Audio Source Separation via Multi-Scale Learning with Dilated Dense U-Nets

Vivek Sivaraman Narayanaswamy[†*], *Sameeksha Katoch*[†*], *Jayaraman J. Thiagarajan*[‡], *Huan Song*[+] *and Andreas Spanias*[†]

[†]Arizona State University, [‡]Lawrence Livermore National Labs, [+]Bosch Research North America

vnaray29@asu.edu, skatoch1@asu.edu, jjayaram@llnl.gov, huan.song@us.bosch.com, spanias@asu.edu

Abstract

Modern audio source separation techniques rely on optimizing sequence model architectures such as, 1D-CNNs, on mixture recordings to generalize well to unseen mixtures. Specifically, recent focus is on time-domain based architectures such as Wave-U-Net which exploit temporal context by extracting multi-scale features. However, the optimality of the feature extraction process in these architectures has not been well investigated. In this paper, we examine and recommend critical architectural changes that forge an optimal multi-scale feature extraction process. To this end, we replace regular $1-$D convolutions with adaptive dilated convolutions that have innate capability of capturing increased context by using large temporal receptive fields. We also investigate the impact of dense connections on the extraction process that encourage feature reuse and better gradient flow. The dense connections between the downsampling and upsampling paths of a U-Net architecture capture multi-resolution information leading to improved temporal modelling. We evaluate the proposed approaches on the MUSDB test dataset. In addition to providing an improved performance over the state-of-the-art, we also provide insights on the impact of different architectural choices on complex data-driven solutions for source separation.

Index Terms: Source separation, U-Net, dilated convolutions, dense connections, multi-scale feature extraction.

1. Introduction

Audio source separation refers to the problem of extracting constituent sound sources from a given audio mixture. Despite being a critical component of several audio enhancement and retrieval systems [1], the task of source separation is severely challenged in practice due to variabilities in acoustic conditions. Mathematically, this is posed as an inverse problem, and classical regularized optimization techniques such as independent component analysis (ICA) [2] and matrix factorization are often employed [3]. However, such unsupervised approaches are known to be effective only under specific conditions (e.g. fully determined) and hence several state-of-the-art solutions [4], [5], [6], [7] increasingly rely on supervisory deep learning techniques, that directly learn the inverse mapping using *mixture-source* pairs. This was motivated by the success of deep learning in solving several highly ill-conditioned inverse tasks in computer vision, such as image completion and super-resolution [8]. A recurring idea in the broad class of recent source separation techniques is to adopt an *encoder-decoder* style architecture, powered by convolutional or generative adversarial networks, for end-to-end optimization of the inversion process. While these data-driven solutions have produced unprecedented success in audio source separation, their performance depends heavily on the choice of data processing strategies and network architectures.

Until recently, majority of source separation techniques operated in the spectral domain, in particular based on the magnitude spectra. However, by ignoring the crucial phase information, these methods required extensive tuning of the front-end spectral transformation for producing accurate separation results. Recently, in [6], Stoller *et. al.* argued that the need for optimizing spectral transformations can be entirely eliminated by directly operating in the time domain, and that the source recovery quality can be significantly improved by not rejecting the phase information. On the other hand, such a fully time-domain approach necessitates the need to deal with very long temporal contexts at high sampling rates, thus making the network training quite challenging. Stoller *et. al.* addressed this critical limitation by proposing the *Wave-U-Net* model that leverages multi-scale features obtained using a combination of $1-$D-convolutions and resampling strategies in a U-Net, which is a fully convolutional network widely adopted in semantic segmentation [9]. In general, U-Nets are comprised of a *downstream* and an *upstream* module, wherein the former module produces multi-scale features by successively downsampling the audio signals while the latter utilizes resampling in order to produce appropriate context information for subsequent layers. In order to obtain meaningful gradients at different temporal scales, the network allows information propagation between the downstream and upstream layers using skip connections. Though Wave-U-Net outperformed several existing baselines, the optimality of the multi-scale feature extraction process has not been studied yet. Further, conventional upsampling was found to produce undesirable aliasing artifacts, thus requiring the design of an adaptive interpolation scheme.

