SMoLnet-T: An Efficient Complex-spectral Mapping Speech Enhancement Approach with Frame-wise CNN and Spectral Combination Transformer for Drone Audition

Zhi-Wei Tan* and Andy W. H. Khong†
* Nanyang Technological University, School of Electrical and Electronic Engineering, Singapore
† Nanyang Technological University, School of Electrical and Electronic Engineering and Lee Kong Chian School of Medicine, Singapore
E-mail: {zhiwei.tan, andykhong}@ntu.edu.sg

***Abstract*—Speech enhancement for drone audition applications is challenging due to the low SNR with large spectra feature overlap and limited computing resources. We propose SMoLnet-T, a complex spectral mapping approach with frame-wise CNN and newly-formulated spectral combination transformers. SMoLnet-T incorporates dilated CNN to extract spectral maps of high-frequency resolution for its transformers. This allows it to focus on a higher level of abstraction and determine the combination of spectral maps is crucial for enhancement across a large temporal context. Experiment results with noise recorded from a hovering drone highlight the efficacy of SMoLnet-T over DPTNet with significantly lower computational requirements and speech distortion while achieving improved speech intelligibility under SNR < -23 dB.**

***Index Terms*—Convolution neural network, deep learning, transformer, drone audition, speech enhancement**

I. Introduction

Drone audition has profound applications in search and rescue, surveillance, and package delivery. These applications require enabling (acoustic) technologies such as source localization [1, 2], sound classification [3], noise suppression [4], and tracking [5]. Unlike conventional microphone array systems for indoor environments [6], speech enhancement for unmanned aerial vehicles (UAVs) is challenging due to the ego-noise generated during flight [7]. Operating under such an adverse low signal-to-noise ratio (SNR) environment limits the enhancement performance when the spectra feature of speech and drone noise overlap. Furthermore, the limited onboard computing resources necessitate the algorithm to be "light-weight".

Deep learning approaches for speech enhancement under low SNR conditions have been gaining attention in recent years [8–10]. One such approach involves the estimation of rotor noise power spectral density by exploiting UAV rotor characteristics and signals from the received microphones [8]. More recently, a small model on low SNR network (SMoLnet) has been proposed for single-channel drone noise reduction [9]. Employing exponentially dilated convolutions [11, 12], the convolutional neural network (CNN) within SMoLnet requires fewer learnable parameters while achieving significant noise reduction and improved speech intelligibility comparable to larger models. An independent study [13] that focuses on single-channel speech enhancement showed that the resource-efficient SMoLnet achieves reasonable speech quality (in terms of perceptual evaluation of speech quality (PESQ) [14] and scale-invariant signal-to-distortion ratio (SI-SDR) [15]) compared to the more extensive single-channel dual-path transformer network (DPTNet) [16] and the deep complex U-net (DCUNet) [17] at very low SNR. However, due to the shorter temporal context of SMoLnet, it suffers from lower speech intelligibility performance in terms of the extended short-time objective intelligibility (ESTOI) measure [18].

To leverage longer temporal context, recent deep learning approaches incorporate transformers [19]. In particular, unlike CNNs and recurrent neural networks (RNNs), the transformer architecture exploits significant temporal contexts more effectively via its self-attention mechanism and has found applications in speech processing models [16, 20]. Of particular interest is the time-domain DPTNet [16], which employs the intra- and inter-transformer for the exploitation of local and sub-global temporal information, respectively. It is also useful to note that separating the time-domain signal into local frames and global chunks reduces the computational complexity required to process all the frames.

As an extension, the time-domain Sepformer [21] alleviates the need for the RNN within DPTNet by employing multiple transformers with fixed positional encoding [19]. This facilitates modeling of the temporal positioning information resulting in a highly-parallelized model that does not require step-wise processing by RNN. The model, however,

This research was conducted under project WP6 within the Delta-NTU Corporate Lab with funding support from A*STAR under its IAF-ICP programme (Grant no: I2201E0013) and Delta Electronics Inc.

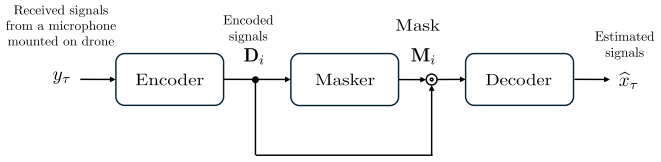

Fig. 1. Masking-based approaches for speech processing algorithms [16, 20, 21, 23]

requires a significantly larger number of learnable parameters to achieve similar performance to that of DPTNet. More recently, a magnitude-domain approach [20], where short-time Fourier transform (STFT) [22] features serve as input to the transformer, has been proposed. This approach performs similarly to the Sepformer with lower computational cost by achieving a trade-off between several global chunks and local frames. Apart from enhancing the magnitude (while leaving the phase unprocessed), the use of multiple transformers require considerable computational resources.

