Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

TSDT-Net: Ultra-Low-Complexity Two-Stage Model Combining Dual-Path-Transformer and Transform-Average-Concatenate Network for Speech Enhancement

Yi Gao[1], Hangting Chen[2], Siyu Zhang[1], Qingshan Yang[1], Jingcong Chen[1]

[1]Tencent WXG Chengdu, Audio and Speech Signal Processing Oteam
[2]Tencent AI Lab, Audio and Speech Signal Processing Oteam
{jackyyigao, erichtchen, serriezhang, noelyang, soapchen}@tencent.com

Abstract

This paper introduces TSDT-Net, a dual-stage ultra-low-complexity architecture for speech enhancement which achieves higher denoising performance with limited parameter number and computational cost. Its first stage utilizes a simplified Dual-Path-Transformer(DPT) structure. In the second stage, the first-stage output and original noisy signal are treated as dual-channel inputs, modeled as a beamforming optimization problem. An enhanced Transform-Average-Concatenate (TAC) network processes these channels through spectral filtering and enhancement. Fast linear transformers ensure ultra-low computational overhead, while gated networks in both stages facilitate complex Ideal Ratio Mask (cIRM) construction. Residual connections between stages enable performance synergies. Evaluations on the INTERSPEECH 2020 DNS Challenge demonstrate TSDT-Net's superiority, achieving state-of-the-art DNSMOS and PESQ scores with significant margins over single-stage models under stringent computational constraints (<700K parameters and <500M/ 200M MACs). This efficiency enables deployment across diverse embedded devices.

Index Terms: speech enhancement, Low Complexity, Two-Stage, DPT, TAC

1. Introduction

Ambient noise degradation in real-world environments drives the need for speech enhancement (SE) to improve communication quality and downstream task performance (e.g., speech recognition, diarization). While SE serves as a critical front-end for robust systems, achieving high denoising efficacy under strict computational constraints remains essential for deployment on resource-limited embedded devices.

In speech enhancement, most model architectures employ convolution, recurrent neural network (RNN) and/or attention mechanisms. U-Net architectures, characterized by its encoder-decoder structure with skip connections, demonstrate superior performance in noise reduction while preserving speech quality and intelligibility [1] - [4]. Recent advancements, exemplified by FullSubNet [5][6], reveal that RNN-based frameworks achieve comparable enhancement efficacy with reduced computational demands. Though these models achieve high performance, they are unsuitable for resource constrained devices due to the large computational cost.

Compared with single-stage model, multi-stage networks have been extensively investigated to partition single-step optimization problems into progressive subtasks, facilitating hierarchical refinement through task decomposition. For instance, [7] adopts a dual-stage approach: binary mask prediction followed by complex spectrum restoration. Similarly, magnitude and complex spectrum enhancement are decoupled in frameworks like GaGNet [8], TEA-PSE [9] and FDFNet [10]. Architectures such as Gesper [11] and RaD-Net [12] further decompose tasks into restoration and denoise phases. Notably, U-Net structures are integrated in one or both stages of these models, contributing to computational overheads ranging from several to tens of Giga MACs.

To optimize parameter count and computational complexity, Meng23 [13] and THLNet [14] compress model parameters to below 600K, employing simplified U-Net architectures with Mel-scale frequency bands. However, their computational costs remain above 1G MACs. TaErLite [15] replaces the original U-Net in TaylorSENet [16] with lightweight grouped GRU layers, achieving reduced complexity while maintaining competitive performance. Dual-path strategies [17][18] significantly shrink parameter sizes by reusing compact RNNs across dimensions. Band splitting [19][20] and frame skipping [21][22] techniques further lower computation costs in such models. DeepfilterNet [23] leverages speech harmonic structures to enable computationally efficient enhancement, but contains over 2.3M parameters and exhibits degraded performance in reverberant environments. Most models fail to balance model parameters and computational cost.

