DASFORMER: DEEP ALTERNATING SPECTROGRAM TRANSFORMER FOR MULTI/SINGLE-CHANNEL SPEECH SEPARATION

Shuo Wang^{1,2*}*, Xiangyu Kong*^{2}*, Xiulian Peng*^{2}*, Hesam Movassagh*^{3}*, Vinod Prakash*^{3}*, Yan Lu*^{2}

[1]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]Microsoft Research Asia, Beijing, China
[3]Microsoft Corporation, Redmond WA 98052, USA

ABSTRACT

For the task of speech separation, previous study usually treats multi-channel and single-channel scenarios as two research tracks with specialized solutions developed respectively. Instead, we propose a simple and unified architecture - DasFormer (**D**eep **a**lternating **s**pectrogram trans**Former**) to handle both of them in the challenging reverberant environments. Unlike frame-wise sequence modeling, each TF-bin in the spectrogram is assigned with an embedding encoding spectral and spatial information. With such input, DasFormer is then formed by multiple repetition of simple blocks each of which integrates *1) two multi-head self-attention (MHSA) modules alternately processing within each frequency bin & temporal frame of the spectrogram 2) MBConv before each MHSA for modeling local features on the spectrogram*. Experiments show that DasFormer has a powerful ability to model the time-frequency representation, whose performance far exceeds the current SOTA models in multi-channel speech separation, and also achieves single-channel SOTA in the more challenging yet realistic reverberation scenario.

***Index Terms*—** multi-channel speech separation, single-channel speech separation, multi-head self-attention

1. INTRODUCTION

Deep neural networks (DNNs) based speech separation systems have received widespread attention since Deep Clustering (DC) [1] and Permutation Invariant Training (PIT) [2] were proposed. When multi-microphone data are available, many approaches merging spatial cues with deep models successively achieve SOTA results on multi-channel separation tasks[3, 4, 5]. However, the majority of the methods are specially designed for each scenario where cross-generalization usually becomes suboptimal. By contrast, in this paper we explore a network backbone capable of handling both single/multi-channel speech separation well.

Early studies on speech separation mainly rely on the sparsity of speech in the time-frequency (TF) domain. This process can be performed by assigning speaker label to each TF-bin [1], either estimating a ratio mask and product with the original input, or directly estimating the complex coefficients [6]. Later, TasNet [7] and its variants [8, 9, 10] have grown to be dominant for speech separation under anechoic conditions. This often attributes to a learnable analytic basis instead of the fixed Fourier basis [7] and a more powerful network structure like Transformer [10, 11, 12]. These methods have achieved satisfactory performance in anechoic environments, while degradation is evident when room reverberation is not neglected [13, 14].

When multi-microphone data are available, the task falls into multi-channel speech separation. A remarkably successful solution is to combine an optimal beamformer with a neural network (NN) like TasNet, such as the spatial covariance matrices (SCMs) [15], and the steering (eigen) vector [16]. Subsequently, this cascade design was extended to the iterative refinement framework with significantly improved performance [5]. Another common approach is end-to-end network design, which attempts to incorporate multi-channel cues such as Inter-channel Phase Difference (IPD) as additional features into the input of the network [3]. Recently, narrow-band conformer (NBC) [4] handles speech separation with impressive gains in a narrow-band mode where all subbands share the same parameters. Note that valuable spatial information even exists for single-channel separation due to the presence of room reflections and reverberations [17].

As far as the authors view, few efforts have been made on a common backbone network handling both tasks well. For instance, Sepformer[11], which performs better than TasNet in single-channel, does not work as well as BeamTasNet when multiple microphones are available, and degrades severely in reverberant environments. Similarly, multi-channel methods like NBC [4] surely handle reverberation cases, but they perform even worse than majority of single-channel systems when less channels are available. Besides, some multi-channel architectures are customized, which becomes incompatible with single-channel case.

In this paper, we argue that with proper modeling and deeper backbone networks, embedding with minimum units

[*] This work was done at Microsoft Research Asia.