We propose the SMoLnet-T, consisting a frame-wise CNN with newly formulated spectral-attention transformers. The proposed approach is based on complex spectral mapping, which extracts high-resolution spectral feature maps through exponentially increasing frequency-dilated convolutions. These spectral maps, along with their associated frequency position information, are retained for the subsequent spectral combination transformers. In contrast to existing transformer-based approaches such as DPTNet that focuses on alternating between the local or sub-global information, SMoLnet-T focuses on a higher level of abstraction by determining which combination of high-resolution spectral maps generated by the frame-wise CNN is suitable for enhancement in the global aspect. This avoids the need for masking, which, in general, extracts features ineffectively.

II. Problem formulation and literature review

The received signal of a single-channel microphone mounted on a UAV can be expressed as

$$y_\tau = x_\tau + v_\tau, \tag{1}$$

where $x_\tau, v_\tau \in \mathbb{R}$ are the (clean) speech and noise signals, respectively, $1 \leq \tau \leq \mathcal{T}$ is the sample index with $\mathcal{T}$ being the number of samples. Due to the significant computation associated with the high-fidelity speech signal, transformer-based approaches [16, 20, 21], in general, focus on reducing the processing of large $\mathcal{T}$ samples. The general framework for the use of these transformer networks is depicted in Fig. 1.

The DPTNet [16] employs an encoder

$$\underline{\mathbf{D}}^{[1]} = \mathfrak{f}_{\mathsf{seg}_2}\Big(\mathfrak{f}_{\mathsf{relu}}\Big(\mathfrak{f}_{\mathsf{conv}}\big(\mathfrak{f}_{\mathsf{seg}_1}\left(\mathbf{y}\right)\big)\Big)\Big), \tag{2}$$

where $\mathbf{y} = [y_1, \ldots, y_\mathcal{T}]^\mathsf{T}$ with $(\cdot)^\mathsf{T}$ being the transpose operator. Here, $\mathfrak{f}_{\mathsf{seg}_1} : \mathbb{R}^{1\times\mathcal{T}} \to \mathbb{R}^{L\times I}$ segments the received signal into I overlapping vectors each with L samples. The convolution function $\mathfrak{f}_{\mathsf{conv}} : \mathbb{R}^{L\times I} \to \mathbb{R}^{N\times I}$ subsequently filters these samples with N number of filters each of length L. The function $\mathfrak{f}_{\mathsf{relu}}$ is the rectified linear unit (ReLU) [24] activation unit, and $\mathfrak{f}_{\mathsf{seg}_2} : \mathbb{R}^{N\times I} \to \mathbb{R}^{N\times K\times P}$ further segments the activated features to P overlapping chunks each of length K. It is important to note that $N < L \leq \mathcal{T}$.

With $\mathbf{D}_p^{[1]} \in \mathbb{R}^{N\times K}$ being the pth matrix of $\underline{\mathbf{D}}^{[1]}$, DPTNet employs B consecutive dual-path transformers such that [16]

$$\widetilde{\mathbf{D}}_p^{[b]} = \mathfrak{f}_{\mathsf{intra-t}}\left(\mathbf{D}_p^{[b]}\right), \tag{3}$$

$$\mathbf{D}_k^{[b+1]} = \mathfrak{f}_{\mathsf{inter-t}}\left(\widetilde{\mathbf{D}}_k^{[b]}\right), \tag{4}$$

where $1 \leq b \leq B$ is the index of the dual-path transformer, $\mathbf{D}_p^{[b]} \in \mathbb{R}^{N\times K}$ is the encoded signal, and $\widetilde{\mathbf{D}}_p^{[b]} \in \mathbb{R}^{N\times K}$ and $\widetilde{\mathbf{D}}_k^{[b]} \in \mathbb{R}^{N\times P}$ are the intermediate outputs. Here, $1 \leq p \leq P$ and $1 \leq k \leq K$, are, respectively, the local and global indices with P being the number of chunks and K being the length of chunks. In (4), $\mathbf{D}_k^{[b+1]}$ is the encoded signal for the subsequent dual-path transformers.

The variables $\mathfrak{f}_{\mathsf{intra-t}}^{[b]}$ and $\mathfrak{f}_{\mathsf{inter-t}}^{[b]}$ in (3) and (4) are the bth dual-path transformers, where each computes a self-attention mechanism given by

$$\mathbf{Q}_{i,h} = \mathbf{D}_i\mathbf{W}_h^{\mathsf{Q}}, \quad \mathbf{K}_{i,h} = \mathbf{D}_i\mathbf{W}_h^{\mathsf{K}}, \quad \mathbf{V}_{i,h} = \mathbf{D}_i\mathbf{W}_h^{\mathsf{V}}.$$

We have omitted b for clarity and that $h \in [1, H]$ denotes the head index with H being the number of heads. The variable

$$i = \begin{cases} p, & \text{for } \mathfrak{f}_{\mathsf{intra-t}}; \\ k, & \text{for } \mathfrak{f}_{\mathsf{inter-t}}, \end{cases} \tag{5}$$

denotes the embedding index. The query $\mathbf{Q}_{i,h}$, keys $\mathbf{K}_{i,h}$, and values $\mathbf{V}_{i,h}$ are, respectively, computed using weights

$$\mathbf{W}_h^{\mathsf{Q}}, \mathbf{W}_h^{\mathsf{K}}, \mathbf{W}_h^{\mathsf{V}} \in \begin{cases} \mathbb{R}^{K\times\frac{D}{H}}, & \text{for } \mathfrak{f}_{\mathsf{intra-t}}; \\ \mathbb{R}^{P\times\frac{D}{H}}, & \text{for } \mathfrak{f}_{\mathsf{inter-t}}. \end{cases} \tag{6}$$

