INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Streaming Dual-Path Transformer for Speech Enhancement

*Soo Hyun Bae, Seok Wan Chae, Youngseok Kim, Keunsang Lee, Hyunjin Lim, Lae-Hoon Kim*

Samsung Electronics, Hwaseong, Republic of Korea

{soo.hyun.bae, seokwan.chae, yseok712.kim, keunsang.lee, hj38.lim, laehoon.kim}@samsung.com

## Abstract

Speech enhancement employing a dual-path transformer (DPT) with a dilated DenseNet-based encoder and decoder has shown state-of-the-art performance. By applying attention in both time and frequency paths, the DPT learns the long-term dependency of speech and the relationship between frequency components. However, the batch processing of the DPT, which performs attention on all past and future frames, makes it impractical for real-time applications. To satisfy the real-time requirement, we propose a streaming dual-path transformer (stDPT) with zero look-ahead structure. In the training phase, we apply masking techniques to control the context length, and in the inference phase, caching methods are utilized to preserve sequential information. Extensive experiments have been conducted to show the performance based on different context lengths, and the results verify that the proposed method outperforms the current state-of-the-art speech enhancement models based on real-time processing.

**Index Terms**: speech enhancement, dual-path transformer, streaming transformer, masked multi-head self-attention

## 1. Introduction

The goal of speech enhancement is to improve the quality and intelligibility of speech degraded by background noise. It is widely used in speech processing applications such as voice communication, speech recognition and speaker recognition to ensure robust performance. In the past decade, a major breakthrough in speech enhancement was achieved by incorporating powerful deep learning models and has become the prevailing trend in the field [1].

Recently, the transformer has demonstrated success in processing sequential information through multi-head self-attention (MHSA) without relying on recurrent structures [2, 3]. By applying the attention mechanism to sequences and capturing their dependencies, the transformer has achieved better performance than the recurrent neural networks (RNNs) based attention models in the machine translation task. Inspired by the effective sequential modeling of the transformer, many algorithms have been proposed since its first appearance. Furthermore, the dual-path transformer (DPT), which incorporates attention in both time and frequency directions while utilizing convolutional neural networks (CNNs) for the encoder and decoder, has shown better performance in estimating clean speech compared to dual-path RNN models [4, 5, 6].

The DPT-based speech enhancement consists of time-frequency domain methods and time domain methods. In time-frequency domain methods, a noisy input waveform is transformed into a spectrum, and the clean magnitude or complex values are predicted to incorporate phase information [7, 8, 9]. On the other hand, time domain methods directly obtain the clean waveform from a noisy waveform in an end-to-end manner [10, 11, 12]. Both methods include an encoder, consecutive DPTs, and a decoder. The encoder expands the channel dimension using a 1 by 1 convolution and learns the feature representation through densely connected dilated convolution blocks [13, 14]. The decoder is the inverse of the encoder and estimates the enhanced waveform.

Although attention mechanisms are highly capable of learning the long-term dependencies of speech signals and the correlation between frequency components, they are not practical for real-time processing because the transformer requires access to the entire sequences to compute attention. Speech enhancement algorithms based on MHSA have attempted to address this limitation by applying causal attention masks or restricting the time context [15, 16]. However, these approaches are still insufficient for real-time operation as they do not provide a methodology for streaming scenarios. Several attempts have been made to develop streaming frameworks for the transformer, but their use cases were mainly limited to transducers in speech recognition [17, 18]. Motivated by the transformer-XL [19], the authors in [20] designed a chunk-wise streaming transducer that utilizes memory. Streaming techniques employing chunk-wise attention masks and history windows enable low look-ahead latency by segmenting the input into small chunks and combining them with the previous chunk to limit the context length [20, 21, 22].

