ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094992

TF-GRIDNET: MAKING TIME-FREQUENCY DOMAIN MODELS GREAT AGAIN FOR MONAURAL SPEAKER SEPARATION

Zhong-Qiu Wang[1], *Samuele Cornell*[1,2], *Shukjae Choi*[3], *Younglo Lee*[3], *Byeong-Yeol Kim*[3], *Shinji Watanabe*[1]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
[2]Università Politecnica delle Marche, Italy [3]Hyundai Motor Group and 42dot Inc., Seoul, Korea

wang.zhongqiu41@gmail.com

ABSTRACT

We propose TF-GridNet, a novel multi-path deep neural network (DNN) operating in the time-frequency (T-F) domain, for monaural talker-independent speaker separation in anechoic conditions. The model stacks several multi-path blocks, each consisting of an intra-frame spectral module, a sub-band temporal module, and a full-band self-attention module, to leverage local and global spectro-temporal information for separation. The model is trained to perform complex spectral mapping, where the real and imaginary (RI) components of the input mixture are stacked as input features to predict target RI components. Besides using the scale-invariant signal-to-distortion ratio (SI-SDR) loss for model training, we include a novel loss term to encourage separated sources to add up to the input mixture. Without using dynamic mixing, we obtain 23.4 dB SI-SDR improvement (SI-SDRi) on the WSJ0-2mix dataset, outperforming the previous best by a large margin.

Index Terms— Complex spectral mapping, speaker separation.

1. INTRODUCTION

Dramatic progress has been made in monaural talker-independent speaker separation since the invention of deep clustering [1] and permutation invariant training (PIT) [2]. Early studies train DNNs to perform separation in the magnitude domain, with or without using magnitude-based phase reconstruction [3–7]. Subsequent studies perform separation in the complex T-F domain through complex ratio masking [8] or in the time domain through the encoder-separator-decoder scheme proposed in TasNet [9–11]. Since 2019, TasNet and its variants [11–23], which feature learned encoder and decoder operating on very short windows of signals, have gradually become the most popular and dominant approach for speaker separation in anechoic conditions, largely due to their strong performance and advanced DNN architectures designed for end-to-end optimization. Their performance on WSJ0-2mix [1], the de-facto benchmark dataset for speaker separation in anechoic conditions, has reached an impressive 22.1 dB SI-SDRi in a recent study [24].

In the meantime, T-F domain models, which typically use larger window sizes and hop sizes, have been largely under-explored and under-represented in anechoic speaker separation. Recently, TFPSNet [25], which claims to operate in the complex T-F domain, reports on WSJ0-2mix a strong SI-SDRi at 21.1 dB, which is comparable to the top results achievable by modern time-domain models. It also shows stronger cross-corpus robustness than representative time-domain models. Following DPRNN [15] and DPTNet [17, 26], it leverages a modern dual-path architecture but takes in a complex T-F spectrogram as input [26, 27], and uses the transformer module proposed in DPTNet [17] to model cross-frequency information in each frequency-scanning path and cross-frame information in each time-scanning path. Although TFPSNet claims to operate in the T-F domain [25], it closely follows the encoder-separator-decoder scheme [11] widely adopted in many TasNet variants: it (1) uses a one-dimensional (1D) convolution (Conv1D) layer followed by rectified linear units to encode the RI components of each T-F unit to a higher-dimensional embedding with non-negative values; (2) uses a dual-path separator to produce a non-negative mask to mask the embeddings for separation; and (3) applies a fully-connected layer at each T-F unit to predict the RI components of each speaker. Its performance, however, does not surpass contemporary time-domain models [22, 23], even with an advanced DNN architecture.

Table 1: Comparison with other systems on WSJ0-2mix.

