D3NET: DENSELY CONNECTED MULTIDILATED DENSENET FOR MUSIC SOURCE SEPARATION

Naoya Takahashi, Yuki Mitsufuji

Sony Corporation, Japan

ABSTRACT

Music source separation involves a large input field to model a long-term dependence of an audio signal. Previous convolutional neural network (CNN) -based approaches address the large input field modeling using sequentially down- and up-sampling feature maps or dilated convolution. In this paper, we claim the importance of a rapid growth of a receptive field and a simultaneous modeling of multi-resolution data in a single convolution layer, and propose a novel CNN architecture called densely connected dilated DenseNet (D3Net). D3Net involves a novel multi-dilated convolution that has different dilation factors in a single layer to model different resolutions simultaneously. By combining the multi-dilated convolution with DenseNet architecture, D3Net avoids the aliasing problem that exists when we naively incorporate the dilated convolution in DenseNet. Experimental results on MUSDB18 dataset show that D3Net achieves state-of-the-art performance with an average signal to distortion ratio (SDR) of 6.01 dB.[1]

Index Terms— source separation, DenseNet, SiSEC

1 Introduction

Music source separation (MSS) has been intensively studied and neural-network-based approaches have shown impressive progress in recent years. Many types of neural network architecture have been proposed, including feedforward fully connected networks (FNNs) [1, 2], recurrent neural networks (RNNs) [3], convolutional neural networks (CNNs) [4–9] and their combinations [10, 11]. In particular, CNNs have attracted great attention because of their superior performance, parameter efficiency and generality for different types of data. Takahashi et. al. [10] applied a CNN architecture with a dense skip connectivity pattern, called DenseNet [12], to MSS and obtained state-of-the-art results in SiSEC 2018 [13]. Such dense connectivity allows maximum information flow and deeper CNN while keeping the model size small by efficiently reusing intermediate representations of preceding layers.

One of the benefits of a deeper CNN is its larger receptive field that allows a large context to be modeled, which is important since audio signals can have long time and wide frequency band dependences. Although the receptive field grows linearly with the number of layers stacked, it is not the optimal way to increase the receptive field only by stacking convolutional layers, as it requires too many layers to cover a sufficiently large input field to model global information, making the network training too difficult. A popular approach to incorporating a large context is to repeatedly downsample intermediate network outputs and apply operations in lower resolution representations. The low-resolution representations are again upsampled to recover the lost resolution while carrying over the global perspective from downsampled layers [4, 5, 7, 10]. Another approach is a dilated convolution, which is shown to be effective for audio generation and MSS tasks [7, 11, 14]. The dilation factors are set to grow exponentially with the number of layers stacked and, therefore, the networks cover the large receptive field with a small number of layers. Although the down–upsampling structure and dilated convolution allow a large receptive field, each layer in the network sees only one resolution at a time. However, the simultaneous consideration of the local and global information can be useful, e.g., the local structure can be more precisely estimated by using global structure information and vise versa. DenseNet partially addresses this problem by means of the dense skip connectivity that allows the direct aggregation of features from early layers and features in later layers within a single convolution layer. However, it may still be too slow to transform local features to global features and it is inefficient to have many parameters, especially for high-resolution data.

In this work, we combine the advantages of DenseNet and dilated convolution, and propose a novel network architecture called dilated DenseNet (D2Net). To properly combine DenseNet with the dilated convolution, we propose a multi-dilated convolution layer that has a multiple dilation factor within a single layer. The dilation factor depends on which skip connection the channels come from, as shown in Fig. 1. The multidilated convolution can prevent the aliasing that occurs when a standard dilated convolution is applied to feature maps with receptive fields smaller than the dilation factor. Although a naive combination of DenseNet with dilation has already been proposed [15], standard dilated convolutions are used and dilation factors are determined depends on the layer depth, which causes considerable aliasing. In contrast, we show the effectiveness of the proposed multidilated convolution in our ablation study. Furthermore, we propose a nested architecture of dilated