This work proposes a novel two-stage framework to achieve higher denoising performance under limited parameter number and computational cost. The model synergistically integrates the advantages of DPT-FSNet[18]-based architectures and RNN-based models. The framework achieves ultra-low computational complexity below 500M MACs, enabling seamless deployment on mainstream edge devices. Key contributions include:

1) The first-stage network employs a simplified DPT-FSNet architecture incorporating a dual-path transformer module. Fast autoregressive transformers with linear attention mechanisms [24] replace conventional transformers, while decoder CNN layers are substituted with feedforward layers.
2) The second-stage network reformulates speech enhancement as a beamforming optimization, accepting dual input streams: (i) processed outputs from the first-stage network and (ii) original noisy speech. A Transform-Average-Concatenate (TAC) module [25] serves as the architectural core, selected for its permutation invariance and input compression capabilities across multi-channel inputs, ensuring robust enhancement performance.

10.21437/Interspeech.2025-429

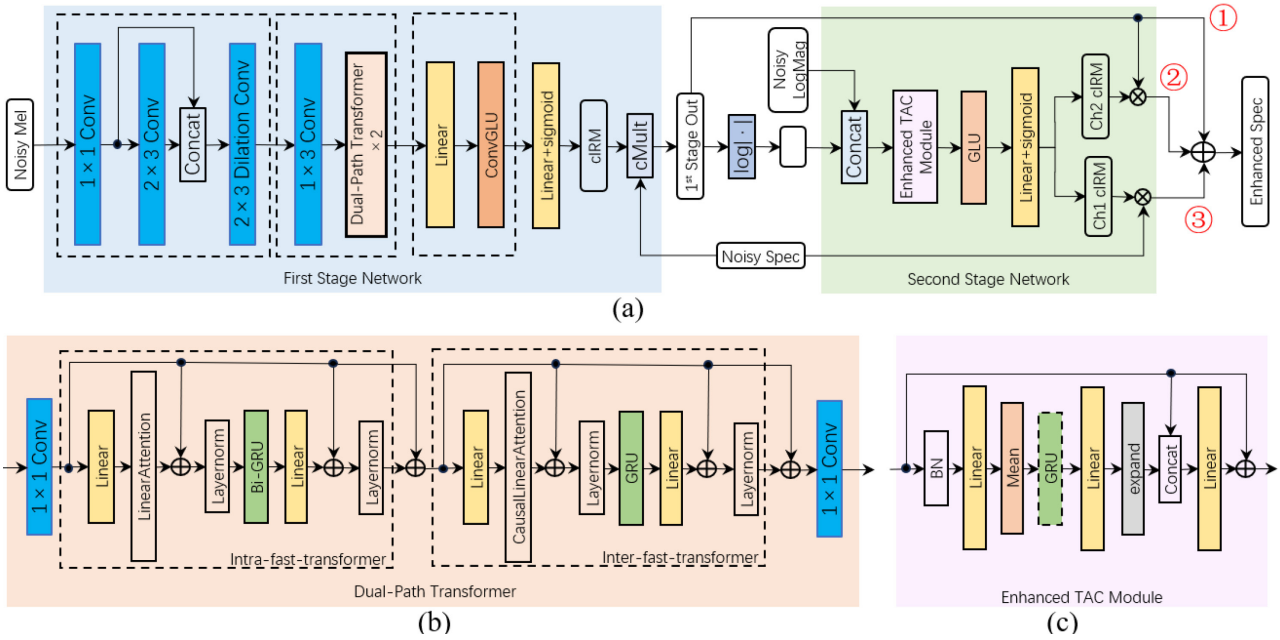


Figure 1: (a) *Overall structure of the TSDT-Net.* (b) *The details of the Dual-Path Transformer(DPT) module enhanced with .* (c) *The details of the enhanced Transform-Average-Concatenate(TAC) module.*

3) To preserve the synergistic advantages of both network stages, Gated Linear Unit (GLU) [26] or Convolutional GLU (ConvGLU) [2] layers are integrated prior to constructing complex Ideal Ratio Masks (cIRMs) in the first and second stages. Residual connections are applied to outputs from both stages, enhancing performance robustness and training stability.