Systems	Domain	Year	#params (M)	SI-SDRi (dB)	SDRi (dB)
DPCL++ [3]	T-F	2016	13.6	10.8	-
uPIT-BLSTM-ST [2]	T-F	2017	92.7	-	10.0
ADANet [28]	T-F	2018	9.1	10.4	10.8
WA-MISI-5 [5]	T-F	2018	32.9	12.6	13.1
Sign Prediction Net [7]	T-F	2019	56.6	15.3	15.6
Conv-TasNet [11]	Time	2019	5.1	15.3	15.6
Deep CASA [8]	T-F	2019	12.8	17.7	18.0
Conv-TasNet-MBT [12]	Time	2020	8.8	15.6	-
FurcaNeXt [13]	Time	2020	51.4	-	18.4
SUDO RM -RF [14]	Time	2020	2.6	18.9	-
DPRNN [15]	Time	2020	2.6	18.8	19.0
Gated DPRNN [16]	Time	2020	7.5	20.1	20.4
DPTNet [17]	Time	2020	2.7	20.2	20.6
DPTCN-ATPP [18]	Time	2021	4.7	19.6	19.9
SepFormer [19]	Time	2021	26.0	20.4	20.5
Sandglasset [20]	Time	2021	2.3	20.8	21.0
Wavesplit [21]	Time	2021	29.0	21.0	21.2
TFPSNet [25]	T-F	2022	2.7	21.1	21.3
MTDS (DPTNet) [22]	Time	2022	4.0	21.5	21.7
SFSRNet [23]	Time	2022	59.0	22.0	22.1
QDPN [24]	Time	2022	200.0	22.1	-
TF-GridNet	T-F	2022	14.4	**23.4**	**23.5**

In this context, our study makes the following contributions to improve the performance of complex T-F domain approaches:

- We propose to use complex spectral mapping for speaker separation in anechoic conditions. Complex spectral mapping, initially proposed in [29–31], has shown strong potential on noisy-reverberant speech separation when combined with modern DNN architectures and loss functions [32–37], exhibiting strong robustness to noise and reverberation in both single- and multi-microphone conditions. Its potential on anechoic speaker separation, however, has not been studied, especially in an era when time-domain models, which usually perform masking in learned embedding space, have become so popular and dominant on this task. This paper is the first study to explore this direction.
- We propose a novel DNN architecture, named TF-GridNet, for speech separation. It operates in the complex T-F domain to model the spectro-temporal patterns in two-dimensional (2D), grid-like spectrograms. Besides refining TFPSNet [25], our study adds

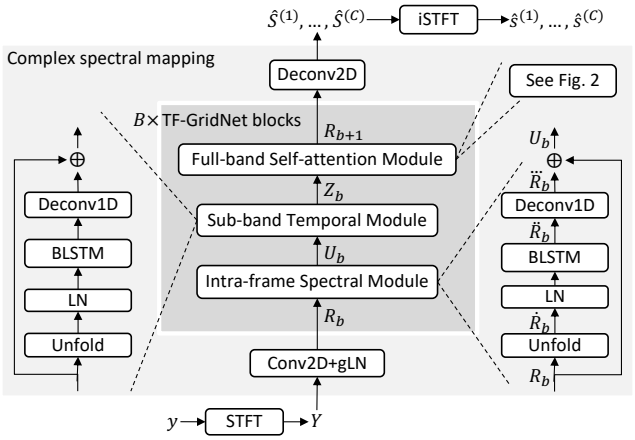

Fig. 1: Overview of proposed system.

a full-band self-attention path for dual-path models to leverage cross-frame global information, resulting in a multi-path model.

- Building upon the popular SI-SDR loss [11,38], we devise a novel time-domain loss term to encourage the summation of estimated sources to be close to the mixture.

Without using data augmentation and dynamic mixing, on WSJ0-2mix we obtain 23.4 dB SI-SDRi, which significantly surpasses the previous best (at 22.1 dB) attained by time-domain approaches [24] and is slightly better than a theoretical SI-SDRi upper bound (at 23.1 dB) suggested in [39] for models with non-overlapping analysis filterbanks. Both of these indicate the strong potential of complex T-F domain approaches also for anechoic speaker separation. The code of TF-GridNet has been released in the ESPnet-SE++ toolkit [40].

2. PROPOSED ALGORITHMS

Given a C-speaker mixture recorded in anechoic conditions, the physical model in the time domain can be formulated as $y[n] = \sum_{c=1}^{C} s^{(c)}[n]$, where y denotes the mixture and $s^{(c)}$ source c, and n indexes N time samples. In the short-time Fourier transform (STFT) domain, the physical model is formulated as $Y(t,f) = \sum_{c=1}^{C} S^{(c)}(t,f)$, where Y and $S^{(c)}$ respectively denote the complex spectra of y and $s^{(c)}$, t indexes T frames, and f indexes F frequencies. C is assumed known in this study. Our goal is to recover each source $s^{(c)}$ based on y. An overview of the proposed system is provided in Fig. 1. This section describes complex spectral mapping, DNN architectures, and loss functions.

