off frequency. Input signals were 12 commercial recordings of 5 to 15 seconds length each of rock, pop, and jazz, for which references for center signals are not available. The input signal was included in the test to let listeners know where source are panned in the input to better understanding the effect of the processing. We found it not helpful to apply loudness normalization, because processed signals with, e.g., more residuals will be attenuated more and thereby desired signals are reduced in level.

Nine expert listeners participated. The tests were done using headphones and listeners were free to choose the headphone model. Similarly to [18], listeners were

asked to rate according to two aspects: the attenuation of the off-center signal components and over-all sound quality of the signals. They were instructed to grade the unprocessed 2-channel stereo signal at 50 points on a scale from 0 to 100 and to grade all other conditions relative to that.

Figure 4 shows median and confidence interval of listener ratings per item and for all items. The proposed method is rated significantly best. GSDR and S2 are rated similar and better than S1, and it should be noted that S2 applies less attenuation than all other conditions, as described for the calibration in Subsection 3.1.

### 3.3 Illustrative Example for training with and without reverberation

Figure 5 shows spectrograms of both channel signals of input and output of models CEG and CEG-R to illustrate differences of training with and without reverberation in target signals. The source panned to center (at around second 2) is present in both output signals, whereas a source panned to right (at second 5 to 7) is attenuated. A third source, which is according to informal listening partially decorrelated, is attenuated by model CEG and only partially attenuated by model CEG-R. In general, it can be seen that CEG has more attenuation than CEG-R in the upper half of the spectrum and in the frequency range between 500 and 3000 Hz.

## 4 Conclusion

We presented Center Extraction GAN for processing two-channel audio signals to obtain a center signal that comprises direct sources panned to center. An important aspect is the generation of data required for training the models using pairs of input and reference signals which enables to control the attenuation of reverberation.

In this work training data and test data for objective evaluation were restricted to intensity stereo (zero ICTDs), however, the listening test data did not have this restriction. The results of our evaluation indicate an improvement over methods based on classical signal processing in terms of computational metrics for source separation and listeners preference.

## References

- [1] Rumsey, F., "Signal processing for 3-D audio," *J. Audio Eng. Soc.*, 56(7/8), 2008.
- [2] Theile, G., "On the performance of two-channel and multi-channel stereophony," in *Proc. AES 88th Conv.*, 1990.
- [3] Holman, T., "New factors in sound for cinema and television," *J. Audio Eng. Soc.*, 39(7-8), 1991.
- [4] Park, M., Härmä, A., Van De Par, S., and Tryfou, G., "Comparison of the width of sound sources in 2-channel and 3-channel sound reproduction," in *Proc. AES 128th Conv.*, 2010.
- [5] Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J., "The precedence effect," *J. Acoust. Soc. Am.*, 106(4), 1999.
- [6] Shirley, B., Kendrick, P., and Churchill, C., "The effect of stereo crosstalk on intelligibility: comparison of a phantom stereo image and a central loudspeaker source," *J. Audio Eng. Soc.*, 55(10), 2007.
- [7] Irwan, R. and Aarts, R., "Two-to-Five Channel Sound Processing," *J. Audio Eng. Soc.*, 50, 2002.
- [8] Avendano, C. and Jot, J.-M., "A frequency-domain approach to multi-channel upmix," *J. Audio Eng. Soc.*, 52, 2004.
- [9] Faller, C., "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.*, 54, 2006.
- [10] Härmä, A. and Park, M., "Extraction of Voice from the Center of the Stereo Image," in *Proc. AES 130th Conv.*, 2011.
- [11] 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.
- [12] Geiger, J., Grosche, P., and Parodi, Y., "Dialogue Enhancement of Stereo Sound," in *Proc. 23rd IEEE European Signal Process. Conf.*, 2015.
- [13] Klipsch, P. W., "Stereophonic sound with two tracks, three channels by means of a phantom circuit (2PH3)," *J. Audio Eng. Soc.*, 6(2), 1958.