2. TSDT-Net

As illustrated in Figure 1, our two-stage speech enhancement system operates as follows. The first-stage network processes noisy Mel-band energies via an encoder-decoder architecture to generate a cIRM, which is complex-multiplied with the noisy spectrum to produce the initial clean spectrum estimate. This output computes logarithmic power spectra(LPS), concatenated with the noisy LPS to form dual-channel features for the second-stage input. To reduce computational overhead, a TAC-based network predicts dual-channel cIRMs, yielding the enhanced clean spectrum. Residual connections combine first-stage outputs with second-stage dual-channel results, enabling automatic fusion of three outputs(①~③ in Figure 1(a)).

Subsequent sections detail each stage's architecture, omitting reshape/permute operations for brevity.

2.1. The First Stage Network

The first-stage network architecture comprises three core components delineated by dashed boxes in the schematic: an encoder, DPT module, and decoder. The encoder transforms noisy Mel-band energy segments into intermediate feature representations. These features are processed by the DPT layer to generate source-specific masks. Departing from the DPT-FSNet implementation in [18], computational optimization is achieved by replacing CNN-based dense decoder blocks with fully-connected(FC) layers and ConvGLUs. The final cIRM prediction is executed through FC layers followed by sigmoid activation.

2.1.1. Encoder

Initially, the input Mel-spectrogram undergoes a 2-D convolution operation, transforming single-channel input into a C-channel tensor $X \in \mathbb{R}^{C\times T\times F}$, where F denotes Mel-frequency bins and T represents frame count. The dilated dense block employs a multi-stage process: the initial convolutional layer preserves channel dimensionality C. Subsequent concatenation operations expand channels to $2C$, followed by a dilated convolutional layer (dilation factors (2,1)) that compresses channels back to C, yielding spectral feature maps $U \in \mathbb{R}^{C\times T\times F}$. Each convolutional output undergoes Layer Normalization and PReLU activation before propagating to subsequent layers, ensuring feature stability and nonlinear transformation.

2.1.2. Dual-Path Transformer(DPT)

Prior to DPT processing, encoder outputs are channel-adapted via a 1×3 convolutional layer, generating tensor $H \in \mathbb{R}^{E\times T\times F}$. Post LayerNorm and PReLU activation, H enters the DPT module. The DPT architecture initiates with an input convolutional layer reducing channel dimensions by half, producing 3D tensor $D \in \mathbb{R}^{K\times T\times F}$, $(K = E/2)$. As illustrated in Fig.1(b), each DPT unit contains dual transformers: intra-transformer capturing sub-band dependencies and inter-transformer modeling full-band correlations, with alternating time-frequency path processing as per [18]. Computational efficiency is enhanced through linear attention mechanisms and causal linear attention [24]. A concluding convolutional layer outputs tensor $G \in \mathbb{R}^{K\times T\times F}$, completing spectral feature refinement.

2.1.3. Decoder

The decoder architecture employs a linear layer to transform the DPT output into a compressed representation $G \in \mathbb{R}^{K\times T\times 80}$, followed by ConvGLU layers for computational efficiency. Each ConvGLU branch utilizes 1×1 convolutional layers to

Table 1: *Parameter sizes and computation costs of tested models and objective results of them on DNS testset*