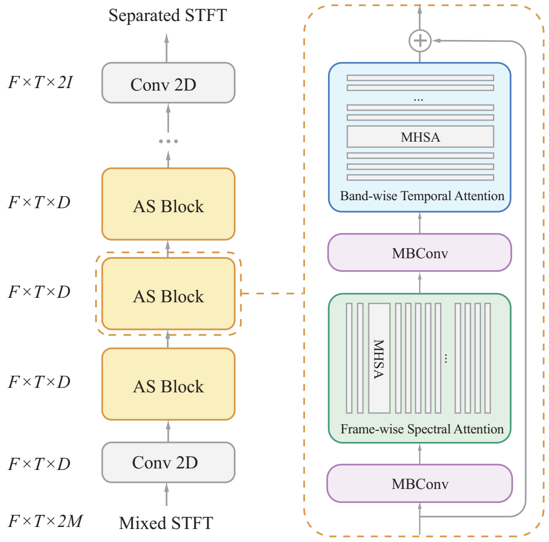

Fig. 1: The architecture of the proposed DasFormer.

in TF-bins can encode enough information to separate speakers. Such embedding includes the dependencies of one TF unit with its surroundings and even distant area, and the spatial information introduced in the reverberant environment or/and multi-channel sampling. Specifically, spectrogram-like features are then alternately fed into a block including two convolution modules, a spectral attention module and a temporal attention module. In this way, we can facilitate the aggregation of stable local features, spectral features and spatial features including direct path and reverberation respectively. Furthermore, by repeating such processing block, a deeper network - DasFormer allows each TF embedding to gather global clues from distant frames/sub-bands.

Experiments show that DasFormer achieves scale-invariant signal-to-distortion ratio improvement (SI-SDRi) far exceeding SOTA both on the spatialized WSJ0-2mix dataset (multi-channel) and a challenging single-channel dataset WHAMR! with room reverberation.

2. THE PROPOSED APPROACH

2.1. Formulation

Assuming that I speakers are recorded by M microphones in a room, the received signal of m-th microphone in frequency domain $Y_m(t,f)$ can be represented as

$$Y_m(t,f) = \sum_{i=0}^{I-1} X_{i,m}(t,f) \qquad (1)$$

$$= \sum_{i=0}^{I-1}\sum_{l=0}^{L-1} S_i(t-l,f)H_{i,m}(l,f), \qquad (2)$$

where $X_{i,m}(t,f)$ denotes the contribution of i-th speech component $S_i(t,f)$ to the m-th microphone, and $H_{i,m}(l,f)$ models the convolution process due to room reflections.

2.2. DasFormer

Within a single frame, the spectral regions occupied by one speaker have inter-dependencies between these sub-bands, such as harmonic trains, despite their overlap [18]. We employ the multi-head self-attention (MHSA) to model the sequence globally to seek such dependencies and aggregate the components from the same speaker. This is similar to [12] which also used sequence modeling on TF domain.

Within a sub-band, frames dominated by the same speaker share a consistent convolution process (as in Eq.(2)), which would be stronger when considering the phase difference introduced by the direct path when multiple channels are available. So MHSA is also applied to temporal sequences. As illustrated by Fig.1, we propose the Alternating Spectrogram Block (AS Block) including two MBConvs, a frame-wise temporal attention module and a band-wise spectral attention module as the basic processing unit. Multiple AS Blocks are then repetitively stacked together as the deep alternating spectrogram transformer (DasFormer). Such deep repetition and the alternating processing pipeline proves to be crucial for obtaining better separation results. Specific design of each component is as followed.

2.2.1. *Frame-wise Spectral Attention (FSA)*

Denoting the output of the encoder as TF embeddings, $\mathbf{e}(t,f)$. The sequence $\mathbf{e}(t,\cdot)$ consisting of all sub-bands in frame t is fed into a MHSA module, which expressed as

$$\mathbf{e}(t,\cdot) \leftarrow \mathbf{e}(t,\cdot) + Dropout(MHSA(LN(\mathbf{e}(t,\cdot)))) \qquad (3)$$

where LN denotes the Layer Normalization. The above is repeated for all frames and the same MHSA module is shared.

2.2.2. *Band-wise Temporal Attention (BTA)*

Similar to the FSA, the sequence consisting of all frames at the f-th sub-band passes through an MHSA, and is denoted as

$$\mathbf{e}(\cdot,f) \leftarrow \mathbf{e}(\cdot,f) + Dropout(MHSA(LN(\mathbf{e}(\cdot,f)))) \qquad (4)$$

All sub-bands repeated and share the same MHSA module.

2.2.3. *MBConv*

Before each Attention module, we add a 3×3 MBConv block with Squeeze-Excitation (SE) module [19]. This is similar to the convolution augmented (FFN) in Conformer [20] and NBC [4], except that 2D convolution is used here, as follows:

$$\mathbf{e}(\cdot,\cdot) \leftarrow \mathbf{e}(\cdot,\cdot) + Pw_2(SE(Dw(Pw_1(BN(\mathbf{e}(\cdot,\cdot)))))) \qquad (5)$$

where Pw_1 and Pw_2 are both Point-wise Conv2D, which implement the expansion and shrinkage projections with factor set to 4, respectively. The BN denotes Batch Normalization and DW denotes 3×3 Depth-wise Conv2D [19].