---

[1] The conference version of this paper, which includes extended works, is available. Please refer to Naoya Takahashi et al. "Densely connected multidilated convolutional networks for dense prediction tasks", CVPR2021
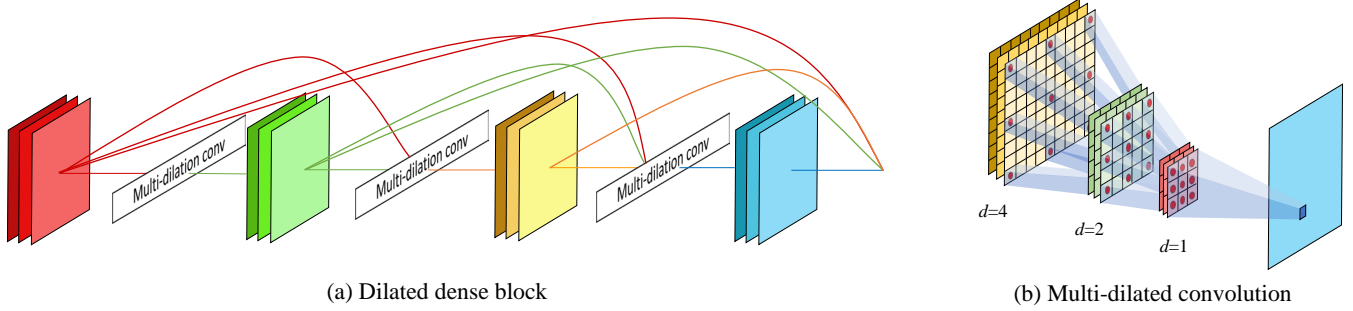
(a) Dilated dense block   (b) Multi-dilated convolution

Fig. 1. Illustration of D2 block. (a) The connectivity pattern is the same as in DenseNet except that the D2 block involves the multi-dilated convolution. (b) Illustration of the multi-dilated convolution at the third layer. To produce a single feature map, it involves multiple dilation factors depending on the input channel. For clarity, we omit the normalization and nonlinearity from the illustration.

dense blocks to effectively repeat dilation factors multiple times with dense connections that ensure the sufficient depth required for modeling each resolution. We call the nested architecture densely connected dilated DenseNet (D3Net).

The contributions of this work are summarized below:

1. We claim that a naive incorporation of dilation in DenseNet architecture can cause a significant aliasing problem, and propose a multidilated convolution layer to properly incorporate the dilated convolution into DenseNet.

2. We further introduce the D3Net architecture of nested dilated dense blocks to effectively apply different dilation factors multiple times.

3. We experimentally show the effectiveness of the proposed architectures. The D3Net achieves state-of-the-art results on MUSDB18 dataset.

2   Multidilated convolution for DenseNet

In DenseNet, the outputs of the lth convolutional layer x_l are computed using filters k_l and outputs of all preceding layers as

$$x_l = \psi([x_0, x_1, \cdots, x_{l-1}]) \circledast k_l, \qquad (1)$$

where $\psi()$ denotes the composite operation of batch normalization and nonlinearity, $[x_0, x_1, \cdots, x_{l-1}]$ the concatenation of feature maps from $1, \cdots, l-1$ layers, and $\circledast$ the convolution. A naive way of incorporating dilated convolution is to replace the convolution $\circledast$ with the dilated convolution $\circledast_d$ with the dilation factor $d = 2^{l-1}$. However, this causes a severe aliasing problem; for instance, at the third layer, input is subsampled with 4 sample intervals without any anti-aliasing filtering because of the skip connections. Assuming that the kernel size is 3, only the path that passes through all convolution operations without any skip connection covers the input field without omission and all other paths from skip connections have *blind spots* in their receptive fields that inherently make it impossible for proper ant-aliasing filters to be learned in the preceding layers (Fig. 2a). To overcome this problem, we propose the multidilated convolution $\circledast_l^m$

defined as

$$Y_l \circledast_l^m k_l = \sum_{i=0}^{l-1} y_i \circledast_{d_i} k_l^i, \qquad (2)$$

where $Y_l = [y_0, y_1, \cdots, y_{l-1}] = \psi([x_0, x_1, \cdots, x_{l-1}])$ is the composite layer output, k_l^i the subset of filters that correspond to the ith skip connection, and $d_i = 2^i$. As depicted in Fig. 2b, DenseNet with the proposed multidilated convolution has different dilation factors depending on which layer the channel comes from. This allows the receptive field to cover the input field without the loss of coverage between the samples the filters to be applied and, hence, to learn proper filters to prevent aliasing. One advantage of the dilated dense block (D2 block) is its ability to integrate information from very local to exponentially large receptive field within a single layer. This fast information flow provides more flexibility in modeling information in a wide range of resolutions.