Model	Para. (M)	MACs (M/s)	No-Reverb			With-Reverb			DNSMOS		
			PESQ	SISNR[dB]	eSTOI[%]	PESQ	SISNR[dB]	eSTOI[%]	SIG	BAK	OVL
Noisy	/	/	1.58	9.23	80.99	1.82	9.19	77.61	4.19	2.4	2.76
Inf-Fullsubnet	1.58	158.67	2.46	**15.62**	89.09	2.56	14.41	82.22	**3.87**	3.93	3.42
NS-DeepVQE	0.16	154.89	2.21	14.06	86.92	2.24	13.06	79.49	3.74	3.85	3.28
TSDT_1stNet	0.11	164.78	2.33	15.05	88.70	**2.58**	**14.73**	**82.89**	3.83	4.06	3.44
TSDT-SumAll	0.66	153.55	**2.55**	**15.84**	89.61	2.55	14.39	82.18	**3.84**	**4.12**	**3.48**
-Sum①&②			2.40	15.40	88.81	**2.62**	14.32	81.89	3.82	4.14	3.46
-Sum①&③			**2.53**	15.48	**89.93**	2.54	14.23	81.71	3.82	**4.14**	**3.47**
-Sum②&③			2.50	**15.71**	**89.79**	2.56	**14.67**	**82.25**	3.81	**4.12**	3.45
Inf-Fullsubnet	4.50	450.82	2.55	**16.02**	89.98	2.64	14.91	83.12	3.88	4	3.46
Fast-Fullsubnet	1.63	469.56	2.42	15.51	88.70	2.55	14.33	82.08	3.85	3.93	3.4
NS-DeepVQE	0.61	450.95	2.57	15.60	89.68	2.54	14.35	81.39	**3.89**	4.13	3.52
TSDT_1stNet	0.15	482.17	2.55	15.85	89.89	**2.74**	**15.22**	83.65	3.88	4.18	**3.53**
TSDT-SumAll	0.56	479.90	**2.68**	**16.55**	**90.59**	**2.76**	**15.25**	**84.10**	**3.89**	4.19	**3.55**
-Sum①&②			2.58	15.79	90.13	2.71	14.99	83.02	3.86	**4.21**	**3.53**
-Sum①&③			2.62	15.68	90.38	**2.74**	14.79	83.32	3.87	4.19	**3.53**
-Sum②&③			**2.65**	**16.37**	**90.56**	**2.74**	15.17	**83.82**	3.85	**4.20**	3.52

reduce *K*-channel features to 3 channels. Post-reshape operations yield 240-dimensional per-frame features. These undergo a (240×322) FC layer and sigmoid activation, generating dual 161-dimensional masks corresponding to cIRM real and imaginary components.

2.2. The Second Stage Network

To mitigate potential performance degradation in the first-stage network caused by Mel-band spectral merging, a secondary pathway incorporates the original noisy speech's 161-dimensional LPS(16kHz sampling rate, FFT length 320). The Transform-Average-Concatenate (TAC) module applies feature fusion and dimensional compression to these dual streams, followed by feedforward processing for joint cIRM estimation.

2.2.1. Transform-Average-Concatenate (TAC)

The TAC module receives concatenated inputs from two signal pathways: (1) logarithmic power spectra derived from first-stage network outputs, and (2) raw noisy signal logarithmic power spectra. This concatenation produces input tensor $Z \in \mathbb{R}^{2\times160\times T}$, where the dimensionality 2 corresponds to dual input streams, 160 represents DC component removed spectral dimensions.

The TAC module implementation follows the methodology detailed in [25], preserving input-output dimensional equivalence. The *transform* and *average* stages align with parameter-sharing principles and permutation-invariant pooling operations, and the *concatenate* stage facilitates beamforming-inspired feature fusion. Residual connections within the TAC phase enhance gradient flow, enabling seamless integration into deep networks without optimization instability.

To optimize parameter efficiency, the GRU layer is strategically positioned after TAC's *average* operation, operating on pooled features (dimensionality reduced from $\mathbb{R}^{2\times160\times T}$ to $\mathbb{R}^{1\times160\times T}$). This optimized implementation, termed the *enhanced TAC* module, achieves computational economy by halving recurrent processing overhead while preserving discriminative performance through feature dimensionality reduction.

Post-TAC processing employs feedforward layers to estimate dual-path denoised complex spectra.

2.2.2. Feedforward layers

The TAC module outputs tensor $Y \in \mathbb{R}^{2\times160\times T}$. Post-processing involves reshaping Y into Y', $Y' \in \mathbb{R}^{T\times320}$ dimensionality prior to GLU integration, optimizing temporal feature alignment and nonlinear gating operations. GLU integration mitigates gradient vanishing in deep networks while maintaining nonlinear activation capabilities through multiplicative gating mechanisms.

The final processing applies linear projection with sigmoid activation to produce dual-channel 160-D cIRM. Zero-padding restores DC components (T×161 per channel). Complex multiplication between cIRMs and noisy spectra yields dual clean speech estimates, with inverse STFT reconstructing temporal signals.

Inter-network residual connections enable automatic combinatorial optimization across three output pathways, potentially enhancing model efficacy and training stability through multi-path signal fusion. This architecture facilitates implicit weighting of spectral enhancement contributions from distinct processing branches.