Proposed Work: In this paper, we propose crucial architectural changes to the Wave-U-Net model to improve performance of time-domain based source separation systems. First, in lieu of carefully designed resampling schemes, we advocate the use of dilated convolutions to obtain robust multi-scale features. Dilated convolutions have been widely adopted in a variety of sequence modeling tasks [10] and are known to be effective in capturing temporal dependencies without any additional parameterizations to standard $1-$D convolutions. More specifically, we show that an adaptive strategy that incrementally leverages information from larger temporal contexts can produce highly robust features. Second, we propose the use of *dense* connections to improve the flexibility in exploiting long-range dependencies. In computer vision applications, DenseNets [11] have produced state-of-the-art recognition per-

*The first two authors contributed equally. This work was supported in part by the ASU SenSIP Center, Arizona State University. Portions of this work was performed under the auspices of the U.S. Depart-ment of Energy by Lawrence Livermore National Laboratory underContract DE- AC52-07NA27344.

formance, through dense connections between convolution layers within a block. While this design was primarily motivated from the standpoint of feature reuse and combating the vanishing gradient problems in very deep networks, we show that this can enable sophisticated modeling of temporal dependencies.

Using the publicly available MUSDB18 dataset [12], we investigate the effects of these architectural changes on the ill-posed inverse task of audio source separation. Following the setup in [6], we assume that the mixing process and the number of sources are known *a priori*. We first examine the impact of the proposed adaptive dilation scheme against other alternative design choices. Subsequently, we show that using dense connections can be particularly beneficial when the depth of the network increases. Our experiments show that the proposed approach, which combines both adaptive dilation and dense connections, significantly outperforms the state-of-the-art baseline.

2. Related Work

In this section, we briefly review existing approaches in the literature that utilize deep neural networks for audio source separation. There exists a large body of prior work for source separation using time-frequency representations typically, short-time Fourier transforms (STFTs) [13, 14, 15]. While [13] and [14] operated with spatial covariance matrices for source separation in the STFT domain, Luo *et al.* [15] used the magnitude spectrogram as the representation for a mixture and its constituent sources. Due to inherent challenges in phase spectrum modification, much of existing literature has focused on the magnitude spectrum, while including an additional step for incorporating the phase information, which often leads to inaccurate determination of source signals [6]. Furthermore, with low-latency systems, large window lengths are needed for effective separation in the STFT domain.

A common approach to address these drawbacks is to entirely dispense the spectral transformation step and build the estimation algorithm in the time-domain directly. Popular instantiations of this idea include the MultiResolution Convolutional Auto-Encoder (MRCAE) [4], TasNet [16] and the Wave-U-Net [6]. MRCAE [4] is an autoencoder-style architecture comprised of multiple convolution and transpose convolution layers, wherein each layer supports filters of different sizes. Note that, this is analogous to capturing audio frequencies with multi-scale resolutions. A crucial limitation of this approach is its inability to deal with long temporal sequences - results reported were with 1024-length sequences, which is often insufficient to model the complex dependencies at high sampling rates. On the other hand, TasNet [16], which is also an encoder-decoder style framework, represents an audio mixture as a weighted sum of basis signals, wherein the estimated weights indicate the contribution of each source signal and the filters from the decoder form the basis set. However, given that the architecture is designed for low-latency scenarios, similar to MRCAE, it deals with only short sequences.

In order to support the use of long temporal sequences, Stoller *et al.* [6] proposed the Wave-U-Net model, which uses a U-Net based architecture and can deal with even $80,000$-sample long sequences. While the contracting *downstream* part captures features at different scales, the expanding *upstream* part successively produces high-resolution features. Furthermore, skip connections are used between downstream and upstream layers, in order to obtain meaningful gradients at different temporal scales. However, as we show in this paper, the design of the multi-layer feature extraction process plays a critical role in the performance of this architecture. Furthermore, the training of such multi-scale feature learning networks, particular with very deep *downstream* and *upstream* modules, can be significantly challenging. We propose to incorporate dense connections, that are known to implicitly encourage feature-use [11], to alleviate this challenge.

3. Proposed Approach

The task of audio source separation involves separating a given mixture waveform $M \in \mathbb{R}^{L_m \times C}$ into K constituent source waveforms $[S_1, \dots S_K]$, where each $S_i \in \mathbb{R}^{L_n \times C}$. Here, L_m and L_n denote the lengths of the mixture and the sources respectively, and C represents the number of channels. In our formulation, we consider $L_m = L_n$, $C = 2$ implying stereo and the mixing process is a unweighted sum of sources.

3.1. Background: The Wave-U-Net Model

The proposed approach is based on the recent Wave-U-Net architecture in [6], which utilizes an *encoder-decoder* style architecture. This model follows a standard U-Net design and is comprised of 12 convolutional layers, in both the *downstream* and *upstream* parts. Each convolutional layer is followed by a factor 2 decimation to obtain successively higher resolution information along the *downstream* path. Similarly, in the *upstream path*, bilinear interpolation coupled with an $1-$D convolution layer is used to perform upsampling. In addition, skip connections are included between every convolutional layer in the downstream and upstream paths.