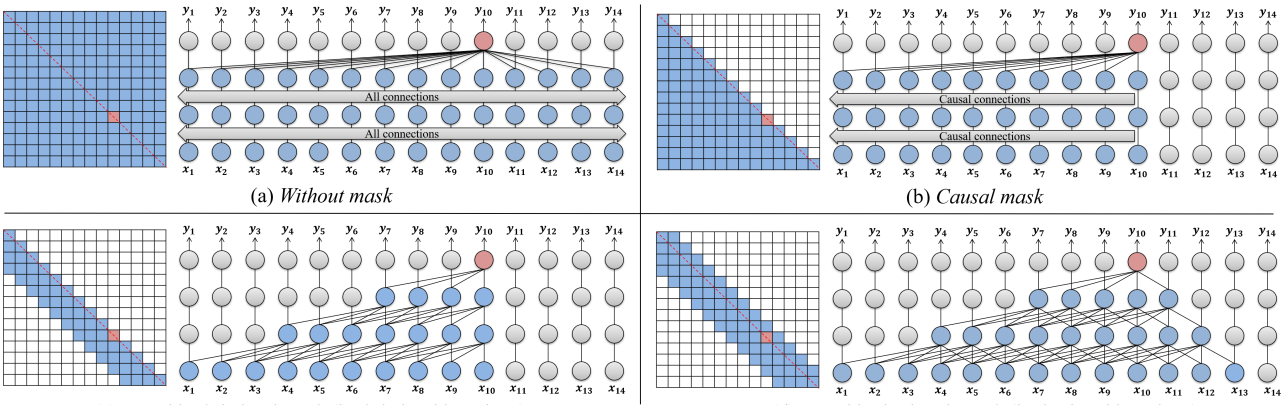
The current batch processing models for speech enhancement based on either transformer or conformer architectures, which have demonstrated promising performance, have a critical limitation: they cannot be applied to real-time applications [7, 8, 23]. To satisfy the real-time constraint, we propose a streaming dual-path transformer (stDPT) for speech enhancement. To the best of our knowledge, this is the first work that enables the DPT to operate in a streaming manner. Our proposed approach, utilizing caching techniques, not only alleviates the drawback of streaming systems that cannot fully utilize sequential information but also enables zero look-ahead streaming. We conducted ablation experiments to examine the impact of context length on the performance of streaming processing. The experimental results show that the proposed approach achieves better results than the current state-of-the-art real-time or causal speech enhancement models, while also having a significantly smaller model size.

## 2. DPT-based speech enhancement

### 2.1. Dilated DenseNet-based encoder and decoder

A dilated convolution can effectively increase the receptive field as the layers become deeper, enabling it to capture long-term information and temporal dependencies in speech signals [24]. In addition, a DenseNet not only effectively utilizes the flow of feature information but also mitigates the problem of vanishing gradient by incorporating the concatenated outputs of all previous layers as input to the current convolutional layer [14]. The

10.21437/Interspeech.2023-751


$y_1$ $y_2$ $y_3$ $y_4$ $y_5$ $y_6$ $y_7$ $y_8$ $y_9$ $y_{10}$ $y_{11}$ $y_{12}$ $y_{13}$ $y_{14}$
All connections
All connections
Causal connections
Causal connections
$x_1$ $x_2$ $x_3$ $x_4$ $x_5$ $x_6$ $x_7$ $x_8$ $x_9$ $x_{10}$ $x_{11}$ $x_{12}$ $x_{13}$ $x_{14}$

(a) *Without mask*

(b) *Causal mask*

(c) *Fixed look-behind mask (look-behind length=3)*

(d) *Fixed look-ahead mask (look-ahead length=1)*

Figure 1: *Different types of attention masks to control the context length. The current frame is depicted in red, while the receptive field applied to the corresponding mask is represented in blue.*

dilated DenseNet, which combines the advantages of aforementioned two models, is commonly employed as both an encoder and a decoder in transformer-based speech enhancement [7].

### 2.2. DPT blocks

The original transformer comprises both encoders and decoders, but in many regression tasks, only the encoder parts are adopted. The main idea of transformer is the MHSA, which allows learning of continuous sequence information without relying on recurrent networks. This attention mechanism involves queries, keys, and values, and it can be organized into multiple heads to capture various aspects, similar to the kernels used in CNNs. The DPT is effective in learning both local and global contextual information from long-term speech sequences where two transformers are connected to perform different types of attention in time and frequency paths. These models have been applied to various speech enhancement tasks and have recently showed outstanding performance [7, 8].