2.1. Complex Spectral Mapping

Our DNNs are trained to perform complex spectral mapping [29–35, 37], where the RI components of Y are concatenated as input features to predict the RI components of each speaker $S^{(c)}$. The loss function, described later in Section 2.3, is defined based on the re-synthesized time-domain signals of the predicted RI components. Our system is non-causal. We normalize the sample variance of the time-domain mixture to one and scale each clean source using the same scaling factor during training.

2.2. TF-GridNet

Fig. 1 shows the proposed system and Table 2 summarizes the notations of the hyper-parameters in TF-GridNet. Given an input tensor with shape $2 \times T \times F$, where 2 is because we stack RI components, we first use a 2D convolution (Conv2D) with a 3×3 kernel

Table 2: Summary of model hyper-parameters.

Symbols	Description
D	Embedding dimension for each T-F unit
B	Number of TF-GridNet blocks
I	Kernel size for Unfold and Deconv1D
J	Stride size for Unfold and Deconv1D
H	Number of hidden units in BLSTMs in each direction
E	Number of output channels in 1×1 Conv2D to obtain query and key tensors in self-attention module
L	Number of heads in self-attention

followed by global layer normalization (gLN) [11] to compute a D-dimensional embedding for each T-F unit, obtaining a $D \times T \times F$ tensor. Next, we feed the tensor to B TF-GridNet blocks, each containing an intra-frame spectral module, a sub-band temporal module, and a full-band self-attention module, to gradually leverage local and global spectro-temporal information to refine the T-F embeddings. At last, a 2D deconvolution (Deconv2D) with $2C$ output channels and a 3×3 kernel followed by linear activation is used to obtain the predicted RI components, and inverse STFT (iSTFT) is applied for signal re-synthesis. The rest of this section details the three modules in each TF-GridNet block.

2.2.1. *Intra-frame Spectral Module*

In the intra-frame spectral module, we view the input tensor $R_b \in \mathbb{R}^{D\times T\times F}$ to the b^{th} block as T separate sequences, each with length F, and use a one-layer bidirectional long short-term memory (BLSTM) to model the local spectral information within each frame. We first use the *torch.unfold* function [41] with kernel size I and stride size J to stack nearby embeddings at each step, after we zero-pad the frequency dimension to $F' = \lceil \frac{F-I}{J} \rceil \times J + I$:

$$\dot{R}_b = \left[\text{Unfold}(R_b[:,t,:]), \text{ for } t = 1,\dots,T\right] \in \mathbb{R}^{(I\times D)\times T\times (\frac{F'-I}{J}+1)}.$$

Note that J can be larger than one so that the sequence length and the amount of computation can be reduced. We then apply layer normalization (LN) along the channel dimension (i.e., the first dimension) of $\dot{R}_b$, and a one-layer BLSTM with H units in each direction is used to model the inter-frequency information within each frame:

$$\ddot{R}_b = \left[\text{BLSTM}\big(\text{LN}(\dot{R}_b)[:,t,:]\big), \text{ for } t = 1,\dots,T\right] \in \mathbb{R}^{2H\times T\times (\frac{F'-I}{J}+1)}.$$

After that, a 1D deconvolution (Deconv1D) layer with kernel size I, stride size J, input channel $2H$ and output channel D (and without subsequent normalization and non-linearity) is applied to the hidden embeddings of the BLSTM:

$$\dddot{R}_b = \left[\text{Deconv1D}(\ddot{R}_b[:,t,:]), \text{ for } t = 1,\dots,T\right] \in \mathbb{R}^{D\times T\times F'}.$$

After removing the zero paddings, we add this tensor to the input tensor via a residual connection to produce the output tensor: $U_b = \dddot{R}_b[:,:,:F] + R_b \in \mathbb{R}^{D\times T\times F}$.

2.2.2. *Sub-band Temporal Module*

In the sub-band temporal module, the procedure is almost the same as that in the intra-frame spectral module. The only difference is that the input tensor $U_b \in \mathbb{R}^{D\times T\times F}$ is viewed as F separate sequences, each with length T, and a BLSTM is used to model temporal information within each sub-band. The output tensor is denoted as $Z_b \in \mathbb{R}^{D\times T\times F}$.