The number of filters in the first *downstream* layer is fixed at $f = 15$ and is increased in the subsequent layers as $f + f \times (i-1)i$ where i represents the layer index. The kernel size for the filters was chosen to be 15 in all layers. The *upstream* path also has similar filter configurations except that the kernel size was chosen to be 5. Finally, the model contains a bottleneck block consisting of a convolution layer with $f + f \times (i)$ filters and kernel size 15. Note that all convolutional layers included the LeakyReLU activation function. The final source prediction layer uses the *tanh* activation. The loss function for training the model includes the Mean Squared Error (MSE) for each of the sources. Furthermore, an energy conservation constraint is imposed by directly estimating only $K-1$ sources and obtaining the K^{th} source as the difference between the input mixture and the sum of estimates for $K-1$ sources.

3.2. Dilated U-Net

As discussed earlier, the performance of source separation approaches that operate directly in the time-domain rely heavily on the quality of the feature extraction process. In particular, building a generalizable model requires the ability to model a wide-range of temporal dependencies, which in turn requires effective multi-scale feature extraction. Furthermore, it was found in [6] that the choice of resampling scheme was very sensitive. Hence, we propose to employ dilated convolutions to seamlessly incorporate multi-scale features, thereby dispensing the need for explicit resampling. The proposed Dilated U-Net architecture is illustrated in Figure. 1.

This model consists of 6 convolutional blocks in the *downstream* path, where every block contains 3 dilated convolutions with filter configurations similar to that of Wave-U-Net. Within each block, the dilation rate of the layers increases exponentially by a factor of 2. We have chosen the dilation rate of the first layer in the consecutive block to be the same as the dilation

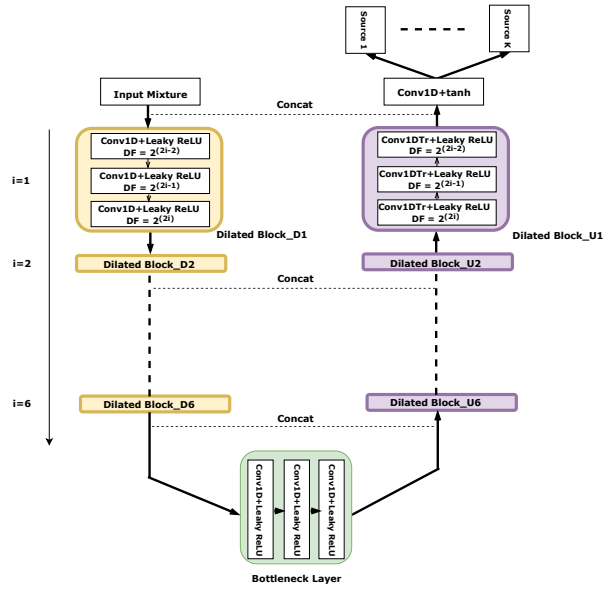


Figure 1: *Dilated U-Net - Each convolutional block consists of three* $1-D$ *convolutions with exponentially increasing dilation factors. Note that, the upstream part utilizes dilated transposed convolutions to recover the sources.*

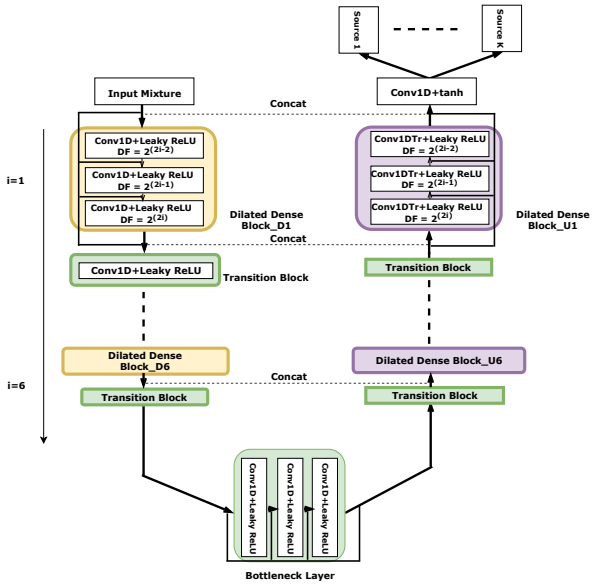


Figure 2: *Dilated Dense U-Net - Similar to Figure 1, every convolutional block is comprised of three* $1-D$ *convolutions with exponentially increasing dilation factors. In addition, we allow dense connections between convolutions within each block as well as across the downstream and upstream paths.*

rate of the last layer in preceding block. This strategy results in providing a wide range of dilation rates from $[1 \ldots 4096]$ which increases the effective receptive field, thereby producing improved multi-scale features from the audio excerpt. Note that, all layers perform convolution with a stride of 1 and employ same padding. The bottleneck block consists of three 1D convolution layers with dilation rate 1, stride 1 and same padding. Correspondingly, the upstream path also consists of 6 blocks of transposed dilated convolutions, wherein the configurations were chosen to reflect the *downstream* path. The use of skip connections, and the process of source estimation follow [6]. By retaining the training protocol and loss functions, we hope to quantify the impact of the proposed architectural changes.

