Beyond Performance Plateaus: A Comprehensive Study on Scalability in Speech Enhancement

Wangyou Zhang[1,2], Kohei Saijo[3], Jee-weon Jung[2], Chenda Li[1,2], Shinji Watanabe[2], Yanmin Qian[1]

[1]Shanghai Jiao Tong University, China   [2]Carnegie Mellon University, USA
[3]Waseda University, Japan

wyz-97@sjtu.edu.cn

Abstract

Deep learning-based speech enhancement (SE) models have achieved impressive performance in the past decade. Numerous advanced architectures have been designed to deliver state-of-the-art performance; however, their scalability potential remains unrevealed. Meanwhile, the majority of research focuses on small-sized datasets with restricted diversity, leading to a plateau in performance improvement. In this paper, we aim to provide new insights for addressing the above issues by exploring the scalability of SE models in terms of architectures, model sizes, compute budgets, and dataset sizes. Our investigation involves several popular SE architectures and speech data from different domains. Experiments reveal both similarities and distinctions between the scaling effects in SE and other tasks such as speech recognition. These findings further provide insights into the under-explored SE directions, e.g., larger-scale multi-domain corpora and efficiently scalable architectures.

Index Terms: speech enhancement, scalability, robustness, generalizability

1. Introduction

Speech enhancement (SE) is the task of removing undesired signals from the input speech [1], which often includes noise and reverberation. The last decade has witnessed significant progress in deep learning-based SE approaches in both frequency and time domains. The state-of-the-art (SOTA) SE approaches often feature a carefully designed architecture with sophisticated interactions between different features. Typical architectures of SE models include the U-Net structure [2, 3], dilated convolution-based structure [4], and dual-path structure with recurrent neural network (RNN) or attention [5–8].

While a wide variety of SE models have been developed in recent years and demonstrate surprisingly strong performance on specific datasets, there remain several issues. 1) Most SE research has been conducted on small-scale datasets (e.g., Voice-Bank+DEMAND [9]) with limited diversity. The limited data amount hinders us from understanding the scalability of the SE approach. The limited evaluation condition also makes it difficult to understand the generalizability and robustness in realistic applications. 2) While it has been commonly observed that more parameters and higher complexity often result in better performance, the scaling effect of model complexity along with data scales has not been well studied in the SE literature. In particular, it is still unclear whether a large amount of data and high-complexity models are the best way to achieve the best performance. 3) Compared to the recent trend of large models showing unprecedented performances in automatic speech recognition (ASR) and large language models (LLMs) [10–13], this area of research is under-explored in SE. This calls for an extensive investigation into the scaling law of SE models as an important step in building a large SE foundation model.

As the first step to address the above issues, in this paper, we aim to explore the scalability of single-channel SE models. Following existing studies on scaling laws in ASR [14, 15], natural language processing [16], and computer vision [17], we explore several typical scaling factors. Specifically, they include model architectures, model sizes, compute budgets, and dataset sizes. We further take the model setup (causal or non-causal) into account to encompass both real-time and offline applications.

While our investigation is driven by motivations akin to [18], it distinguishes itself through several novel findings and complements the existing work in five aspects: 1) larger model complexity; 2) extensive investigation of both causal and non-causal setups; 3) multiple data scales; 4) extended coverage of SE sub-tasks; and 5) comprehensive multi-domain evaluation. Our key contributions are summarized below:

1) By covering a wider range of model complexity, we reveal the *distinct potentials of different SE architectures in low and high complexity regions*. Among them, BSRNN and TF-GridNet show exceptional scalabilities respectively in low and (relatively) high complexity ranges.

2) We show that *scaling model complexity along with multi-domain data sizes consistently improves all metrics*. This aligns with the findings in large-scale ASR [11, 12] and LLMs [10].

3) SE models, compared to ASR and LLMs, *particularly suffer from data mismatch* due to the limitations of simulation-based training data. As a result, largely increasing simulated data based on the same source corpora is not as effective as increasing data diversity. Although we successfully scaled the data up to 157 h, further scaling up encountered various issues such as scarcity of high quality data, limited diversity, and highly imbalanced data distribution. This calls for the construction of larger-scale multi-domain SE corpora, as we are still far away from the data scale in ASR research (e.g., 1M h).

4) Although RNN-based architectures [8, 19] have excelled in popular SE benchmarks, they are highly inefficient in scaling up, demanding much larger compute due to limited parallelization. *We are still lacking an SE counterpart to transformers in ASR and LLMs that have well-recognized scalability and outstanding performance.*

5) Finally, we show that model parameters are not very informative when comparing different architectures, as they usually have distinct parameter efficiency. Instead, *model complexity in terms of computational costs (#MACs) should be compared*.

2. Model and experimental design

2.1. Models

We aim to investigate SOTA SE models with diverse architectures to gain insights into their potential scalability. Specifically,

Table 1: *Configurations of BSRNN models[1].*

Model	Causal	#Params (M)	#MACs (G/s)	
			16 kHz	48 kHz
BSRNN		(sampling-frequency-independent)		
xtiny	✓	0.5	0.1	0.4
	✗	0.5	0.2	0.6
tiny	✓	1.3	0.6	1.7
	✗	1.5	0.7	2.2
small	✓	4.1	2.1	6.4
	✗	4.8	2.8	8.5
medium	✓	14.3	8.4	25.2
	✗	16.9	11.2	33.4
large	✓	52.9	33.4	99.9
	✗	63.1	44.3	132.5
xlarge	✓	83.6	66.1	197.7
	✗	104.1	87.9	262.3