## 3. Proposed stDPT

In this section, we propose a streaming structure for real-time speech enhancement, which includes an encoder, DPT blocks, and decoders, as illustrated in Figure 2.

### 3.1. Encoder

The spectrum magnitude, real and imaginary components of the noisy speech are denoted as $Y_m$, $Y_r$, $Y_i \in \mathbb{R}^{T\times F\times 1}$ respectively, where $T$ and $F$ indicate time and frequency dimensions. $Y_m$, $Y_r$, and $Y_i$ are concatenated and used as an input $Y_{in} \in \mathbb{R}^{T\times F\times 3}$ to the encoder. The encoder consists of a $1 \times 1$ convolutional layer that expands the channel dimension and dilated dense blocks comprising four densely connected dilated convolutional layers with the dilation rate of $\{1, 2, 4, 8\}$. The encoder output $Y_{enc} \in \mathbb{R}^{T\times F\times C}$ is extracted to retrieve intermediate representation of the noisy speech, where $C$ is the channel dimension.

### 3.2. Masked attention for stDPT

During the training of the transformer, a masking method is adopted to control the extent of attention applied among sequences. The masked MHSA is defined as follows:

$$Q_h = ZW_h^Q,\ K_h = ZW_h^K,\ V_h = ZW_h^V,\ h \in [1, H] \quad (1)$$

$$head\_attention_h = softmax\left(\frac{Q_h K_h^T}{\sqrt{d}} + M\right) V_h \quad (2)$$

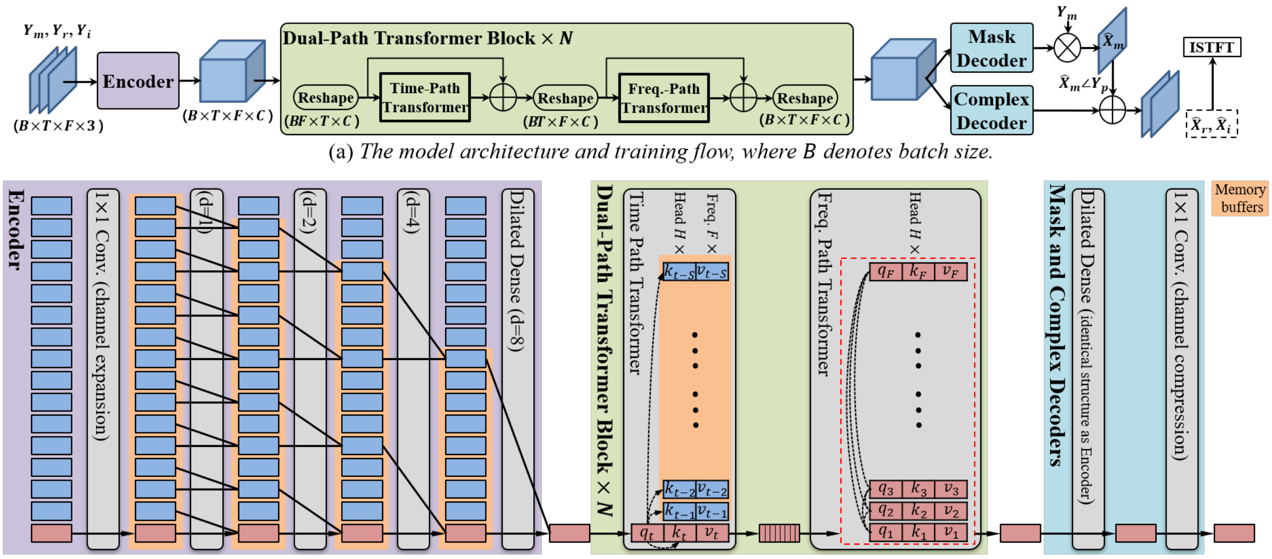
$$M_{i,j} = \begin{cases} 0, & if \text{ attention sequence} \\ -\infty, & else \end{cases} \quad (3)$$