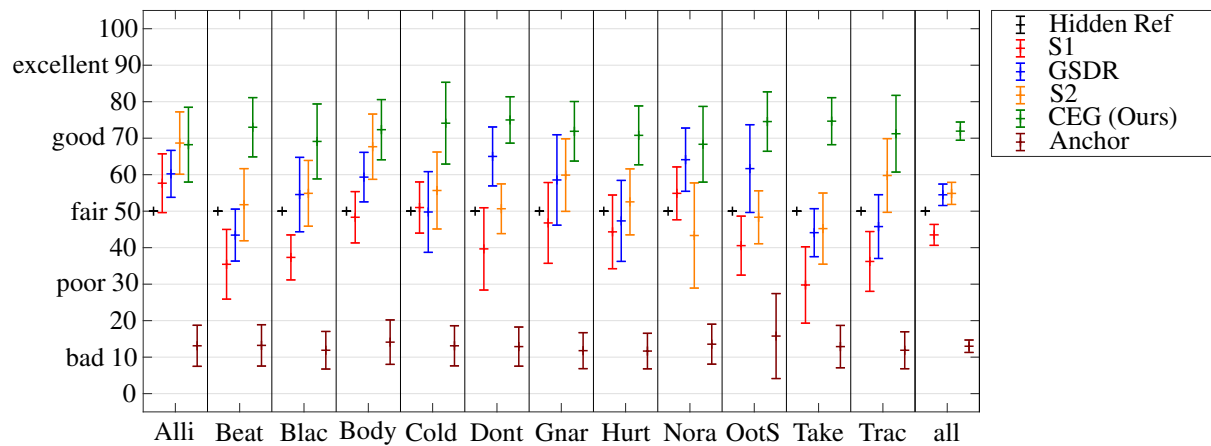


**Fig. 4:** Listening test results of multi-stimulus test with 9 listeners and 12 items, with median and confidence intervals per item and aggregated over all.

- [14] 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.
- [15] Jang, D., Hong, J., Jung, H., and Kang, K., “Center Channel Separation Based on Spatial Analysis,” in *Proc. Int. Conf. Digital Audio Effects, DAFx*, 2008.
- [16] Vickers, E., “Two-to-three Channel Upmix for Center Channel Derivation and Speech Enhancement,” in *Proc. AES 127th Conv.*, 2009.
- [17] Uhle, C., “Center Signal Scaling using Signal-to-Downmix Ratios,” in *Proc. Int. Conf. Digital Audio Effects, DAFx*, 2013.
- [18] Uhle, C. and Habets, E., “Subband Center Signal Scaling Using Power Ratios,” in *Proc. AES 53rd Int. Conf. Semantic Audio*, 2014.
- [19] 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.
- [20] 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.
- [21] Dosovitskiy, A., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [22] 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.
- [23] Ba, J. L., Kiros, J. R., and Hinton, G. E., “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] Kong, J., Kim, J., and Bae, J., “Hifi-GAN: Generative adversarial networks for efficient and high

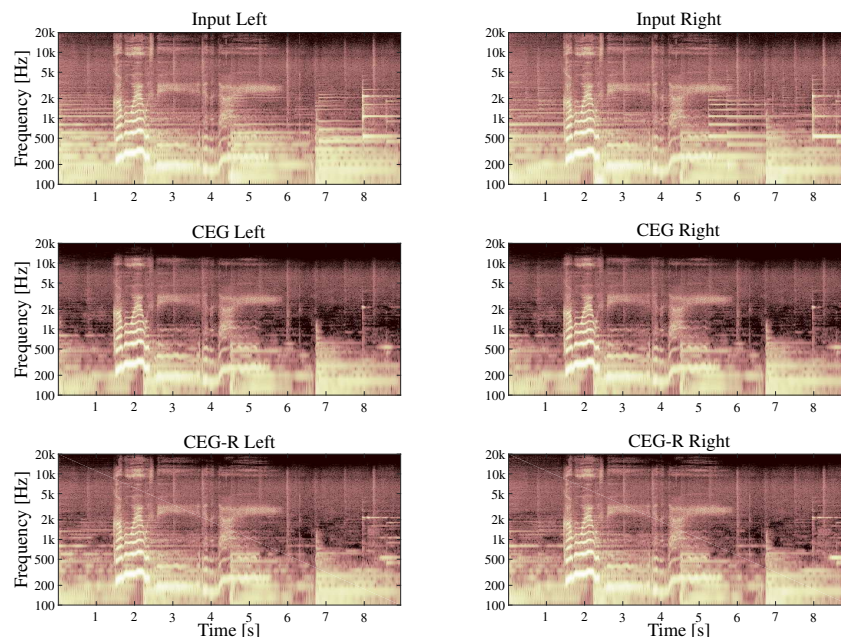


**Fig. 5:** Example illustration of center extraction by methods CEG and CEG-R discussed in Section 3.3

- fidelity speech synthesis,” *Advances in neural information processing systems*, 33, 2020.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] Rafi, Z., Liutkus, A., Stoeter, F.-R., Mimitakis, S. I., and Bittner, R., “MUSDB18-HQ - an uncompressed version of MUSDB18,” 2019, doi: 10.5281/zenodo.3338373.
- [32] 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.
- [33] 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.
- [34] 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.