Thereafter, the multi-head attention is achieved via

$$\mathbf{A}_i = \mathfrak{f}_{\mathsf{cat}}\Big([\mathbf{A}_{i,1}, \ldots, \mathbf{A}_{i,h}, \ldots, \mathbf{A}_{i,H}]\Big)\mathbf{W}^{\mathsf{O}}, \tag{7}$$

where $\mathfrak{f}_{\mathsf{cat}}$ concatenates the single-head attention $\mathbf{A}_{i,h} \in \mathbb{R}^{N\times\frac{D}{H}}$ along the second dimension and $\mathbf{W}^{\mathsf{O}} \in \mathbb{R}^{D\times D}$ denotes output weights to achieve $\mathbf{A}_i \in \mathbb{R}^{N\times D}$. Here,

$$\mathbf{A}_{i,h} = \mathfrak{f}_{\mathsf{softmax}}\left(\frac{\mathbf{Q}_{i,h}\mathbf{K}_{i,h}^\mathsf{T}}{\sqrt{D}}\right)\mathbf{V}_{i,h}. \tag{8}$$

With (7), the DPTNet employs the improved transformer [25] described by

$$\dot{\mathbf{D}}_i = \mathfrak{f}_{\mathsf{ln}_1}\left(\mathbf{A}_i + \mathbf{D}_i\right), \tag{9}$$

$$\ddot{\mathbf{D}}_i = \mathfrak{f}_{\mathsf{relu}}\left(\mathfrak{f}_{\mathsf{RNN}}\left(\dot{\mathbf{D}}_i\right)\right)\mathbf{W}_2 + \mathbf{b}_2, \tag{10}$$

$$\dddot{\mathbf{D}}_i = \mathfrak{f}_{\mathsf{ln}_2}\left(\dot{\mathbf{D}}_i + \ddot{\mathbf{D}}_i\right), \tag{11}$$

where $\mathfrak{f}_{\mathsf{ln}_1}$ and $\mathfrak{f}_{\mathsf{ln}_2}$ denote layer normalizations, $\mathbf{W}_2 \in \mathbb{R}^{D\times D}$ and $\mathbf{b}_2 \in \mathbb{R}^{1\times D}$ denote the weight and the bias, respectively. The variables $\dot{\mathbf{D}}_i$ and $\ddot{\mathbf{D}}_i$ are the intermediate outputs of the

improved transformer, and the output $\dddot{\mathbf{D}}_i$ is used as the input of (3) or (4) such that

$$\ddot{\mathbf{D}}_i = \begin{cases} \widetilde{\mathbf{D}}_i, & \text{if subsequent transformer is } \mathfrak{f}_{\mathsf{inter-t}}; \\ \mathbf{D}_i, & \text{if subsequent transformer is } \mathfrak{f}_{\mathsf{intra-t}}. \end{cases} \tag{12}$$

The function $\mathfrak{f}_{\mathsf{RNN}}$ in (10) is an RNN that learns the order information of the input. We note that the improved transformer differs from the transformer [19] in two ways—a feed-forward network is employed instead of $\mathfrak{f}_{\mathsf{RNN}}$ and that a fixed position encoding is added to $\mathbf{D}_i$.

If there are no subsequent transformer (i.e., $b = B$) in (12), the estimated signal is decoded using

$$\mathbf{M} = \mathfrak{f}_{\mathsf{ola}}\Big(\mathfrak{f}_{\mathsf{conv}}\left(\ddot{\mathbf{D}}_i\right)\Big), \tag{13}$$

$$\widehat{\mathbf{x}} = \mathfrak{f}_{\mathsf{deconv}}\left(\mathbf{D}_p^{[1]} \odot \mathbf{M}\right), \tag{14}$$

where $\mathbf{M}$ is the mask, and $\mathfrak{f}_{\mathsf{deconv}}$, $\mathfrak{f}_{\mathsf{ola}}$, and $\odot$ are transpose convolution, overlap-add, and Hadamard product operators. With reference to (8), it is important to note that the sequential information due to the attention mechanism is on the first dimension, and its computational requirement scales quadratically with this dimension. For DPTNet, this dimension has N elements determined by $\mathfrak{f}_{\mathsf{conv}}$.