Note that the multidilation convolution is not equivalent to applying the multibranch convolution where convolutions with different dilation factors are applied to the same input feature maps, similar to the Inception block [16–18], again causing the aliasing problem.

3   D3Net

Although the D2 block provides an exponentially large receptive field as the number of layers increases, it is also worthwhile to provide sufficient flexibility to transform feature maps in each resolution. In WaveNet [14], dilation factors are reset to one after several layers are stacked and repeated; that is, the dilation factor in the lth layer is given by $d_l = 2^{l-1 \bmod M}$, where mod is the modulo operation and M is the number of layers at which the dilation factor is doubled. Inspired by this work, we propose a nested architecture of D2 blocks as shown in Fig. 3. D2 blocks are considered as single composite layers and are densely connected in the same way as within the D2 block itself. We also employ a channel reduction mechanism at the end of each D2 block to mitigate the growth of an excessive number of channels and thus improve computational efficiency. The channel reduc-
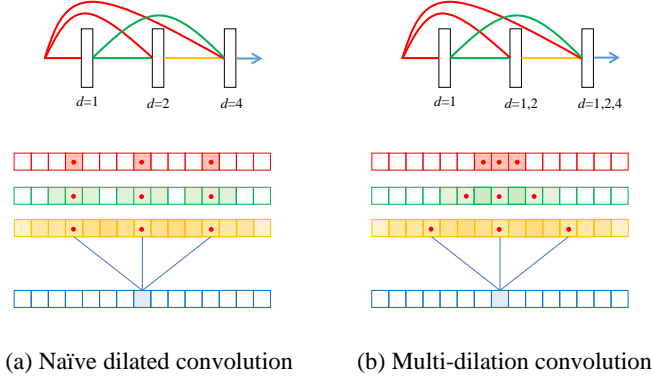
(a) Naïve dilated convolution   (b) Multi-dilation convolution

Fig. 2. Visualization of receptive fields at the third layer of (a) naive integration of dilated convolution and (b) proposed multi-dilated convolution (in the case of one dimension). Red dots denote the points to which filters are applied, and the colored background shows the receptive field covered by the red dot.
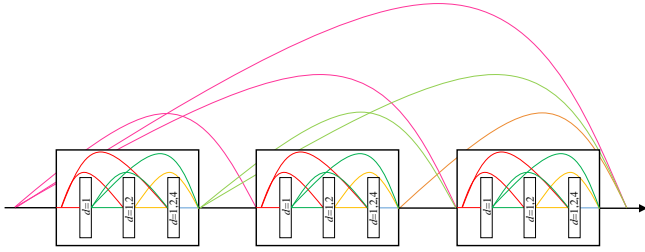

Fig. 3. D3 block densely connects D2 blocks with repeated dilation pattern.

tion can be performed by either a 1×1 convolution or simply passing the output of the last N layers' to the next block. In this work, we take the latter approach since performance characteristics of both methods are similar, but the former approach requires slightly more computations. Note that without the channel reduction, the architecture is reduced to a standard dense connection with repeated multidilation factors.

4 Experiments

Dataset We evaluated the proposed method using the MUSDB18 dataset, prepared for SiSEC 2018 [13]. In the dataset, approximately 10 hours of professionally recorded 150 songs in stereo format at 44.1kHz are available. For each song, a mixture and its four sources, *bass, drums, other* and *vocals*, are provided and thus, the task is to separate the four sources from the mixture. We adopted the official split of 100 and 50 songs for *Dev* and *Test* set, respectively. Short-time Fourier transform (STFT) magnitude frames of the mixture, windowed at 4096 samples with 75% overlap, with data augmentation [3] were used as inputs.

Training The four networks for each source instrument were trained to estimate the source spectrogram by minimizing the mean square error with the Adam optimizer for 50 epochs. The patch length was set to 256 frames; thus, the dimensions of input were $2 \times 256 \times 2049$. The batch size was set to 6. The learning rate was initially set to 0.001 and annealed to 0.0001 at 40 epochs.