3.3. Dilated Dense U-Net

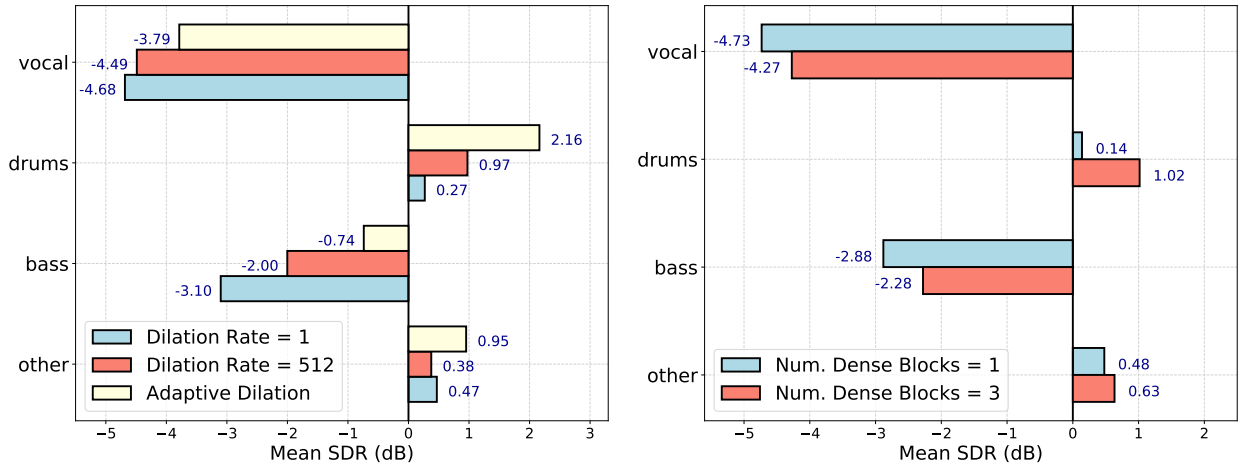
While the Dilated U-Net enables seamless incorporation of multi-scale features, with increasing depths in *downstream* and *upstream* paths, the network training becomes very challenging. We propose to improve this by employing dense connections in the networks, that supports feature reuse and protects against vanishing gradients. The Dilated Dense U-Net architecture proposed in this work is illustrated in Figure. 2. The architecture is very similar to the previous case, with the key difference that each block (a.k.a dense block), contains dense connections between the dilated convolutional layers. More specifically, within every dense block, the feature maps produced by each layer are concatenated to the subsequent layers in the block to exploit the advantages of feature reuse and improved gradient flow. This can however lead to a large number of feature maps which may be computationally infeasible to process. In order to control the growth of the number of feature maps, we include a transition block which performs dimensionality reduction at the end of every dense block.

The bottleneck block consists of three 1D convolution layers that are densely connected with the dilation rates and stride equal to 1 with same padding. Correspondingly, the *upstream* path consists of 6 dense blocks where each block contains 3 transposed convolution layers with dilation rates same as the corresponding block in the *downstream* path. Furthermore, in this model, the skip connections between the respective blocks along the paths are made dense, implying that the feature maps from the block in the *downstream* path are concatenated to all following layers in the corresponding dense block at the upstream path. Finally, the process of source extraction is identical to the Dilated U-Net.

4. Experiments

In this section, we evaluate the proposed approaches using the publicly available MUSDB18 dataset and present comparisons to the state-of-the-art Wave-U-Net model [6]. Before presenting the performance evaluation, we will first discuss the impact of different design choices on the overall performance. This study provides important insights into the behavior of source separation approaches that operate directly in the time-domain.