In contrast to DPTNet, the Sepformer [21] alleviates the need of RNN in (10) by incorporating sinusoidal positional encoding [19] to $\mathbf{D}_p^{[b]}$ in (4) and $\widetilde{\mathbf{D}}_k^{[b]}$ in (3), respectively, via

$$\mathbf{D}_p^{'[b]} = \mathbf{D}_p^{[b]} + \mathbf{E}^{N\times K}, \qquad \widetilde{\mathbf{D}}_k^{'[b]} = \widetilde{\mathbf{D}}_k^{[b]} + \mathbf{E}^{N\times P}, \tag{15}$$

where the nth row and kth column of $\mathbf{E}^{N\times K}$ is expressed as

$$e_{n,k} = \begin{cases} \sin\left(\frac{k}{10000^{2n/N}}\right), & \text{if } n \text{ is even;} \\ \cos\left(\frac{k}{10000^{2n/N}}\right), & \text{if } n \text{ is odd,} \end{cases} \tag{16}$$

and similarly for $\mathbf{E}_{N\times P}$, i.e.,

$$e_{n,p} = \begin{cases} \sin\left(\frac{p}{10000^{2n/N}}\right), & \text{if } n \text{ is even;} \\ \cos\left(\frac{p}{10000^{2n/N}}\right), & \text{if } n \text{ is odd.} \end{cases} \tag{17}$$

Instead of employing a learned encoder-decoder in (2) and (14) for DPTNet and SepFormer, the magnitude-based SepFormer (Mag-SepFormer) [20] operates on the spectral-temporal domain by re-formulating (1) as

$$y_{t,f} = x_{t,f} + v_{t,f}, \tag{18}$$

where $1 \le t \le T$ is the time index with T being the number of time frames, while $1 \le f \le F$ is the frequency index with F being the number of the frequency bins. Defining $\mathbf{y}_f = [y_{1,f}, \ldots, y_{T,f}]^{\mathsf{T}}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_F] \in \mathbb{R}^{T\times F}$, the Mag-SepFormer encodes the received signals as

$$\mathbf{D}^{\mathsf{mag}} = \left|\mathbf{Y}^{\mathsf{T}}\right|, \tag{19}$$

where $|\cdot|$ is the element-wise magnitude operator. The estimated signal is then given by

$$\mathbf{M}^{\mathsf{mag}} = \mathfrak{f}_{\mathsf{relu}} \circ \mathfrak{f}_{\mathsf{conv-b}} \circ \mathfrak{f}_{\mathsf{ola}} \circ \mathfrak{f}_{\mathsf{conv}} \circ \mathfrak{f}_{\mathsf{p-relu}} \circ \mathfrak{f}_{\mathsf{sep}}\left(\mathbf{D}^{\mathsf{mag}}\right), \tag{20}$$

$$\mathbf{X}^{\mathsf{T}} = \mathbf{D}^{\mathsf{mag}} \odot \mathbf{M}^{\mathsf{mag}} \odot \exp\left(\jmath\angle\mathbf{Y}^{\mathsf{T}}\right), \tag{21}$$

where $\circ$, $\mathfrak{f}_{\mathsf{sep}}$, and $\mathfrak{f}_{\mathsf{p-relu}}$ are, respectively, the function composite, the SepFormer, and parametric ReLU [26] operations. The function $\mathfrak{f}_{\mathsf{c-branch}}(\cdot) = \mathfrak{f}_{\mathsf{sigmoid}}\big(\mathfrak{f}_{\mathsf{conv}}(\cdot)\big) + \mathfrak{f}_{\mathsf{tanh}}\big(\mathfrak{f}_{\mathsf{conv}}(\cdot)\big)$ with $\mathfrak{f}_{\mathsf{sigmoid}}$ and $\mathfrak{f}_{\mathsf{tanh}}$ being the sigmoid and hyperbolic tangent operations, respectively. Here, $\exp(\jmath\angle\mathbf{Y}^{\mathsf{T}})$ defines the noisy (unprocessed) phase information with $\exp(\cdot)$ being the exponential function, $\jmath = \sqrt{-1}$, and $\angle$ being the associated angle. The signal $\mathbf{X} \in \mathbb{C}^{T\times F}$ in (21) is then decoded back to time-domain via the inverse short-time Fourier transform.

It is worth noting that a time-domain signal with a high sampling rate and long duration results in $P >> K$ for the time-domain dual-path transformers and that computational complexity associated with $\mathfrak{f}_{\mathsf{inter-t}}$ in (4) is significantly larger than $\mathfrak{f}_{\mathsf{intra-t}}$ in (3). In contrast to the time-domain DPTNet and Sepformer, Mag-Sepformer employs a larger frame length $F > K$ to reduce the number of chunks (i.e., $P > T$) and that $P > F$ is set to strike a balance between the computation complexity of $\mathfrak{f}_{\mathsf{inter-t}}$ and $\mathfrak{f}_{\mathsf{intra-t}}$.

While aforementioned approaches, in general, employ dual-path transformers to reduce computational complexity, these approaches still rely on multiple transformers (i.e., $B > 2$). Furthermore, they employ either features extracted from only a single layer of convolution layer in (2) or the unprocessed magnitude spectrum in (19). The use of such encoded features are limiting since they contain largely noisy signals which are subsequently masked via (13) or (20), respectively.

III. The Proposed SMoLnet-T model

With reference to Fig. 2, the proposed small model on low SNR network with post transformer (SMoLnet-T) consists of a frame-wise CNN and a spectral combination transformer. The frame-wise CNN is based on the first ten layers of the SMoLnet, where it extracts high-resolution spectral feature maps. Since the spectral order in each frame is maintained, the subsequent transformer focuses on the extracted feature maps that are crucial based on the overall temporal information.