2.2.3. *Full-band Self-attention Module*

In the full-band self-attention module (illustrated in Fig. 2), given Z_b produced by the sub-band temporal module, we first compute frame-level embeddings based on the T-F embeddings within each frame, and then use whole-sequence self-attention on these frame

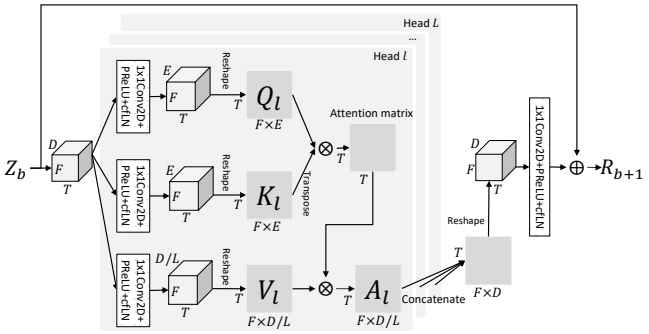

Fig. 2: Proposed full-band self-attention module.

embeddings to capture long-range global information. The motivation is that the intra-frame and sub-band BLSTMs can only model local information within each frame and each frequency, and a whole-sequence self-attention module enables each frame to attend to any frames of interest to exploit long-range information. Our self-attention module follows the attention mechanism proposed in [42, 43], which is used with U-Net for music source separation and speech denoising. The key differences include (1) we use multi-head attention instead of single-head attention; and (2) we use the attention mechanism with dual-path models for speaker separation.

In detail, the self-attention module has L heads, and, in each head l, we apply 1×1 Conv2D followed by PReLU, LN along the channel and frequency dimensions (denoted as cfLN), and reshape layers to respectively obtain the 2D query $Q_l \in \mathbb{R}^{T\times(F\times E)}$, key $K_l \in \mathbb{R}^{T\times(F\times E)}$ and value $V_l \in \mathbb{R}^{T\times(F\times D/L)}$ tensors. The Conv2D layers for obtaining the query and key tensors both have E output channels, leading to $F \times E$-dimensional query and key vectors at each frame after stacking the T-F embeddings within each frame, and similarly the Conv2D layer for computing the value tensor has D/L output channels, resulting in an $F \times D/L$-dimensional value vector at each frame. All the three 1×1 Conv2D layers has D input channels. Following [44], the attention output $A_l \in \mathbb{R}^{T\times(F\times D/L)}$ is computed as

$$A_l = \text{softmax}(\frac{Q_l K_l^{\mathsf{T}}}{\sqrt{F \times E}})V_l.$$

We then concatenate the attention outputs of all heads along the second dimension, reshape it back to $D \times T \times F$, apply 1×1 Conv2D with D input and D output channels followed by PReLU and cfLN to aggregate cross-head information, and add it to the input tensor Z_b via a residual connection to obtain the output tensor R_{b+1}, which is fed into the next TF-GridNet block.

This self-attention mechanism has two major advantages. First, it only introduces a negligible number of parameters through the Conv2D layers. Second, it operates at the frame level and the memory cost on attention matrices is $\mathcal{O}(B \times L \times T^2)$. In contrast, TF-PSNet [25] applies multi-head self-attention in each path-scanning module, and the memory cost on attention matrices is $\mathcal{O}\big(B \times L \times F \times T^2\big) + \mathcal{O}\big(B \times L \times T \times F^2\big)$, which is annoyingly high.

2.3. Loss Functions

Our models are trained with utterance-level PIT [2]. The loss function follows the SI-SDR loss [11, 38], but with two differences.