3. EXPERIMENT

3.1. Setup

We evaluate DasFormer on both multi/single-channel tasks.

Multi-channel dataset. The commonly used dataset - spatialized WSJ0-2MIX is selected [3]. All utterances are segmented into 4-second lengths and convolved with randomly generated room impulse responses (RIR). The 28,000 RIRs are generated using same parameters with [3], including room size, reverberation time (T60), speaker location, and array geometry. These clips are then mixed in a fully overlapped manner according to [1]. To align with [15, 5], the first microphone is used as reference and only the first 4 of the 8 microphones are used as inputs. The sampling rate is 8 kHz. To compare with [4], we added another settings adopted in [4]. The main differences include, 8 microphone arrays with fixed geometry, larger T60, 16 kHz sampling rate. In two above settings, we focus on pure separation by using reverberated speech as the training target.

Single-channel dataset. This is actually a special case when $M = 1$, e.g. using only one microphone data. For comparison with reported results, we choose a widely used and more challenging single-channel dataset - WHAMR! [13], whose goal is to predict each speaker's clean signal from reverberant and noisy input and hence is closer to the real-world scenario.

Model implementations. The MHSA module in FSA and BTA modules both applies a dimension $D = 64$ and number of heads $H = 4$. The parameters of the two MBConv modules are: the kernel size of DW Conv is 3×3, the number of channels $D = 64$, the expansion factor is 4, and the shrinkage factor in the SE module is 0.25. The AS Block repeats $L = 12$ layers. The initial encoder is a 3×3 Conv2D with $2M$ input channels and D output channels. The final decoder is also a 3×3 Conv2D with D input channels and $2I$ output channels. The model architecture is consistent by default on both tasks, except for the number of input channels. A frame length of 32 ms with 16 ms frame shift is used in STFT.

Training configurations. An Adam optimizer is used with an initial learning rate of 0.001. The signal metric SI-SDR is used as the loss function. The learning rate halved when no lower SI-SDR is found for 7 consecutive epochs. When no lower metric is found for 15 consecutive epochs, training stops. Gradient clipping with a maximum norm of 5 is used to avoid gradient explosion.

3.2. Performance comparison

Multi-channel speech separation We use SDR improvement (SDRi) and narrow-band PESQ to evaluate DasFormer and each baseline system. Our proposed DasFormer achieves an SDR improvement of 25.9 dB with the RIR setting in [3], which is a 4.4 dB improvement compared to BeamGuided-TasNet, an approach with an iterative refinement framework combined with spatial filters and single-channel separation in the time domain. The DasFormer also obtained an SDR improvement of 20.5 dB under the RIR setting in [4], a 5.2 dB improvement over the NBC approach with narrow-band complex mapping. We also notice that the NBC model working in narrowband mode degrades significantly in the task with random array geometry, while DasFormer achieves a higher SDR improvement under this challenging scenario instead. Considering the diversity and phase inconsistency of arrays in practice, this random array geometry setting is a more robust and adaptive way to training models.

Model	Params.	PESQ	SDRi
	(M)		(dB)
RIR settings as [15, 5]			
Mixture	–	1.80	0.0
FaSNet-TAC [21]	2.8	2.90	11.7
NBC [4]	2.0	2.95	13.3
Beam-TasNet [15]	–	–	16.8
BeamGuided-TasNet [5]	5.4	–	21.5
DasFormer (ours)	2.2	**4.33**	**25.9**
RIR settings as [4]			
Mixture	–	1.80	0.0
NBC [4]	2.0	3.53	15.3
DasFormer (ours)	2.2	**4.11**	**20.5**

Table 1: Experiment results on spatialized WSJ0-2Mix.

Model	Params.	SI-SDRi	SDRi
	(M)	(dB)	(dB)
TasNet-BLSTM[13]	23.6	9.2	–
Conv-TasNet [13]	8.8	8.3	–
DPRNN [9]	2.6	10.3	–
DPTNET [10]	2.7	12.1	11.1
Sepformer [11]	26	11.4	–
WaveSplit [22]	29	12.0	11.1
WaveSplit (DM) [22]	29	13.2	12.2
Sudo rm -rf (U=16) [14]	6.3	12.1	–
Sudo rm -rf (U=36) [14]	26.6	13.5	–
DasFormer	2.2	16.0	14.6
DasFormer Plus	6.4	**17.3**	**15.7**

Table 2: Experiment results on WHAMR!.