A. Frame-wise SMoLnet

With $y_{t,f}$ defined in (18), and that $\mathbf{y}_{\Re,t}$ and $\mathbf{y}_{\Im,t} \in \mathbb{R}^{F\times 1}$ are the real and imaginary component of $\mathbf{y}_t = [y_{t,1}, \ldots, y_{t,F}]^{\mathsf{T}}$, respectively, the proposed SMoLnet-T with $L = \log_2(F) + 1$ convolution layers is expressed as

$$\mathbf{Z}_t^{[l]} = \mathfrak{f}_{\mathsf{CNN}}^{[l]}\left(\mathbf{Z}_t^{[l-1]}\right). \tag{22}$$

Here, $1 \le l \le L$ is the layer index and $\mathbf{Z}_t^{[l]} = \left[\mathbf{z}_{t,1}^{[l]}, \ldots, \mathbf{z}_{t,F}^{[l]}\right] \in \mathbb{R}^{C\times F}$ consists the C spectral feature with

$$\mathbf{Z}_t^{[0]} = [\mathbf{y}_{\Re,t}, \mathbf{y}_{\Im,t}]^{\mathsf{T}} \tag{23}$$

Fig. 2. Proposed direct-mapping approach with frame-wise SMoLnet and spectral combination transformer.

being the complex spectral input. The lth convolution layer in (22) is given by

$$\mathfrak{f}_{\mathsf{CNN}}^{[l]} = \mathfrak{f}_{\mathsf{relu}}^{[l]} \circ \mathfrak{f}_{\mathsf{ln}}^{[l]} \circ \mathfrak{f}_{\mathsf{di-conv}}^{[l]} \circ \mathfrak{f}_{\mathsf{pad}}^{[l]} \tag{24}$$

with $\mathfrak{f}_{\mathsf{relu}}^{[l]}, \mathfrak{f}_{\mathsf{ln}}^{[l]}, \mathfrak{f}_{\mathsf{di-conv}}^{[l]}$, and $\mathfrak{f}_{\mathsf{pad}}^{[l]}$ being the ReLU activation, layer normalization [27], convolution, and zero padding functions, respectively. Here, filter $\mathfrak{f}_{\mathsf{di-conv}}^{[l]}$ is of length $k = 3$, and a dilation rate $\mathcal{D}(l) = 2^{l-1}$ is set across the frequency dimension to achieve a receptive field larger than F. In addition, $\mathbf{Z}_t^{[l]}$ has been padded on both ends such that $\mathfrak{f}_{\mathsf{pad}}^{[l]}\left(\mathbf{Z}_t^{[l]}\right) \in \mathbb{R}^{C\times(F+\mathcal{D}(l)+1)}$. Hence, the fth node of the lth layer is achieved by leveraging the corresponding node of the $l-1$th layer. This process facilitates the positional consistency of the frequency information and maintains the high-resolution frequency information throughout. In contrast to employing a transformer that relies on adding sinusoidal encoding to the input or RNN to maintain positional consistency, the proposed CNN ensures position consistency while efficiently exploiting global high-resolution frequency information.

B. Spectral combination transformer

Since high-frequency spectral features have been achieved in $\mathbf{Z}_t^L \in \mathbb{R}^{C\times F}$ in (22), subsequent transformer layers are formulated to focus on the temporal aspect. Given $\mathbf{D}_f^{[1]} = \left[\mathbf{z}_{1,f}^{[L]}, \ldots, \mathbf{z}_{t,f}^{[L]}, \ldots, \mathbf{z}_{T,f}^{[L]}\right]^{\mathsf{T}} \in \mathbb{R}^{T\times C}$ with $\mathbf{z}_{t,f}^{[L]}$ from (22), the proposed SMoLnet-T employs M consecutive transformers

$$\mathbf{D}_f^{[m+1]} = \mathfrak{f}_{\mathsf{SCT}}^{[m]}\left(\mathbf{D}_f^{[m]}\right), \tag{25}$$

where the mth transformer is computed from

$$\mathfrak{f}_{\mathsf{SCT}}^{[m]} = \mathfrak{f}_{\mathsf{ln}_2}^{[m]} \circ \mathfrak{f}_{\mathsf{ffn}_2}^{[m]} \circ \mathfrak{f}_{\mathsf{ffn}_1}^{[m]} \circ \mathfrak{f}_{\mathsf{ln}_1}^{[m]} \circ \mathfrak{f}_{\mathsf{mha}}^{[m]} \circ \mathfrak{f}_{\mathsf{pe}}^{[m]}. \tag{26}$$

For succinctness in notation, we omit m henceforth. Similar to Sepformer [21] and Mag-Sepformer [20], we avoid the need for RNN in DPTNet via position-encoded spectral maps

$$\mathbf{D}'_f = \mathfrak{f}_{\mathsf{pe}}\left(\mathbf{D}_f\right) = \mathbf{D}_f + \mathbf{E}^{T\times C}, \tag{27}$$

where the tth row and cth column of the fixed position encoding $\mathbf{E}^{T\times C}$ is given by [19]

$$e_{t,c} = \begin{cases} \sin\left(\frac{t}{10000^{2c/C}}\right), & \text{if } c \text{ is even;} \\ \cos\left(\frac{t}{10000^{2c/C}}\right), & \text{if } c \text{ is odd.} \end{cases} \tag{28}$$

Compared with RNN and CNN models, where the sequence order is inherent, a transformer-based approach requires positional encoding such as (28) to achieve such information.