3. Experimental Setup

3.1. Configuration

All evaluated models shared identical temporal configurations: zero lookahead, 20ms Hanning windows, and 10ms frame shifts. With 16kHz sampling rate, FFT length was fixed at 320 points. GRU units uniformly replaced LSTMs across architectures to standardize recurrent network implementations. Model variants were categorized into two computational tiers: larger configurations approximating 460M MACs and compact

versions operating at ~160M MACs, enabling comparative analysis of performance-scaling tradeoffs.

• Fast-FullSubNet[22]：utilizes a shrinkage factor of 4 and 64 Mel-frequency bands with encoder input dimensionality 161. Architectural parameters include encoder/decoder hidden dimensions (256/256) and bottleneck layer dimensionality 144, resulting in 469.56M MACs. Comparative evaluations exclude smaller model variants due to observed significant performance degradation under parameter reduction.

• Inf-Fullsubnet[22]: adopts $+\infty$ downsampling factor to bypass bottleneck layers, enabling full-band sequential processing with 64 Mel-frequency bins. encoder/decoder hidden dimensions : (1) Larger model (450.82M MACs): 440, (2) Compact model(158.67M MACs): 256.

• NS-DeepVQE: removes the far-end branch from DeepVQE [27], demonstrating superior denoising performance over CRUSE [28]. Two variants are implemented: (1) Larger model(450.95M MACs): Four encoder blocks (channel transitions: 2→16, 16→32, 32→40, 40→24), bottleneck (input/hidden: 264→256), four decoder blocks (24→40, 40→32, 32→16, 16→27), culminating in a complex convolutional mask block. (2) Compact model(154.89M MACs): Three encoder blocks (2→16, 16→16, 16→8), bottleneck (168→128), three decoder blocks (8→16, 16→16, 16→27).

• TSDT: (1) Larger model (479.90M MACs) with 64 Mel-frequency bins (F=64), input convolution channel C=16, and DPT adapter convolution channel E=40. The second stage uses 1-layer GRU. (2) Compact model (153.55M MACs) using C=8, E=16, 2-layer GRU. TSDT-SumAll aggregates all three branches, while TSDT-Sum①&② fuses only the first and second branches as shown in Figure 1(a).

• TSDT-1stNet: Only the first stage network of TSDT model, reset C and E to make the computation complexty comparable: (1) Larger model (482.17M MACs) C=24, E=40. (2) Compact model (164.78M MACs) using C=8 and E=24.

The TSDT-Net framework employs the $\mathcal{L}_{PLCPA-ASYM}$ loss proposed in [29] during model training.

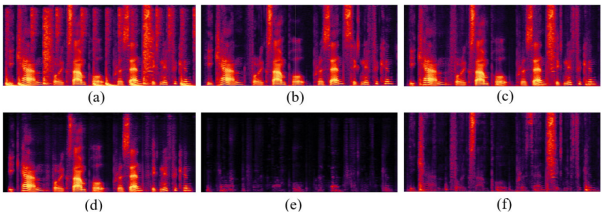


Figure 2: *Spectrograms*. (a) *Noisy input.* (b) Clean. (c) *Denoised.* (e)~(f) *The output of branch* ①~③ in the *TSDT-SumAll model shown in Figure 1(a)*.

3.2. Dataset

The speech enhancement models are trained on the DNS Challenge dataset[30], a benchmark corpus containing 500 hours of clean speech from 2,150 speakers and 180 hours of noise. During training, noisy mixtures are dynamically generated with signal-to-noise ratios (SNRs) spanning -5 dB to 30 dB. Prior to each epoch, 20% of clean utterances are convolved with randomized room impulse responses (RIRs) from [31]. Evaluation employs two official non-blind test subsets (reverberant/non-reverberant), each containing 150 noisy-clean pairs. DNSMOS scores are reported on 700 real-world 10-second noisy recordings.

3.3. Evaluation Metrics

The output of the speech enhancement algorithm is evaluated in terms of: (1) wide-band PESQ (dubbed PESQ)[32]; (2) eSTOI [33]; (3) Scale-invariant signal-to-noise ratio (SISNR) and (4) DNSMOS [34]. For all evaluation metrics, higher values denote better enhancement performance.