Model architecture Following [4,10], in which the best results obtained in SiSEC 2018 were reported, we used the multiscale multiband architecture in which band-dedicated modules and a full band module, each with a bottleneck encoder–decoder architecture with skip connections, are placed. The network configuration is shown in Table 1. The network outputs are used to calculate the multichannel Wiener filter (MWF) to obtain the final separations, as commonly performed in frequency domain audio source separation methods [3,8,10,11].

Results The signal-to-distortion ratio (SDR) of our proposed method and existing state-of-the-arts methods are compared in Table 2. The SDRs were computed using the *museval* package [13] and median SDRs are reported. TAK1 [10] and UHL2 [3] are the two best performing methods in SiSEC 2018 (among submissions that do not use external data) and the network architectures are the combination of DenseNet and recurrent units for TAK1 and an ensemble of bi-directional LSTM models for UHL2. The proposed D3Net exhibited the best performance for *vocals*, *drums* and *accompaniment* (the summation of *drums*, *bass* and *other*) and performed comparably to the best method for *other*. The average SDR of four instruments is significantly better than all baseline values. The primally difference between MMDenseLSTM (TAK1) and the proposed method is that MMDenseLSTM incorporates LSTM units to further expand the receptive field, whereas the proposed method uses the multidilated convolution. Comparison of these methods indicate the effectiveness of the multidilated convolution. On the other hand, GRU dilation 1 [11] consists of dilated convolution and dilated GRU units without a down–up-sampling path. This also highlights the effectiveness of the multiresolution modeling of the multidilation convolution with the dense connection. For *bass*, approaches that operate in the time domain perform better, as they are capable of recovering the target phase, which is easier in the low frequency range. Among the frequency domain approaches, D3Net performs the best.

We also conducted an ablation study to validate the effectiveness of the multidilated convolution. By replacing the multidilated convolutions with the standard convolutions without dilation, we obtained comparable results as the best performing model in SiSEC2018, TAK1 (MMDenseLSTM). When we replaced the multidilated convolution with the standard dilated convolution, we obtained a decent improvement over the D3Net without dilation even though the aliasing problem arises. However, the proposed multidilated convolution clearly outperforms the standard dilated convolution, showing the importance of handling the aliasing problem in order to incorporate dilation in DenseNet.

We further investigated the effects of the multidilated con-

Table 1. Proposed architectures. All D3 blocks have 3×3 kernels with growth rate k, L layers, and M D2 blocks.