where $Z \in \mathbb{R}^{l\times d}$ are input sequences of each transformer block with length $l$ and dimension $d$, and $Q_h$, $K_h$, $V_h \in \mathbb{R}^{l\times d/H}$, $H$ are the mapped queries, keys, values of $h$-th head, and the number of head, respectively. $W_h^Q$, $W_h^K$, $W_h^V \in \mathbb{R}^{d\times d/H}$ are linear transformation matrices, and $M$ indicates the attention mask which determines the range of sequences to be involved for attention. Masking is specifically applied to the transformers in the time path, as the frequency path is not relevant to causality. Figure 1 shows different types of attention masks. In each section of the figure, the left side represents the attention mask, while the right side provides a schematic representation of the receptive field applied to the corresponding mask.

A general transformer performs attention to all input sequences without any masking, as depicted in Figure 1(a). To achieve causal attention, the mask shown in Figure 1(b) is utilized, which only allows attention to the past input sequences. We apply the attention mask shown in Figure 1(c) when training the model. By applying a limited context length $S$ with zero look-ahead masking, a streaming system has been achieved during the inference phase. If there is a small room for additional look-ahead frames, the masking shown in Figure 1(d) is applied. However, using a fixed look-ahead mask in Figure 1(d) for all layers of the transformers has a drawback because the number of future frames increases linearly with the increasing number of transformer layers. Therefore, a dual-mask method has been applied to maintain a fixed size of look-ahead frames. Specifically, the first layer of the transformer utilizes the look-ahead mask to incorporate limited future context, while the remaining layers adopt look-behind masking to prevent the expansion of the future receptive field. A DPT block consists of two sequentially connected transformers that perform attention in both time and frequency paths. By stacking these blocks $N$ times, the model captures both time and frequency dependencies in an alternating manner. The overall model architecture and training flow are depicted in Figure 2(a).

### 3.3. Mask and complex decoders

Phase information is a significant factor that improves the performance of speech enhancement [25]. Inspired by the previous works in [8] and [26], two decoders have been con-



(a) *The model architecture and training flow, where B denotes batch size.*

(b) *The detail of streaming inference, where red boxes represent the flow of current frame.*

Figure 2: *The overall diagram of the proposed approach.*

structed to predict the magnitude mask and complex components. Both decoders employ the dilated DenseNet, similar to the one in the encoder, and are followed by a 1 × 1 convolution to reduce the channel dimension. The masked magnitude $\widehat{X}_m \in \mathbb{R}^{T\times F\times 1}$ is obtained by element-wise multiplication of the mask decoder output with the input noisy magnitude $Y_m$. The complex decoder estimates the real and imaginary components $(\widehat{X}'_r, \widehat{X}'_i) \in \mathbb{R}^{T\times F\times 2}$, and further enhances the complex parts obtained by the mask decoder as follows:

$$\widehat{X}_r = \widehat{X}_m \cos Y_p + \widehat{X}'_r,\ \widehat{X}_i = \widehat{X}_m \sin Y_p + \widehat{X}'_i \tag{4}$$

where $Y_p$ is the phase of the input noisy spectrum, and $(\widehat{X}_r, \widehat{X}_i)$ are final complex spectrum estimated by the model.

### 3.4. Loss function

In order to estimate both the magnitude and phase information of speech, the loss function is composed of a linear combination of the magnitude loss $\mathcal{L}_{mag}$, complex loss $\mathcal{L}_{comp}$, and waveform loss $\mathcal{L}_{wav}$ as follows:

$$\begin{aligned}\mathcal{L}_{mag} &= \mathbb{E}\left[\left|X_m - \widehat{X}_m\right|^2\right]\\ \mathcal{L}_{comp} &= \mathbb{E}\left[\left|X_r - \widehat{X}_r\right|^2\right] + \mathbb{E}\left[\left|X_i - \widehat{X}_i\right|^2\right]\\ \mathcal{L}_{wav} &= \mathbb{E}\left[\left|x - \widehat{x}\right|\right]\end{aligned} \tag{5}$$

where $X_m$, $X_r$, $X_i$, $x$ and $\widehat{x}$ refer to the clean target magnitude, real part, imaginary part, waveform and enhanced waveform, respectively. The total training loss is defined as follows:

$$\mathcal{L}_{total} = \alpha_1\mathcal{L}_{mag} + \alpha_2\mathcal{L}_{comp} + \alpha_3\mathcal{L}_{wav} \tag{6}$$

where $\alpha_1$, $\alpha_2$, and $\alpha_3$ are the weights of the corresponding losses, which are set to 0.5, 0.2 and 0.3 respectively.