Single-channel separation with reverberation DasFormer and its plus version (D=96, L=16) achieved SI-SDR improvements of 16.0 dB and 17.3 dB, respectively, which is a 3.8 dB performance improvement compared to the SOTA model on WHAMR! dataset [14].

3.3. Model scalability on microphone number

We compare DasFormer and current mainstream models on various microphone numbers. The WHAMR! in Fig. 2 can be viewed as a more difficult single microphone setting with noise added. For aligning, we extend Sepformer by increasing the input channels of encoder. However, from the limited

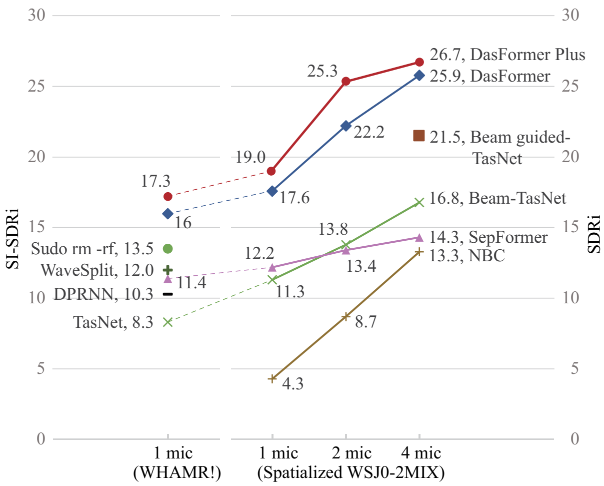

Fig. 2: Results on different microphone numbers. All results are on spatialized WSJ0-2MIX with SI-SDRi as metric except for the first column which is on WHAMR! with SDRi metric.

Layers (L)	Dims (D)	Conv	Params. (M)	SDRi (dB)
12	64	-	2.2	**25.9**
8	80	-	2.2	24.1
4	128	-	2.2	23.4
12	64	w/o SE	1.4	24.3
12	64	1×1, w/o SE	1.3	22.2

Table 3: Results on ablation study. '-' means that the MB-Conv module is not changed. 'w/o SE' means remove the SE module. '1×1' means that the 3×3 Conv2D in the MBConv module is replaced with a point-wise Conv2D.

performance gain after increasing channels, it is obvious that the spatial cues are less effectively utilized by SepFormer. And Beam-TasNet is far inferior to other SOTA algorithms when degraded to single-channel (when MVDR is not available). The NBC method is severely degraded, which can be explained by the fact that its narrow-band mode relies heavily on spatial information, which is susceptible to channel reduction. The proposed DasFormer not only achieves the highest performance on both tasks, but also increases the performance significantly as the microphone number increases.

3.4. Ablation Study

To investigate the contribution of deeper layers and the MB-Conv module, we conducted two sets of ablation experiments (all trained and tested on the spatialized WSJ0-2Mix dataset [3]). The first one gradually makes the DasFormer shallower and keeps the model size constant by increasing the embedding dimension. It shows that deeper networks are easier to achieve higher performance for DasFomer. The second one is to remove the SE module and the 3×3 Conv from MBConv consecutively. The 3×3 Conv is replaced by a point-wise Conv, becoming similar to the FFN module in Transformer. It can be found that both of them bring significant degradation. And the 3×3 Conv2D is more crucial for better performance.

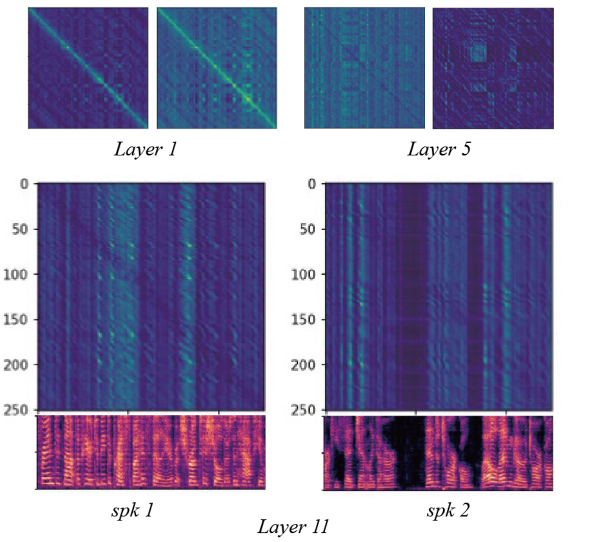

Fig. 3: Results of attention scores from different layers in AS Block. They are two heads from the shallow (layer 1), middle (layer 5) and deep (layer 11) layer respectively.