4. Results

Table 1 presents speech enhancement results for all tested models on the DNS test set, with the top two entries in each metric for larger and compact models are highlighted in bold.

Among models with comparable computational complexity, two-stage TSDT architectures demonstrate clear superiority over single-stage variants, particularly under no-reverb conditions. For larger models, the TSDT-SumAll achieves superior PESQ (2.68) and DNSMOS OVL (3.55), outperforming RNN-based Inf-Fullsubnet (2.55/3.46), Fast-subnet (2.42/3.40), and TSDT_1stNet (2.55/2.53). The CNN-based NS-DeepVQE achieves competitive OVL (3.52) but lower PESQ (2.57), with reduced noise suppression capability (lower BAK scores) in compact configurations.

Figure 2 illustrates spectrograms of the three output branches in the TSDT-SumAll model. The second branch compensates for low-frequency energy loss in Branch 1, while Branch 3 rectifies unvoiced speech degradation observed in Branch 1. This demonstrates that residual connections facilitate task-specific collaboration between dual-stage networks.

For compact models, TSDT-SumAll maintains dominance with PESQ=2.55 and OVL=3.48, surpassing Inf-Fullsubnet (2.46/3.42), TSDT_1stNet (2.33/3.44), and NS-DeepVQE (2.21/3.28). The performance gap highlights the critical role of residual connections integrating the TAC-based second stage, as TSDT-SumAll significantly outperforms the single-stage TSDT_1stNet under matched computational budgets.

Under reverberant conditions, TSDT_1stNet achieves relatively higher scores across both model scales, potentially due to more residual reverberation after enhancement.

Branch fusion analysis demonstrates that combining Branches 1 and 2 yields the lowest PESQ scores (2.58 for larger, 2.40 for compact models), highlighting Branch ③'s critical role in compensating for Mel-band resolution loss while preserving high-frequency granularity through raw spectral fidelity. TSDT-SumAll model achieves optimal metric consistency across all evaluation scenarios.

5. Conclusion

We propose TSDT-Net, an ultra-low complexity architecture designed for embedded platforms with constrained computational resources. The framework reduces computational costs through a simplified DPT architecture while enhancing performance via residual-connected second-stage processing with TAC modules. Experimental results demonstrate that TSDT-Net outperforms single-stage RNN and U-Net architectures of comparable complexity. Future research directions will explore the integration of the proposed model with complementary speech enhancement frameworks (e.g., DeepfilterNet) and investigate its potential extensions to cross-domain adaptation scenarios.

6. References

[1] K. Tan and D. Wang, “Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6865–6869.

[2] K. Tan and D. Wang. “Learning Complex Spectral Mapping with Gated Convolutional Recurrent Networks for Monaural Speech Enhancement.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 380-390.

[3] 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 – 21th Annual Conference of the International Speech Communication Association, 2020.

[4] X. Le, H. Chen, K. Chen, and J. Lu, “DPCRN: Dual-path convolution recurrent network for single channel speech enhancement,” in Proc. INTERSPEECH 2021 – 22th Annual Conference of the International Speech Communication Association, 2021.

[5] X. Hao, X. Su, R. Horaud, and X. Li, “Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6633–6637.

[6] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, “Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7857– 7861.

[7] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, W. Chen, “Masking and Inpainting: A Two-Stage Speech Enhancement Approach for Low SNR and Non-Stationary Noise,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6959-6963.

[8] Andong Li, Chengshi Zheng, Lu Zhang, and Xiaodong Li, “Glance and gaze: A collaborative learning framework for single-channel speech enhancement,” Applied Acoustics, vol. 187, 2022.