### 3.5. Streaming inference methods

The trained model, which uses masked MHSA to limit the context length as described in Section 3.2, enables the DPT to operate as a streaming system during inference. Our proposed streaming inference method is illustrated in Figure 2(b). In the encoder and decoders, the causal DenseNet block is applied only for the current frame, considering limited future information. As layers are stacked, the convolution results are cached to concatenate the outputs of the previous layers consecutively. At the DPT, the proposed masking-based training technique is used only for the time path, meaning there is no caching for the frequency path transformer. In the frequency path, attention among the frequency components is obtained within the current frame. The query in the current frame is then applied to the cached previous keys and values, and the MHSA is obtained as follows:

$$head\_attention_{h,t} = softmax\left(\frac{Q_{h,t}K^T_{h,t-S:t}}{\sqrt{d}}\right)V_{h,t-S:t} \tag{7}$$

where $t$ is the current time index. In comparison to Equation (2), the steaming method truncates the history sequences to the context length $S$, which is the number of look-behind frames of the mask applied during the training phase.

## 4. Experiments and ablation study

### 4.1. Dataset

We conducted experiments on the Voice Bank+DEMAND dataset, which is widely used in speech enhancement research, to evaluate the performance [27]. The dataset was selected from the Voice Bank corpus [28] and consists of 11,572 utterances from 28 speakers as the training set, and 872 utterances from 2 unseen speakers as the test set. In the training set, clean utterances were artificially mixed with 10 different noise types (8 types from DEMAND [29] and 2 artificial types) at 0, 5, 10, and 15 dB SNRs. In the test set, 5 unseen noise types were mixed at 2.5, 7.5, 12.5, and 17.5 dB SNRs. All utterances were down-sampled from 48 kHz to 16 kHz.

### 4.2. Experimental setup

The input frame is applied with a Hanning window of 25$ms$ (400-point FFT), and a hop size of 6.25$ms$ (75% overlap). The channel dimension $C$, which is extended by 1 × 1 convolution, is set to 64. Four DenseNets with a dilated rate of 2 are used for the encoder and decoders. All convolutional layers are followed

Table 1: *The performance variation with the number of look-behind frames (0-frame look-ahead). "$\leftarrow S$" denotes the history context length.*

| #($\leftarrow S$) | **PESQ** | **CSIG** | **CBAK** | **COVL** | **SSNR** |
|---|---|---|---|---|---|
| Full | 3.126 | 4.503 | 3.776 | 3.891 | 10.691 |
| Causal | 3.046(97.5) | 4.403(97.8) | 3.683(97.5) | 3.798(97.6) | 9.918(92.8) |
| 128 | 3.031(97.0) | 4.402(97.8) | 3.674(97.3) | 3.789(97.4) | 9.894(92.5) |
| 96 | 2.965(94.9) | 4.383(97.3) | 3.637(96.3) | 3.738(96.1) | 9.823(91.9) |
| 64 | 2.947(94.3) | 4.356(96.7) | 3.644(96.5) | 3.715(95.5) | 10.039(93.9) |
| 32 | 2.912(93.2) | 4.360(96.8) | 3.623(96.0) | 3.697(95.0) | 9.971(93.3) |
| 16 | 2.908(93.0) | 4.337(96.3) | 3.620(95.9) | 3.683(94.7) | 9.951(93.1) |