First, in the paper proposing the SI-SDR metric [38], there are two equivalent formulations of SI-SDR, one scaling the *target* to equalize its energy level with that of the estimate and the other instead scaling the *estimate*. The SI-SDR loss used in the seminal Conv-TasNet study [11] and almost all the follow-up studies employ the former formulation, while our study uses the latter, i.e.,

$$\mathcal{L}_{\text{SI-SDR-SE}} = -\sum_{c=1}^{C} 10\log_{10}\frac{\|s^{(c)}\|_2^2}{\|\hat{\alpha}^{(c)}\hat{s}^{(c)} - s^{(c)}\|_2^2}, \quad (1)$$

where $\hat{s}^{(c)}$ denotes the re-synthesized signal based on the predicted RI components for speaker c, $\hat{\alpha}^{(c)} = \text{argmin}_\alpha \|\alpha\hat{s}^{(c)} - s^{(c)}\|_2^2 = (\hat{s}^{(c)})^{\mathsf{T}}s^{(c)}/(\hat{s}^{(c)})^{\mathsf{T}}\hat{s}^{(c)}$, and the "SE" in $\mathcal{L}_{\text{SI-SDR-SE}}$ means "scaling estimate". We observe that this loss leads to faster convergence and very similar performance, compared with the former.

Second, the latter formulation motivates us to add a mixture-constraint (MC) loss between the mixture and the summation of the scaled estimated sources, defined, following [35], as

$$\mathcal{L}_{\text{SI-SDR-SE+MC}} = \mathcal{L}_{\text{SI-SDR-SE}} + \frac{1}{N}\Big\|\sum_{c=1}^{C}\hat{\alpha}^{(c)}\hat{s}^{(c)} - y\Big\|_1, \quad (2)$$

where the sample variance of y has been normalized to one beforehand. The second term results in better separation in our experiments. It is motivated by a trigonometric perspective [7] in source separation, which suggested that constraining the separated sources to sum up to the mixture can lead to better phase estimation. It should be noted that, at run time, $\sum_{c=1}^{C}\hat{\alpha}^{(c)}\hat{s}^{(c)}$ would not equal y. This distinguishes our loss from mixture consistency [45], which strictly enforces the separated sources to sum up to the mixture. Our loss is also different from another mixture consistency loss proposed in [46], where the DNN is trained for real-valued magnitude masking without any phase estimation and the task is target speaker extraction based speech recognition in meeting scenarios.

In Eq. (2), we do not add a weight between the two terms for two reasons. First, this can simplify the practical application of this loss function, as we can avoid a weight to tune. Second, in modern speaker separation systems [47], it is not uncommon for the SI-SDRi to surpass 10 dB, and when the sample variance of the input mixture has been normalized to one (which is the case in our study), the second term in our experiments has a very small scale (less than 0.01 when the models are converged). This way, the second term would not dominate the overall loss, since it has a smaller influence compared with the first term, which is directly related to the final separation performance.

3. EXPERIMENTAL SETUP

We validate the proposed algorithms on WSJ0-2mix [1], the most popular dataset to date to benchmark monaural talker-independent speaker separation algorithms in anechoic conditions. It contains 20,000 (∼30h), 5,000 (∼10h), and 3,000 (∼5h) two-speaker mixtures respectively in its training, validation and test sets. The clean utterances are sampled from the WSJ0 corpus. The speakers in the training and validation sets are not overlapped with the speakers in the test set. The two utterances in each mixture are fully overlapped, and their relative energy level is sampled uniformly from the range $[-5, 5]$ dB. The sampling rate is 8 kHz.

The STFT window size is 32 ms and hop size 8 ms. The square-root Hann window is used as the analysis window. A 256-point discrete Fourier transform is applied to extract 129-dimensional complex spectra at each frame. We use $B = 6$ blocks and E is set to 4 for 8 kHz (see Table 2 for their definitions). This way, the dimension of frame-level embeddings (i.e., $F \times E$) used for self-attention is reasonable. In each epoch, we sample a 4-second segment from each mixture for model training. Adam is used as the optimizer. The norm for gradient clipping is set to 1. The learning rate starts from 0.001 and is halved if the validation loss does not improve in 3

Table 3: Comparison between masking and mapping based on WSJ0-2mix.

Row	Systems	Masking or Mapping?	#params (M)	SI-SDRi (dB)
1	DPRNN [15]	Masking learned embeddings	2.6	18.8
2	TFPSNet (BLSTM) [25]	Masking learned embeddings	2.6	19.7
3	TF-GridNet	Masking learned embeddings	2.8	20.7
4	TF-GridNet	Complex ratio masking	2.6	20.8
5	TF-GridNet	Complex spectral mapping	2.6	**21.2**

Table 4: Ablation SI-SDRi (dB) results on WSJ0-2mix.