Layer	scale	Vocals, Other			Drums			Bass		
		low	high	full	low	high	full	low	high	full
band split index		1-256	257-1600	-	1-128	128-1600	-	1-192	192-1600	-
conv (t×f,ch)	1	3×3, 32	3×3, 8	3×3, 32	3×3, 32	3×3, 8	3×3, 32	3×3, 32	3×3, 8	3×3, 32
D3 block 1 (k,L,M)		16, 5, 2	2, 1, 1	13, 4, 2	16, 5, 2	2, 1, 1	13, 4, 2	16, 5, 2	2, 1, 1	10, 4, 2
down sample	$\frac{1}{2}$	avg. pool 2 × 2			avg. pool 2 × 2			avg. pool 2 × 2		
D3 block 2 (k,L,M)		18, 5, 2	2, 1, 1	14, 5, 2	18, 5, 2	2, 1, 1	14, 5, 2	18, 5, 2	2, 1, 1	10, 5, 2
down sample	$\frac{1}{4}$	avg. pool 2 × 2			avg. pool 2 × 2			avg. pool 2 × 2		
D3 block 3 (k,L,M)		20, 5, 2	2, 1, 1	15, 6, 2	20, 5, 2	2, 1, 1	15, 6, 2	18, 5, 2	2, 1, 1	12, 6, 2
down sample	$\frac{1}{8}$	avg. pool 2 × 2			avg. pool 2 × 2			avg. pool 2 × 2		
D3 block 4 (k,L,M)		22, 5, 2	2, 1, 1	16, 7, 2	22, 4, 2	2, 1, 1	16, 7, 2	20, 5, 2	2, 1, 1	14, 7, 2
down sample	$\frac{1}{16}$	avg. pool 2 × 2			avg. pool 2 × 2			avg. pool 2 × 2		
D3 block 5 (k,L,M)		-	-	17, 8, 2	-	-	16, 8, 2	-	-	16, 8, 2
up sample	$\frac{1}{8}$	t.conv 2 × 2			t.conv 2 × 2			t.conv 2 × 2		
concat.		-	-	D3 block 4	-	-	D3 block 4	-	-	D3 block 4
D3 block 6 (k,L,M)		-	-	16, 6, 2	-	-	16, 6, 2	-	-	14, 6, 2
up sample	$\frac{1}{4}$	t.conv 2 × 2			t.conv 2 × 2			t.conv 2 × 2		
concat.		D3 block 3	D3 block 3	D3 block 3	D3 block 3	D3 block 3	D3 block 3	D3 block 3	D3 block 3	D3 block 3
D3 block 7 (k,L,M)		20, 4, 2	2, 1, 1	14, 5, 2	20, 4, 2	2, 1, 1	14, 6, 2	18, 4, 2	2, 1, 1	12, 6, 2
up sample	$\frac{1}{2}$	t.conv 2 × 2			t.conv 2 × 2			t.conv 2 × 2		
concat.		D3 block 2	D3 block 2	D3 block 2	D3 block 2	D3 block 2	D3 block 2	D3 block 2	D3 block 2	D3 block 2
D3 block 8 (k,L,M)		18, 4, 2	2, 1, 1	12, 4, 2	18, 4, 2	2, 1, 1	12, 4, 2	16, 4, 2	2, 1, 1	8, 4, 2
up sample	1	t.conv 2 × 2			t.conv 2 × 2			t.conv 2 × 2		
concat.		D3 block 1	D3 block 1	D3 block 1	D3 block 1	D3 block 1	D3 block 1	D3 block 1	D3 block 1	D3 block 1
D3 block 9 (k,L,M)		16, 4, 2	2, 1, 1	11, 4, 2	16, 4, 2	2, 1, 1	11, 4, 2	16, 4, 2	2, 1, 1	8, 4, 2
concat. (axis)		freq		-	freq		-	freq		-
concat. (axis)	1	channel			channel			channel		
d2 block (k,L)		12, 3			12, 3			12, 3		
gate conv (t×f,ch)		3 × 3, 2			3 × 3, 2			3 × 3, 2		

Table 2. *SDR values for MUSDB18 dataset. SDR values are median of median SDR of each song. '*' denotes method operating in time domain.*

Method	SDR in dB					
	Vocals	Drums	Bass	Other	Acco.	Avg.
TAK1 (MMDenseLSTM) [10]	6.60	6.43	5.16	4.15	12.83	5.59
UHL2 (BLSTM ensemble) [3]	5.93	5.92	5.03	4.19	12.23	5.27
GRU dilation 1 [11]	6.85	5.86	4.86	**4.65**	13.40	5.56
UMX [19]	6.32	5.73	5.23	4.02	-	5.33
demucs* [7]	6.29	6.08	5.83	4.12	-	5.58
Meta-TasNet* [8]	6.40	5.91	5.58	4.19	-	5.52
Nachmani et. al.* [20]	6.92	6.15	**5.88**	4.32	-	5.82
D3Net w/o dilation	6.86	6.37	4.97	4.21	13.19	5.60
D3Net standard dilation	7.12	6.61	5.19	4.53	13.39	5.86
D3Net (proposed)	**7.24**	**7.01**	5.25	4.53	**13.52**	**6.01**

Fig. 4. Comparison of normalized L1 norm of weights in the last layer of first d3 block.