Table 2: *The performance variation with the number of look-ahead frames (32-frames look-behind). "$\rightarrow S$" denotes the future context length.*

| #($\rightarrow S$) | **PESQ** | **CSIG** | **CBAK** | **COVL** | **SSNR** |
|---|---|---|---|---|---|
| 0 | 2.912 | 4.360 | 3.623 | 3.697 | 9.971 |
| 1 | 2.919(100.2) | 4.350(99.8) | 3.625(100.1) | 3.694(99.9) | 9.986(100.1) |
| 2 | 2.930(100.6) | 4.372(100.3) | 3.629(100.2) | 3.714(100.5) | 9.934(99.6) |
| 4 | 2.981(102.4) | 4.387(100.6) | 3.676(101.5) | 3.747(101.4) | 10.270(103.0) |
| 8 | 2.990(102.7) | 4.390(100.7) | 3.659(101.0) | 3.756(101.6) | 10.019(100.5) |
| 16 | 2.998(102.9) | 4.393(100.8) | 3.671(101.3) | 3.764(101.8) | 10.072(101.0) |
| 32 | 3.019(103.7) | 4.405(101.0) | 3.701(102.1) | 3.780(102.2) | 10.363(103.9) |

by instance normalization and PReLU activation function [30]. The number of DPT blocks $N$ is 4, and the number of heads $H$ is 4. Training is conducted using Adam optimization, and gradient clipping is applied with L2-norm of 5 to avoid gradient explosion. The learning rate starts at 0.008 and decreases exponentially. The total number of training epochs is set to 100.

### 4.3. Evaluation metrics

We have selected metrics to evaluate the performance of enhanced speech in various aspects. Higher values indicate better performance for all metrics. The objectives of each metric are summarized as follows:

- PESQ: Perceptual evaluation of speech quality defined in the ITU-T P.862.2 standard [31].
- SSNR: Segmental SNR, which is the average of the SNR per frame for two speech signals.
- CSIG, CBAK, and COVL: Mean Opinion Score (MOS) prediction of the signal distortion, the intrusiveness of background noise, and the overall effect, respectively [32].

### 4.4. Results and ablation analysis

We conducted extensive experiments to evaluate the performance based on different context lengths, and the results were analyzed to ensure that the performance of the proposed stDPT is as expected. Table 1 shows the results of applying attention only to history frames with zero look-ahead, where $(\cdot)$ represents the percentage degradation compared to the performance of the batch processing method that performs attention on all past and future frames. It is observed that with a smaller number of past frames, there is more degradation in terms of PESQ and SSNR compared to MOS. When using 16 frames of history masking, the MOS scores showed approximately 5% degradation compared to the batch processing results, while PESQ and SSNR showed about 7% degradation in performance.

Table 2 compares the performance of increasing the look-ahead sequences from 1 to 32 while keeping the past masking fixed at 32, where $(\cdot)$ indicates the percentage of improvement. The longer look-ahead frames yield slightly higher performance in all metrics, but the improvement is not as significant as expected. The reason for this is that in the transformer blocks of the time path, the previous keys and values are already cached, allowing for the full utilization of history information even as the transformer layers increase. However, in the case of look-ahead, only the first layer utilizes future information to prevent linearly growing receptive field. This leads to the problem that the future information becomes more diluted as the transformer is stacked, resulting in limited performance improvement. Therefore, the proposed streaming structure is more sensitive to past frames than future information. If we can leverage sufficient past information, the stDPT can achieve real-time speech enhancement with only processing time delay and without significant performance degradation. In the case of zero look-ahead, the total algorithmic latency (frame length + frame shift) is 25 + 6.25 = $31.25ms$.

Table 3: *Comparison with other state-of-the-art casual and real-time speech enhancement models on the Voice Bank+DEMAND dataset.*

| **Model** | **#para(M)** | **PESQ** |
|---|---|---|
| Noisy | - | 1.97 |
| RNNoise [33] | 0.06 | 2.29 |
| PercepNet [34] | 8 | 2.73 |
| DCCRN [35] | 3.7 | 2.68 |
| DCCRN+ [36] | 3.3 | 2.84 |
| FSFusionNet [37] | 3.1 | 2.905 |
| stDPT ($S$=16) | 1.14 | **2.908** |
| stDPT ($S$=32) | 1.14 | **2.912** |
| stDPT ($S$=64) | 1.14 | **2.947** |
| stDPT ($S$=96) | 1.14 | **2.965** |
| stDPT ($S$=128) | 1.14 | **3.031** |