Experiment Setup: We use the MUSDB18 dataset [12] for our experiments, which is comprised of 75 tracks for train, 25 for validation and 50 for testing. The dataset is encoded in the stems format, and contains multi-stream files of separate sources i.e. bass, drums, other and vocals and resampled to 22050 Hz. In our experiment setup, we use segments of $16,384$ samples each ($\sim$ $1sec$) and adopt a simple additive mixing process, following current practice. Note that, in [6], the authors found that using much larger input contexts ($L_m > L_s$) produces improved results. However, to measure the effective performance of the architectural choices alone, we benchmark without the additional input context. We also performed data augmentation similar to [6], wherein the source signals are scaled using a randomly chosen factor in the interval $[0.7, 1]$. All models reported in the paper were trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 16.



(a) (b)

Figure 3: *Effect of design choices on the source separation performance (Mean SDR (dB)) - (a) Impact of the choice of dilation rates in different layers of the model. An adaptive learning provides a significant performance boost. (b) Impact of the use of dense connections as the depth of the architecture increases.*

Table 1: *Source separation performance obtained using different architectures on the MUSDB18 corpus. We show the mean and median signal-to-distortion ratio (in dB) and in each case the best results are highlighted in bold.*

Source	**Wave-U-Net**		**Dilated U-Net**		**Dilated Dense U-Net**	
	Mean SDR	**Median SDR**	**Mean SDR**	**Median SDR**	**Mean SDR**	**Median SDR**
Vocal	-3.292	2.643	-3.787	2.561	**-2.986**	**2.83**
Drums	1.435	3.310	2.163	**3.977**	**2.449**	**3.934**
Bass	-1.935	1.942	**-0.738**	**3.08**	-1.023	2.711
Other	0.986	1.911	0.953	1.945	**1.187**	**2.039**

While the results for the initial study were obtained by training for only 30 epochs, the actual performance metrics were obtained by training for longer ($\sim$ 80 epochs). The mean and median signal-to-distortion ratio (SDR) for each of the sources over the entire dataset are computed. The SDR metric takes into account the noise arising from interference and other artifacts in the estimated audio sources [17]. The mean SDR is computed after removing silence regions. Since the mean value can be affected by outliers from near-silence regions we also report the median SDR which is known to be more unbiased.

4.1. Impact of Design Choices

As discussed earlier, the source separation performance depends heavily on the architecture choices for multi-scale feature extraction. Hence, we first study the impact of different dilation schemes in the proposed architecture, wherein we entirely eliminate the resampling process using dilated convolutions. As described in the previous section, our architecture is comprised of 6 blocks of convolutional layers. In its simplest form, we use conventional $1-$D convolutions with the dilation rate fixed at 1 in all layers. In addition, we consider the case where it was fixed at a constant value (512) and the case with the proposed adaptive dilation scheme. As observed in Figure 3(a), the sub-optimal performance of conventional $1-$D convolution clearly shows the importance of leveraging multi-scale features. Furthermore, the proposed adaptive dilation scheme provides a significant performance boost compared to using fixed dilation in all layers. Similarly, we analyzed the impact of using dense connections on the separation performance. For this experiment, we fixed the dilation rate at a constant value of 512 and the number of convolutional blocks at 1 and 3 respectively. As showed in Figure 3(b), as the depth of the network increases, using dense connections provides significant gains.

4.2. Performance Evaluation

In this section, we report the overall performance of the proposed approaches, namely Dilated U-Net and Dilated Dense U-Net, on the MUSDB18 dataset. Though a number of baseline techniques exist for time-domain source separation, we chose to compare against the state-of-the-art Wave-U-Net architecture from [6]. Table 1 compares the mean/median SDR (dB) for each of the constituent sources for the testing set in MUSDB18. The first striking observation is that by improving the multi-scale feature extraction process, we could obtain significant performance improvements over the baseline in all cases. In particular, our approaches provide improvements between 0.2dB and 1.2dB. While the dilated variant eliminates the need for explicit resampling by capturing information from exponentially increasing receptive fields, the inclusion of dense connections improves the robustness of the training process. This performance gain clearly evidences the dependence of these complex data-driven solutions for audio processing on the inherent feature extraction mechanism, and the need for improved architecture design.

5. References

[1] A. Spanias, T. Painter, and V. Atti, *Audio signal processing and coding*. John Wiley & Sons, 2006.

[2] S. Makino, S. Araki, R. Mukai, and H. Sawada, "Audio source separation based on independent component analysis," in *2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No.04CH37512)*, vol. 5, May 2004, pp. V–V.

[3] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, "Mixing matrix estimation using discriminative clustering for blind source separation," *Digital Signal Processing*, vol. 23, no. 1, pp. 9–18, 2013.

[4] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1577–1581.

[5] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

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

[7] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," 2017.

[8] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[10] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

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

[13] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 261–265.

[14] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 323–332.

[15] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 61–65.

[16] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[17] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.