volution by assessing the learned weights. We calculated the L1 norm of the convolution weights in the last layer of the first block for both the proposed D3Net with the multidilated convolution and the baseline D3Net with the standard dilated convolution. The norm was calculated separately for each skip connection and norm values were normalized by the norm of the path with no skip connection. Fig. 4 shows that the weights of the skip connection from early layers have smaller norms than later layers. This trend is much more prominent for D3Net with the standard dilated convolution than for D3Net with the proposed multidilated convolution. This results also indicate that applying a dilated convolution to skip connections from early layers without handling the aliasing problem makes it difficult to extract information from them and, therefore, the network assigns a low norm to them.
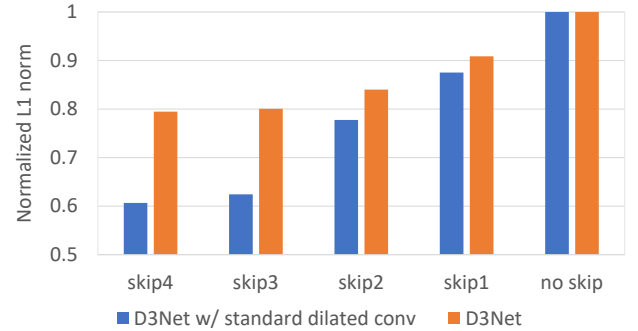
Finally, we trained D3Net with 1500 extra songs to study how much D3Net can be generalized by using larger dataset. In Table 3, we summarize the SDR values of D3Net and other methods that utilize extra data in addition to MUSDB. Although these methods are not directly comparable since the extra data are different for every methods, we observe that the performance of D3Net is greatly improved by the aid of extra data, and obtain state-of-the-art results on *vocals*, *other*, *accompaniment* and the average SDR.

Table 3. *Comparison of models that use external data for training. SDR values are for MUSDB18 test set. '*' denotes method operating in time domain.*

Method	Extra data	SDR in dB					
		Vocals	Drums	Bass	Other	Acco.	Avg.
TAK2 (MMDenseLSTM) [10]	800 songs	7.16	6.81	5.40	4.80	13.73	6.04
demucs* [7]	150 songs	7.05	7.08	6.70	4.47	-	6.33
Spleeter [21]	24,097 songs / 79 hours	6.68	6.71	5.51	4.02	12.54	5.78
TasNet* [22][23]	300 hours	7.34	**7.68**	**7.04**	4.04	13.76	6.52
D3Net (proposed)	-	7.24	7.01	5.25	4.53	13.52	6.01
D3Net (proposed)	1,500 songs / 93 hours	**7.80**	7.36	6.20	**5.37**	**14.26**	**6.68**

5 Conclusion

We proposed a novel neural network architecture called D3Net. D3Net employs the multidilated convolution with dense skip connections that enables the local and global feature information to be modeled simultaneously within a single layer. Experimental results showed that D3Net achieves state-of-the-art results for the MUSDB18 dataset. The ablation study demonstrated the importance of handling the aliasing problem when we combine DenseNet with the dilated convolution.

6 References

[1] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel music separation with deep neural networks," in *Proc. EUSIPCO*, 2015.

[2] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *Proc. ICASSP*, 2015, pp. 2135–2139.

[3] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep networks through data augmentation and network blending," in *Proc. ICASSP*, 2017, pp. 261–265.

[4] N. Takahashi and Y. Mitsufuji, "Multi-scale Multi-band DenseNets for Audio Source Separation," in *Proc. WAS-PAA*, 2017, pp. 261–265.

[5] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proc. ISMIR*, 2018.

[6] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "Phasenet: Discretized phase modeling with deep neural networks for audio source separation," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[7] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[8] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," in *Proc. ICASSP*, 2020.

[9] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and state-of-the art music source separation tool with pre-trained models," Late-Breaking/Demo ISMIR 2019, November 2019, Deezer Research.

[10] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MM-DenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. IWAENC*, 2018.

[11] J.-Y. Liu and Y.-H. Yang, "Dilated convolution with dilated GRU for music source separation," in *International Joint Conferences on Artificial Intelligence Organization (IJCAI)*, 2019.

[12] G. Huang, Z. Liu, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. CVPR*, 2017.

[13] A. Liutkus, F.-R. Stöter, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proc LVA/ICA*, 2018.

[14] A. van den 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.

[15] A. Fuchs, R. Priewald, and F. Pernkopf, "Recurrent dilated densenets for a time-series segmentation task," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2019, pp. 75–80.

[16] B. McMahan and D. Rao, "Listening to the world improves speech command recognition," in *Proc. AAAI*, 2018.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. CVPR*, 2015.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, 2016.

[19] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix - a reference implementation for music source separation," *Journal of Open Source Software*, 2019.

[20] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. ICML*, 2020.

[21] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, pp. 2154, 2020, Deezer Research.

[22] E. Pierson Lancaster and N. Souviraà-Labastie, "A frugal approach to music source separation," working paper or preprint, Nov. 2020.

[23] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *Trans. Audio, Speech, and Language Processing*, 2019.