Row	Systems	Use attention?	L	D	I	J	H	#params (M)	Loss	SI-SDRi
1	TF-GridNet	✗	-	64	1	1	128	2.6	(1)	21.2
2	TF-GridNet	✗	-	16	4	1	128	2.6	(1)	20.5
3	TF-GridNet	✗	-	128	1	1	128	3.6	(1)	21.6
4	TF-GridNet	✗	-	16	8	1	128	3.6	(1)	21.6
5	TF-GridNet	✗	-	16	8	1	128	3.6	(2)	21.8
6	TF-GridNet	✗	-	16	8	1	192	6.5	(2)	21.9
7	TF-GridNet	✗	-	24	8	1	192	8.0	(2)	22.5
8	TF-GridNet	✓	1	24	8	1	192	8.0	(2)	22.6
9	TF-GridNet	✓	4	24	8	1	192	8.0	(2)	22.9
10	TF-GridNet	✓	4	32	8	1	256	14.4	(2)	**23.4**

epochs. SI-SDRi [38] and SDRi [48] are used as the evaluation metrics, following previous studies. The mixture SI-SDR is 0 dB and the mixture SDR 0.2 dB. The number of model parameters is reported in million (M).

4. EVALUATION RESULTS

Table 3 compares the results of TF-GridNet with DPRNN [15] and TFPSNet [25]. All the models are trained with the loss in (1). Each model has almost the same number of parameters and uses almost the same amount of computation. This is realized by using BLSTMs in all the models and unifying the embedding dimension (or the bottleneck dimension in the cases of DPRNN and TFPSNet, which perform masking in the embedded space) to 64 and the hidden dimension of BLSTMs to 128. For DPRNN, the window size is set to 2 samples, hop size to 1 sample, chunk size to 250 frames, and overlap between consecutive chunks to 50%, following the best configuration reported in [15]. For TF-GridNet, we remove the full-band self-attention module in each block and set both I and J to 1 for a fair comparison. From row 1, 2 and 5, we can see that TF-GridNet with complex spectral mapping obtains better results. Table 3 also provides the results of using TF-GridNet with masking in row 3 and 4. In row 3, we perform masking in the learned embedded space, following [11, 15, 25]. We employ the same encoder-separator-decoder modules used in [25], but replace their path-scanning modules with our intra-frame spectral modules and sub-band temporal modules. In row 4, we use TF-GridNet for complex ratio masking [8, 29]. After obtaining the output tensor of the Deconv2D module (see Fig. 1), we truncate the values in the tensor to the range $[-5, 5]$ to obtain a complex mask, and then multiply it with the mixture spectrogram for separation. From row 3, 4 and 5, we observe that using complex spectral mapping is better.

Table 4 presents the SI-SDRi results of our models on WSJ0-2mix using different model configurations. From row 1-4, we observe that, when the kernel size is sufficiently large (i.e., $I = 8$), using the Unfold and Deconv1D mechanism together with a smaller embedding dimension (i.e., $D = 16$) does not degrade the performance, compared with the configuration that uses a larger embedding dimension (i.e., $D = 128$) but does not stack nearby T-F embeddings (i.e., $I = 1$). One key benefit of using the former setup is that the memory consumption is noticeably lower. From row 4 and 5, we notice that the MC loss produces slight improvement from 21.6 to 21.8 dB. From row 5-7, we can see that enlarging the model size by increasing the embedding dimension D and the number of hidden units H in BLSTMs produces clear improvement. The results in row 7, 8 and 9 indicate that the self-attention module is helpful (22.9 and 22.6 vs. 22.5 dB), and using four attention heads is better than using just one (22.9 vs. 22.6 dB). Further enlarging the model size in row 10 pushes up the SI-SDRi to 23.4 dB.

Table 1 compares the performance of our system with others' based on WSJ0-2mix. Our model has a modest number of parameters compared with previous best models such as SepFormer [19], SFSRNet [23] and QPDN [24]. One notable observation is that, compared with time-domain models, T-F domain models, since 2019, have been largely under-explored and under-represented for speaker separation in anechoic conditions, and the research community has been focused on developing time-domain models. The recent TFPSNet study [25] reports a strong performance at 21.1 dB SI-SDRi, but the performance falls within the range of scores (i.e., $[20.0, 22.1]$ dB SI-SDRi) that can be commonly reached by modern time-domain models. Our study, for the first time since 2019, unveils that complex T-F domain models, with a contemporary DNN architecture, can surpass the performance of modern time-domain models by a large margin.