3.5. Visualization

To further understand the behavior of the model, we visualize the attention map. We observe that different layers present different temporal patterns. The shallow layer pays more attention on local scope formed by neighboring frames, as evidenced by higher scores around the diagonal. The middle layer reflects more block-like structure, which may attribute to the similarity of frames within the same phonetic unit. And the heads from deeper layers show distributions with clear correspondence to speaker activity. So as the network becomes deeper, the DasFormer tends to encourage different layers to aggregate different levels of information.

4. CONCLUSION

This work attempt to employ a common architecture for both multi/single-channel speech separation. By alternately performing band-wise and frame-wise MHSA on TF-bin embedding and combining spatial information when multi-channel data are available, the proposed DasFormer can maintain robust separation results against less microphone number and even gets SOTA results on single-channel speech separation in the challenging reverberant environments. The proposed architecture is potential for more challenging scenarios, like adapting a model trained on one microphone array to another (different on both microphone number and array geometry).

5. REFERENCES

[1] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*. IEEE, 2016, pp. 31–35.

[2] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*. IEEE, 2017, pp. 241–245.

[3] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *ICASSP*. IEEE, 2018, pp. 1–5.

[4] Changsheng Quan and Xiaofei Li, "Multichannel speech separation with narrow-band conformer," in *Interspeech*, 2022, pp. 5378–5382.

[5] Hangting Chen, Yi Yang, Feng Dang, and Pengyuan Zhang, "Beam-Guided TasNet: An iterative speech separation framework with multi-Channel output," in *Interspeech*, 2022, pp. 866–870.

[6] Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust asr," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1778–1787, 2020.

[7] Yi Luo and Nima Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP*. IEEE, 2018, pp. 696–700.

[8] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[9] Yi Luo, Zhuo Chen, and Takuya Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP*. IEEE, 2020, pp. 46–50.

[10] Jingjing Chen, Qirong Mao, and Dong Liu, "Dual-Path Transformer Network: Direct context-aware modeling for end-to-end monaural speech separation," in *Interspeech*, 2020, pp. 2642–2646.

[11] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *ICASSP*. IEEE, 2021, pp. 21–25.

[12] Lei Yang, Wei Liu, and Weiqin Wang, "Tfpsnet: Time-frequency domain path scanning network for speech separation," in *ICASSP*. IEEE, 2022, pp. 6842–6846.

[13] Matthew Maciejewski, Gordon Wichern, Emmett McQuinn, and Jonathan Le Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *ICASSP*. IEEE, 2020, pp. 696–700.

[14] Efthymios Tzinis, Zhepei Wang, Xilin Jiang, and Paris Smaragdis, "Compute and memory efficient universal sound source separation," *Journal of Signal Processing Systems*, vol. 94, no. 2, pp. 245–259, 2022.

[15] Tsubasa Ochiai, Marc Delcroix, Rintaro Ikeshita, Keisuke Kinoshita, Tomohiro Nakatani, and Shoko Araki, "Beam-tasnet: Time-domain audio separation network meets frequency-domain beamformer," in *ICASSP*. IEEE, 2020, pp. 6384–6388.

[16] Zhuohuang Zhang, Takuya Yoshioka, Naoyuki Kanda, Zhuo Chen, Xiaofei Wang, Dongmei Wang, and Sefik Emre Eskimez, "All-neural beamformer for continuous speech separation," in *ICASSP*. IEEE, 2022, pp. 6032–6036.

[17] Katharine Patterson, Kevin Wilson, Scott Wisdom, and John R. Hershey, "Distance-Based sound separation," in *Interspeech*, 2022, pp. 901–905.

[18] Thomas W Parsons, "Separation of speech from interfering speech by means of harmonic selection," *The Journal of the Acoustical Society of America*, vol. 60, no. 4, pp. 911–918, 1976.

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[20] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020, pp. 5036–5040.

[21] Yi Luo, Zhuo Chen, Nima Mesgarani, and Takuya Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *ICASSP*. IEEE, 2020, pp. 6394–6398.

[22] Neil Zeghidour and David Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2840–2849, 2021.