To further validate the proposed method, we compared the performance with other state-of-the-art real-time speech enhancement techniques, including RNNoise [33], PercepNet [34], DCCRN [35], DCCRN+ [36] and FSFusionNet [37]. These models are based on either dual-path RNNs or complex convolutional layers. As shown in Table 3, the proposed method outperforms the other models in terms of PESQ score. Across all history context lengths from 16 to 128 with zero look-ahead, the stDPT showed superior performance compared to the other approaches. Note that, the model parameters are reduced by approximately 3 times compared to existing models meaning that the proposed model efficiently estimates clean speech with a relatively smaller model size.

## 5. Conclusions

The DPTs have demonstrated promising performance by incorporating attention mechanisms in both time and frequency paths. However, most of the existing models are based on batch processing, which poses a significant limitation for real-time applications. To overcome this challenge, we proposed a streaming approach to DPT in this paper. Our approach involves training DPTs using masking methods and conducting inference using caching techniques to limit the length of the past context while enabling streaming with zero look-ahead. Experimental results showed that our method outperforms state-of-the-art models in terms of real-time and causal speech enhancement processing, while also achieving a much smaller model size.

## 6. References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[3] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, 2022.

[4] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.

[5] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-path convolution recurrent network for single channel speech enhancement," in *Proc. Interspeech 2021*, pp. 821–825.

[6] C. Li, Y. Luo, C. Han, J. Li, T. Yoshioka, T. Zhou, M. Delcroix, K. Kinoshita, C. Boeddeker, Y. Qian *et al.*, "Dual-path RNN for long recording speech separation," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 865–872.

[7] F. Dang, H. Chen, and P. Zhang, "DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6857–6861.

[8] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric GAN for speech enhancement," in *Proc. Interspeech 2022*, pp. 936–940.

[9] 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.

[10] 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.

[11] K. Wang, B. He, and W.-P. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7098–7102.

[12] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 21–25.

[13] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *ICLR*, 2016.

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

[15] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Communication*, vol. 125, pp. 80–96, 2020.

[16] M. Strake, A. Behlke, and T. Fingscheidt, "Self-attention with restricted time context and resolution in DNN speech enhancement," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[17] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7829–7833.

[18] Z. Tian, J. Yi, Y. Bai, J. Tao, S. Zhang, and Z. Wen, "Synchronous transformers for end-to-end speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7884–7888.

[19] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2978–2988, 2019.

[20] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5904–5908.

[21] V. Garg, W. Chang, S. Sigtia, S. Adya, P. Simha, P. Dighe, and C. Dhir, "Streaming transformer for hardware efficient voice trigger detection and false trigger mitigation," in *Proc. Interspeech 2021*, pp. 4209–4213.

[22] O. O. Rudovic, A. Bindal, V. Garg, P. Simha, P. Dighe, and S. Kajarekar, "Streaming on-device detection of device directed speech from voice and touch-based invocation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 491–495.

[23] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech 2020*, pp. 5036–5040.

[24] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6629–6633.

[25] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[26] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7847–7851.

[27] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech." in *SSW*, 2016, pp. 146–152.

[28] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Oriental International Conference on Speech Database and Assessments COCOSDA*, 2013, pp. 1–4.

[29] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, 2013.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE international conference on computer vision (ICCV)*, 2015, pp. 1026–1034.

[31] ITUT Rec., "P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH–Geneva*, 2005.

[32] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[33] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *IEEE 20th international workshop on multimedia signal processing (MMSP)*, 2018, pp. 1–5.

[34] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," in *Proc. Interspeech 2020*, pp. 2482–2486.

[35] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech 2020*, pp. 2472–2476.

[36] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-wise subband DCCRN with SNR estimation for speech enhancement," in *Proc. Interspeech 2021*, pp. 2816–2820.

[37] Z. Chen and P. Zhang, "Lightweight full-band and sub-band fusion network for real time speech enhancement," in *Proc. Interspeech 2022*, pp. 921–925.