5. CONCLUSIONS

We have proposed TF-GridNet, a novel model operating in the complex T-F domain for monaural speaker separation. By combining it with complex spectral mapping and a time-domain loss with a mixture constraint, we obtain state-of-the-art 23.4 dB SI-SDRi on WSJ0-2mix without using dynamic mixing. We believe that our study will generate broad impact, as it shows the strong performance of T-F domain models even for anechoic speaker separation. There are many follow-up directions to investigate, such as leveraging more advanced DNN architectures, and extensions to noisy-reverberant speech separation and to multi-channel conditions.

In closing, we emphasize that, through years of efforts, the performance on WSJ0-2mix has been largely saturated [39], and the separated signals do not sound too much different after the SI-SDRi surpassed, based on our informal listening tests, 22.0 dB. Nonetheless, the findings in this study suggest that, at a minimum, T-F domain methods modeling complex representations, which implicitly perform phase estimation by simultaneously predicting target RI components, are not sub-optimal compared with time-domain approaches for the task of anechoic speaker separation. The performance gap between these two approaches in earlier studies could be mainly due to differences in DNN architectures, rather than because of the use of over-complete learned filterbanks with very short window length.

6. ACKNOWLEDGMENTS

We would like to thank Dr. Wangyou Zhang at SJTU for generously sharing his reproduced code of TFPSNet [25].

7. REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep Clustering: Discriminative Embeddings for Segmentation and Separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[2] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-Talker Speech Separation with Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[3] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe *et al.*, "Single-Channel Multi-Speaker Separation using Deep Clustering," in *Proc. Interspeech*, 2016, pp. 545–549.

[4] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative Objective Functions for Deep Clustering," in *Proc. ICASSP*, 2018, pp. 686–690.

[5] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, "End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction," in *Proc. Interspeech*, 2018, pp. 2708–2712.

[6] Z.-Q. Wang and D. Wang, "Combining Spectral and Spatial Features for Deep Learning Based Blind Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 457–468, 2019.

[7] Z.-Q. Wang, K. Tan, and D. Wang, "Deep Learning Based Phase Reconstruction for Speaker Separation: A Trigonometric Perspective," in *Proc. ICASSP*, 2019, pp. 71–75.

[8] Y. Liu and D. Wang, "Divide and Conquer: A Deep CASA Approach to Talker-Independent Monaural Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, 2019.

[9] Y. Luo and N. Mesgarani, "TasNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," in *Proc. ICASSP*, nov 2017, pp. 697–700.

[10] ——, "Real-Time Single-Channel Dereverberation and Separation with Time-Domain Audio Separation Network," in *Proc. Interspeech*, 2018, pp. 342–346.

[11] ——, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] M. W. Lam, J. Wang, D. Su, and D. Yu, "Mixup-Breakdown: A Consistency Training Method for Improving Generalization of Speech Separation Models," in *Proc. ICASSP*, 2020, pp. 6374–6378.

[13] L. Zhang, Z. Shi, J. Han, A. Shi *et al.*, "FurcaNeXt: End-to-End Monaural Speech Separation with Dynamic Gated Dilated Temporal Convolutional Networks," in *Proc. ICMM*, 2020, pp. 653–665.

[14] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -RF: Efficient Networks for Universal Audio Source Separation," in *Proc. MLSP*, 2020.

[15] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *Proc. ICASSP*, 2020, pp. 46–50.

[16] E. Nachmani, Y. Adi, and L. Wolf, "Voice Separation with An Unknown Number of Multiple Speakers," in *ICML*, 2020, pp. 7121–7132.

[17] J. Chen, Q. Mao, and D. Liu, "Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation," in *Proc. Interspeech*, 2020, pp. 2642–2646.

[18] Y. Zhu, X. Zheng, X. Wu, W. Liu *et al.*, "DPTCN-ATPP: Multi-Scale End-To-end Modeling for Single-Channel Speech Separation," in *Proc. ICCIS*, 2021, pp. 39–44.

[19] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi *et al.*, "Attention Is All You Need In Speech Separation," in *Proc. ICASSP*, 2021, pp. 21–25.

[20] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "Sandglasset: A Light Multi-Granularity Self-Attentive Network for Time-Domain Speech Separation," in *Proc. ICASSP*, 2021, pp. 5759–5763.