With $\mathbf{D}'_f$, the multi-head attention is given by

$$\begin{aligned} \mathbf{A}_f &= \mathfrak{f}_{\mathsf{mha}}\left(\mathbf{D}'_f\right) \\ &= \mathfrak{f}_{\mathsf{cat}}\left(\left[\mathbf{A}_{f,1}, \ldots, \mathbf{A}_{f,h}, \ldots, \mathbf{A}_{f,H}\right]\right)\mathbf{W}^{\mathsf{O}}, \end{aligned} \tag{29}$$

where $\mathbf{W}^{\mathsf{O}} \in \mathbb{R}^{\frac{C}{H}\times C}$ is the output weights and

$$\mathbf{A}_{f,h} = \mathfrak{f}_{\mathsf{softmax}}\left(\frac{\mathbf{Q}_{f,h}\mathbf{K}_{f,h}^{\mathsf{T}}}{\sqrt{C}}\right)\mathbf{V}_{f,h} \tag{30}$$

is the single-head self-attention. Defining weights $\mathbf{W}_h^{\mathsf{Q}}, \mathbf{W}_h^{\mathsf{K}}, \mathbf{W}_h^{\mathsf{V}} \in \mathbb{R}^{C\times\frac{C}{H}}$ for the query, key, and value, respectively, we have

$$\mathbf{Q}_{f,h} = \mathbf{D}'_f\mathbf{W}_h^{\mathsf{Q}}, \tag{31}$$

$$\mathbf{K}_{f,h} = \mathbf{D}'_f\mathbf{W}_h^{\mathsf{K}}, \tag{32}$$

$$\mathbf{V}_{f,h} = \mathbf{D}'_f\mathbf{W}_h^{\mathsf{V}}. \tag{33}$$

Here, $0 \le h \le H$ denotes the head index with H being the number of heads. The hth head of the proposed attention mechanism in (30), therefore, focuses on the important feature map (out of the C features maps) in $\mathbf{D}_f^{[m]}$ for every t. Hence, $\mathbf{A}_f^{[m]}$ encapsulates the top H important feature maps to focus on, resulting in enhanced network training by highlighting the important spectral features within the temporal period T. The attention mechanism in (29) is translatable across frequencies since the proposed CNN in (22) maintains the frequency position consistency for the feature maps.

With $\mathbf{A}_f$ in (29) and $\mathbf{D}'_f$ in (27), the transformer estimates the feature maps for its subsequent transformer via

$$\dot{\mathbf{D}}_f = \mathfrak{f}_{\mathsf{ln}_1}\left(\mathbf{A}_f + \mathbf{D}'_f\right), \tag{34}$$

$$\ddot{\mathbf{D}}_f = \mathfrak{f}_{\mathsf{ffn}_1}\left(\dot{\mathbf{D}}_f\right) = \mathfrak{f}_{\mathsf{relu}}\left(\dot{\mathbf{D}}_f\mathbf{W}_1 + \mathbf{b}_1\right), \tag{35}$$

$$\dddot{\mathbf{D}}_f = \mathfrak{f}_{\mathsf{ffn}_2}\left(\ddot{\mathbf{D}}_f\right) = \mathfrak{f}_{\mathsf{relu}}\left(\ddot{\mathbf{D}}_f\mathbf{W}_2 + \mathbf{b}_2\right), \tag{36}$$

$$\ddddot{\mathbf{D}}_f = \mathfrak{f}_{\mathsf{ln}_2}\left(\dot{\mathbf{D}}_f + \dddot{\mathbf{D}}_f\right), \tag{37}$$

where $\mathfrak{f}_{\mathsf{ffn}_1}$ and $\mathfrak{f}_{\mathsf{ffn}_2}$ are feed-forward networks corresponding to weights $\mathbf{W}_1 \in \mathbb{R}^{C\times 4C}$, $\mathbf{W}_2 \in \mathbb{R}^{4C\times C}$, and biases $\mathbf{b}_1 \in \mathbb{R}^{C\times 1}$ and $\mathbf{b}_2 \in \mathbb{R}^{4C\times 1}$, respectively. Here, $\ddddot{\mathbf{D}}_f$ serves as the input to the subsequent transformer (i.e, $\mathbf{D}_f^{[m+1]}$ in (25)).