[9] Y. Ju, J. Chen, S. Zhang, S. He, W. Rao, W. Zhu, Y. Wang, T. Yu, and S. Shang, “Tea-pse 3.0: Tencent-ethereal-audio-lab personalized speech enhancement system for icassp 2023 dns-challenge,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–2

[10] Y. Zhang, H. Zou, and J. Zhu, “A two-stage framework in crossspectrum domain for real-time speech enhancement,” in ICASSP 2024- 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 12587–12591

[11] W. Liu, Y. Shi, J. Chen, W. Rao, S. He, A. Li, Y. Wang, and Z. Wu, “Gesper: A Restoration-Enhancement Framework for General Speech Reconstruction,” in Proc. INTERSPEECH 2023 – 24th Annual Conference of the International Speech Communication Association, 2023, pp. 4044–4048

[12] M. Liu, Z. Chen, X. Yan, Y. Lv, X. Xia, C. Huang, Y. Xiao, L. Xie, “RaD-Net 2: A causal two-stage repairing and denoising speech enhancement network with knowledge distillation and complex axial self-attention,” Proc. Interspeech 2024, 1700-1704

[13] L. Meng, J. Coldenhoff, P. Kendrick, T. Stojkovic, A. Harper, K. Ratmanski and M. Cernak. “On Real-Time Multi-Stage Speech Enhancement Systems.” ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024.

[14] F. Dang, H. Qi, and P. Zhang. "THLNet: two-stage heterogeneous lightweight network for monaural speech enhancement." arXiv preprint arXiv:2301.07939, 2023.

[15] A. Li, G. Yu, C. Zheng, W. Liu, and X. Li, “A general unfolding speech enhancement method motivated by Taylor’s theorem,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.

[16] A. Li, S. You, G. Yu, C. Zheng, and X. Li, “Taylor, can you hear me now? a Taylor-unfolding framework for monaural speech enhancement,” in Proc. IJCAI, 2022, pp. 4193–4200.

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

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

[19] J. Yu, Y. Luo, H. Chen, R. Gu, and C. Weng, “High fidelity speech enhancement with band-split rnn,” in Proc. Interspeech 2023, pp.2483-2487

[20] H. Chen, J. Yu, and C. Weng, “Complexity scaling for speech denoising,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Korea, 2024

[21] H. Chen, J. Yu, Y. Luo, R. Gu, W. Li, Z. Lu, and C. Weng, “Ultra dual-path compression for joint echo cancellation and noise suppression,” in Proc. INTERSPEECH 2023 – 24th Annual Conference of the International Speech Communication Association, 2023.

[22] H. Xiang, and X. Li. "Fast FullSubNet: Accelerate Full-band and Sub-band Fusion Model for Single-channel Speech Enhancement." *arXiv preprint arXiv:2212.09019*, 2022.

[23] H. Schr¨oter, T. Rosenkranz, and A. Maier, “DeepFilterNet: Perceptually Motivated Real-time Speech Enhancement,” in INTERSPEECH, 2023

[24] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, “Transformers are rnns: Fast autoregressive transformers with linear attention,” in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, pp. 5156–5165

[25] Y. Luo, Z. Chen, N. Mesgarani and T. Yoshioka, "End-to-end Microphone Permutation and Number Invariant Multi-channel Speech Separation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 6394-6398

[26] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in: International conference on machine learning, PMLR, 2017, pp. 933–941. 4

[27] E. Indenbom, N.-C. Ristea, A. Saabas, T. Parnamaa, J. Guzvin, and R. Cutler, “DeepVQE: Real Time Deep Voice Quality Enhancement for Joint Acoustic Echo Cancellation, Noise Suppression and Dereverberation,” in Interspeech, Dublin, Ireland, August 2023, pp. 3819–3823.

[28] S. Braun, H. Gamper, C. K. Reddy, I. Tashev, “Towards efficient models for real-time deep noise suppression”. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp.656-660

[29] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, “Personalized speech enhancement: New models and comprehensive evaluation,” in ICASSP, 2022, pp.356-360

[30] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun et al., “The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results,” in Proc. INTERSPEECH 2020 – 21th Annual Conference of the International Speech Communication Association, 2020.

[31] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, “A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research,” EURASIP Journal on Advances in Signal Processing, vol. 2016, no. 1, pp. 7

[32] 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.

[33] J. Jensen, and C. H. Taal. “An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers,” IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2016, vol.24. pp. 2009-2022

[34] C. K. Reddy, V. Gopal, and R. Cutler, “DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.