[21] N. Zeghidour and D. Grangier, "Wavesplit: End-to-End Speech Separation by Speaker Clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

[22] S. Qian, L. Gao, H. Jia, and Q. Mao, "Efficient Monaural Speech Separation With Multiscale Time-Delay Sampling," in *Proc. ICASSP*, vol. 2022-May, 2022, pp. 6847–6851.

[23] J. Rixen and M. Renz, "SFSRNet: Super-Resolution for Single-Channel Audio Source Separation," in *Proceedings of AAAI*, 2022.

[24] ——, "QDPN - Quasi-Dual-Path Network for Single-Channel Speech Separation," in *Proc. Interspeech*, 2022, pp. 5353–5357.

[25] L. Yang, W. Liu, and W. Wang, "TFPSNet: Time-Frequency Domain Path Scanning Network for Speech Separation," in *Proc. ICASSP*, 2022, pp. 6842–6846.

[26] F. Dang, H. Chen, and P. Zhang, "DPT-FSNet: Dual-Path Transformer Based Full-Band and Sub-Band Fusion Network for Speech Enhancement," in *Proc. ICASSP*, 2022, pp. 6857–6861.

[27] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement," in *Proc. Interspeech*, vol. 2, 2021, pp. 821–825.

[28] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-Independent Speech Separation with Deep Attractor Network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 787–796, 2018.

[29] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pp. 483–492, 2016.

[30] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex Spectrogram Enhancement By Convolutional Neural Network with Multi-Metrics Learning," in *Proc. MLSP*, 2017, pp. 1–6.

[31] K. Tan and D. Wang, "Learning Complex Spectral Mapping With Gated Convolutional Recurrent Networks for Monaural Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.

[32] Z.-Q. Wang and D. Wang, "Deep Learning Based Target Cancellation for Speech Dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, 2020.

[33] Z.-Q. Wang, G. Wichern, and J. Le Roux, "Leveraging Low-Distortion Target Estimates for Improved Speech Enhancement," *arXiv preprint arXiv:2110.00570*, 2021.

[34] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-Microphone Complex Spectral Mapping for Utterance-Wise and Continuous Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2001–2014, 2021.

[35] Z.-Q. Wang, G. Wichern, and J. Le Roux, "Convolutive Prediction for Monaural Speech Dereverberation and Noisy-Reverberant Speaker Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3476–3490, 2021.

[36] ——, "On The Compensation Between Magnitude and Phase in Speech Separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 2018–2022, 2021.

[37] K. Tan, Z.-Q. Wang, and D. Wang, "Neural Spectrospatial Filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 605–621, 2022.

[38] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-Baked or Well Done?" in *Proc. ICASSP*, 2019, pp. 626–630.

[39] S. Lutati, E. Nachmani, and L. Wolf, "Sepit approaching a single channel speech separation bound," *arXiv preprint arXiv:2205.11801*, 2022.

[40] Y.-J. Lu, X. Chang, C. Li, W. Zhang *et al.*, "ESPnet-SE++: Speech Enhancement for Robust Speech Recognition,Translation, and Understanding," in *Proc. Interspeech*, 2022.

[41] A. Paszke, S. Gross, F. Massa, A. Lerer *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Proc. NIPS*, vol. 32, 2019.

[42] Y. Liu, B. Thoshkahna, A. Milani, and T. Kristjansson, "Voice and Accompaniment Separation in Music using Self-Attention Convolutional Neural Network," in *arXiv preprint arXiv:2003.08954*, 2020.

[43] A. Pandey and D. Wang, "Dense CNN with Self-Attention for Time-Domain Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1270–1279, 2021.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit *et al.*, "Attention Is All You Need," in *Proc. NIPS*, 2017.

[45] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe *et al.*, "Differentiable Consistency Constraints for Improved Deep Speech Enhancement," in *Proc. ICASSP*, 2019, pp. 900–904.

[46] K. Zmolikova, M. Delcroix, D. Raj, S. Watanabe *et al.*, "Auxiliary Loss Function for Target Speech Extraction and Recognition with Weak Supervision Based on Speaker Characteristics," in *Proc. Interspeech*, vol. 6, 2021, pp. 4156–4160.

[47] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, 2018.

[48] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.