With $\mathbf{D}_f^{[M]}$ being the output of the last transformer in (37), the estimated speech is given by

$$[\widehat{\mathbf{x}}_{\Re,f}, \widehat{\mathbf{x}}_{\Im,f}] = \mathbf{D}_f^{[M]}\mathbf{W}_{\mathsf{last}} + \mathbf{b}_{\mathsf{last}}, \tag{38}$$

$$\widehat{\mathbf{x}} = \widehat{\mathbf{x}}_{\Re,f} + \jmath\widehat{\mathbf{x}}_{\Im,f}, \tag{39}$$

where $\mathbf{W}_{\mathsf{last}} \in \mathbb{R}^{C\times 2}$ and $\mathbf{b}_{\mathsf{last}} \in \mathbb{R}^{1\times 2}$ are, respectively, the weights and bias of the last feed-forward neural network. Here, $\widehat{\mathbf{x}}_{\Re,f}$ and $\widehat{\mathbf{x}}_{\Im,f}$ are the real and imaginary

components of the estimated signal, respectively. In contrast to the mask [16] in (13), the proposed transformer performs a direct mapping from the frame-wise spectral maps (i.e. $\mathbf{Z}_t^L \in \mathbb{R}^{C\times F}$ in (22)) to the output. This allows the overall transformer network to decide, at a higher level of abstraction, which subset of high-resolution spectral maps generated by the frame-wise SMoLnet is crucial at t. This is achieved by the consistency in position and the high-resolution frequency information retained by the proposed frame-wise SMoLnet.

IV. Experiment Results

We evaluate our proposed model using a WSJ0 noise dataset collected from a hovering drone and speech utterances from WSJ0-84 [28]. An hour of noise was used and out of the 8714 data points, 6382, 1166, and 1052 were allocated for training, validation, and testing, respectively. The speakers and SNRs for training, validation, and testing were varied to validate the model's generalization for out-of-training speakers and SNRs. More specifically, for training, validation, and testing, the SNRs (in decibels) are drawn from $\{-30, -27, -24, -21, -18\}$, $\{-31, -28, -25, -22, -19\}$, and $\{-32, -29, -26, -23\}$, respectively. A batch size of two was selected to train the proposed SMoLnet-T and the baselines DPTnet [16] and SMoLnet [9]. We note that SMoLnet-T requires low GPU memory usage, and a larger batch size and length can be selected. However, for fair comparison with the baselines, which require significant amount of GPU memory during training, we fixed the batch size to two. For SMoLnet-T and SMoLnet, we set $C = 64$, a Hamming window of length 2048 with 50% overlap, resulting in $F = 1025$ and $T = 123$. For SMoLnet-T, $M \in \{1, 2, 3\}$ was used to highlight the importance of the temporal context achieved via the spectral combination transformer. An Adam optimizer [29] was used with a learning rate of 0.001 and the model with the best validation result over 100 epochs was selected for testing.

We evaluate the speech enhancement performance in terms of speech intelligibility via ESTOI [18] and speech distortion via SI-SDR [15]. As shown in Fig. 3, all models exhibited improved ESTOI and SI-SDR over the received (unprocessed) signal. In general, SMoLnet-T achieves lower speech distortion compared to the baselines for all tested SNRs. In particular, for out-of-training SNR of -32 dB, SMoLnet-T achieves significantly lower speech distortion with an SI-SDR improvement of 2.3 and 9.3 dB over SMoLnet and DPTNet, respectively. Although the SI-SDR is similar across the number of transformers M in SMoLnet-T, ESTOI increases with M. With $M = 3$, SMoLnet-T achieves an ESTOI improvement of 0.02 over SMoLnet at an SNR of -23 dB. The improvement in both speech intelligibility and speech distortion by SMoLnet-T highlights the efficacy of the proposed spectral combination transformer in leveraging temporal context over the CNN approach in SMoLnet. Furthermore, the overall higher performance for SMoLnet-T and SMoLnet compared to DPTnet highlights the viability of direct mapping over masking under low SNR scenarios.

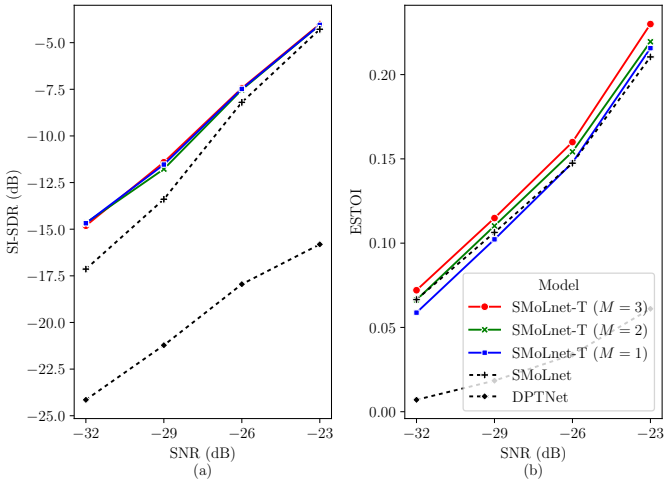

Fig. 3. Speech enhancement performance in terms of (a) SI-SDR [15] and (b) ESTOI [18].

TABLE I
Computational resources requirement for a batch size of 2 each with 64 k samples on the Nvidia RTX6000 Ada.

Method	Train time (mins/epoch)	Model size	GPU mem. (GB)
SMoLnet-T ($M = 1$)	2.8	192 k	4.7
SMoLnet-T ($M = 2$)	4.3	254 k	5.6
SMoLnet-T ($M = 3$)	6.0	317 k	6.7
SMoLnet [9]	1.6	249 k	2.8
DPTnet [16]	40.0	2.6 M	40.0

We evaluate the computation resources required, which includes the training time per epoch, model size in terms of the number of learnable parameters, and the amount of GPU memory usage used during training. The Nvidia RTX6000 Ada with 48 GB of GPU available memory is used. As shown in Table I, the proposed SMoLnet-T requires significantly lower computational resources than DPTNet. More specifically, SMoLnet-T with $M = 1$ requires seven times lower GPU memory to train and is 14 times faster for an epoch. These results highlight that SMoLnet-T can be trained with longer sequences, requires lower GPU resources, and achieves faster training iterations. Although the computation needs are higher than SMoLnet, SMoLnet-T achieves a larger context length ($T = 123$) compared to the nine frames context in SMoLnet, which is beneficial for improving speech intelligibility and reducing speech distortion.

V. Conclusion

We proposed the SMoLnet-T that incorporates a frame-wise CNN with newly formulated spectral-attention transformers. These transformers achieve higher efficacy over CNN in SMoLnet by their ability to leverage high-resolution frequency spectral maps from its frame-wise SMoLnet and

longer temporal context. Experiment results demonstrate that SMoLnet-T achieves improved speech intelligibility and lower speech distortion under low SNR over DPTNet with reduced computational requirements.

REFERENCES

[1] W. Manamperi, T. D. Abhayapala, J. Zhang, and P. N. Samarasinghe, "Drone audition: Sound source localization using on-board microphones," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 21, pp. 508–519, 2022.

[2] M. Wakabayashi, H. G. Okuno, and M. Kumon, "Multiple sound source position estimation by drone audition based on data association between sound source localization and identifications," *IEEE Robot. Autom. Letters*, vol. 5, no. 2, pp. 782–789, 2020.

[3] T. Morito, O. Sugiyama, R. Kojima, and K. Nakadai, "Partially shared deep neural network in sound source separation and identification using a UAV-embedded microphone array," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robot. Syst.*, 2016, pp. 1299–1304.

[4] C. Premachandra and Y. Kunisada, "GAN based audio noise suppression for victim detection at disaster sites with UAV," *IEEE Trans. Services Comput.*, vol. 17, no. 1, pp. 183–193, 2024.

[5] T. Yamada, K. Itoyama, K. Nishida, and K. Nakadai, "Sound source tracking by drones with microphone arrays," in *Proc. IEEE/SICE Int. Symp. Syst. Integration*, 2020, pp. 796–801.

[6] W. Zhang, A. W. H. Khong, and P. A. Naylor, "Adaptive inverse filtering of room acoustics," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 2008, pp. 788–792.

[7] B. Yen, Y. Li, and Y. Hioka, "Rotor noise-aware noise covariance matrix estimation for unmanned aerial vehicle audition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2491–2506, 2023.

[8] B. Yen, Y. Hioka, and B. Mace, "Improving power spectral density estimation of unmanned aerial vehicle rotor noise by learning from non-acoustic information," in *Proc. Int. Work. Acoust. Signal Enhanc.*, 2018, pp. 545–549.

[9] Z.-W. Tan, A. H. T. Nguyen, and A. W. H. Khong, "An efficient dilated convolutional neural network for UAV noise reduction at low input SNR," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 1885–1892.

[10] Z.-W. Tan, A. H. T. Nguyen, Y. Liu, and A. W. H. Khong, "Multichannel noise reduction using dilated multichannel U-net and pre-trained single-channel network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 266–270.

[11] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent.*, 2016, p. 1–13.

[12] 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 Prepr. arXiv1609.03499*, pp. 1–15, 2016.

[13] D. Mukhutdinov, A. Alex, A. Cavallaro, and L. Wang, "Deep learning models for single-channel speech enhancement on drones," *IEEE Access*, vol. 11, pp. 22 993–23 007, 2023.

[14] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU, ITU-T Rec. P. 862, 2000.

[15] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[16] 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.

[17] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–19.

[18] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Info. Process. Syst.*, 2017, pp. 5998–6008.

[20] D. de Oliveira, T. Peer, and T. Gerkmann, "Efficient transformer-based speech enhancement using long frames and STFT magnitudes," in *Proc. Interspeech*, 2022, pp. 2948–2952.

[21] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 21–25.

[22] S. H. Nawab and T. F. Quatieri, "Short-time Fourier transform," in *Adv. Top. Signal Process.* Upper Saddle River, New Jersey: Prentice-Hall, 1987, pp. 289–337.

[23] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[24] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[25] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. Waibel, "Self-attentional acoustic models," in *Proc. Interspeech*, 2018, pp. 3723–3727.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," *in Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1026–1034, 2015.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[28] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," *Linguist. Data Consortium, Philadelphia*, 2007.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 127–142.