Table 2: *Configurations of other SE models[1].*

Model	Causal	#Params (M)	#MACs (G/s)	
			16 kHz	48 kHz
Conv-TasNet		(input is always resampled to 48 kHz)		
small	✗	1.1	-	8.9
medium	✗	14.3	-	18.7
large	✗	52.6	-	47.2
xlarge	✗	103.9	-	85.4
DEMUCS-v4		(input is always resampld to 48 kHz)		
tiny	✗	1.0	-	1.0
small	✗	4.1	-	3.5
medium	✗	16.2	-	13.0
large	✗	26.9	-	17.2
xlarge	✗	79.3	-	40.7
TF-GridNet		(sampling-frequency-independent)		
xxtiny	✗	0.1	1.9	5.6
xtiny	✗	0.5	7.4	21.7
tiny	✗	1.5	24.1	70.5
small	✗	5.7	89.5	261.8

we experiment with band-split RNN (BSRNN) [19, 20], Conv-TasNet [4], DEMUCS-v4 [3], and TF-GridNet [8]. We briefly introduce each model below, while their model configurations are presented in Tables 1 and 2.

BSRNN [19] is a dual-path T-F domain SE model that can handle different sampling frequencies (SF). It reduces the frequency dimension with a hand-crafted band division to lower the complexity. Each T-F bin is projected to a high-dimensional embedding for RNN-based dual-path modeling. We adopt the improved design in [20] which combines masking and mapping to obtain the enhanced spectrum. In [18], BSRNN has shown superior performance over other architectures across different complexity. Therefore, we first conduct extensive experiments with this model to investigate the scaling effect with different model setups (causal or non-causal) as shown in Table 1.

Conv-TasNet [4] is a time-domain SE model, which consists of convolution-based encoder/decoder and stacked temporal convolutional networks (TCN). It features small kernel/stride sizes in the learnable encoder/decoder and a large receptive field thanks to the stacked TCNs.

DEMUCS-v4 [3] is a hybrid time and time-frequency (T-F) domain SE model with a U-Net architecture. It consists of two parallel branches, i.e., time-domain U-Net and T-F domain U-Net. The bottleneck features from both branches are fused via cross-attention in a cross-domain transformer encoder, and both branches' outputs are summed to obtain the enhanced speech.

TF-GridNet [8] is the SOTA single-channel SE approach based on T-F dual-path modeling. It also leverages RNNs for dual-path modeling and further enhances the sequence modeling capability by utilizing adjacent frames/frequency bins. In addition, a cross-frame self-attention module is inserted after each dual-path modeling block to better exploit the global information. In our preliminary experiments, we noticed that the original TF-GridNet already has very high computational costs, making it expensive to scale up. Therefore, we focus on smaller-sized configurations for this model as shown in Table 2.

To allow the processing of different SFs, the simple solution is to always resample the input signal to 48 kHz for all SE models regardless of their original SF f_{org}. The output is then downsampled back to f_{org} for generating the enhanced output as well as for loss computation. But specifically for T-F dual-path models (BSRNN and TF-GridNet), we adopt a different approach by applying the sampling-frequency-independent (SFI) design as proposed in [21]. It uses adaptive short-time Fourier transform

(STFT) window/hop sizes to handle different SFs without the need for resampling. During training, we augment each training sample by randomly downsampling it to a new SF $f_{\text{org}}^{\text{aug}}$ in $\{8, 16, 24, 32, 44.1, 48\}$ kHz. It assists the trained SE models in generalizing to different SFs. The SF flows of both non-SFI and SFI models are summarized below:

$$f_{\text{org}} \xrightarrow{\text{augment}} f_{\text{org}}^{\text{aug}} \xrightarrow{\text{upsample}} 48 \text{ kHz} \xrightarrow[\text{model}]{\text{non-SFI}} 48\text{kHz} \xrightarrow{\text{downsample}} f_{\text{org}}^{\text{aug}}, \quad (1)$$

$$f_{\text{org}} \xrightarrow{\text{augment}} f_{\text{org}}^{\text{aug}} \xrightarrow[\text{model}]{\text{SFI}} f_{\text{org}}^{\text{aug}}. \quad (2)$$

The L1-based time-domain plus frequency-domain multi-resolution loss [22] is adopted to train all SE models. We use four STFT window sizes $\{256, 512, 768, 1024\}$ when calculating the multi-resolution loss. We use the ESPnet-SE toolkit [23, 24] for all experiments. All models are trained until convergence (up to 3M iterations)[1].

2.2. Data details

As shown in Table 3, we combine three datasets from different domains to train all aforementioned SE models, i.e., VoiceBank+DEMAND [9], DNS-2020 challenge data [25], and WHAMR! [26]. They cover various conditions including noise, reverberation, and different SFs. We investigate three different data scales $\{8.8, 98.8, 156.8\}$ h by combining the listed datasets.

For evaluation, we combine five different test sets as listed in Table 4. The first three correspond to the test sets of the corpora mentioned above, representing the matched conditions. The rest two are challenge datasets from CHiME-4 [27] and RE-VERB [28], which provide both simulation and real-recorded data for evaluation. They represent mismatched conditions as they contain unseen speech/noise/reverberation during training.

2.3. Evaluation metrics

We adopt the following evaluation metrics to analyze the scaling effect from multiple views: PESQ-WB [29], short-time objective intelligibility (STOI) [30], signal-to-distortion ratio (SDR) [31], DNSMOS [32] and word accuracy (WAcc)[2]. The first three metrics are intrusive objective measures that require well-aligned reference signals. Therefore, they are only calculated on simulated test samples. The DNSMOS OVRL score is a non-intrusive metric predicted by a pre-trained neural network that does not require the reference signal as input. The WAcc

---

[1]Detailed hyperparameter configurations of all models are available at `https://github.com/Emrys365/se-scaling`.
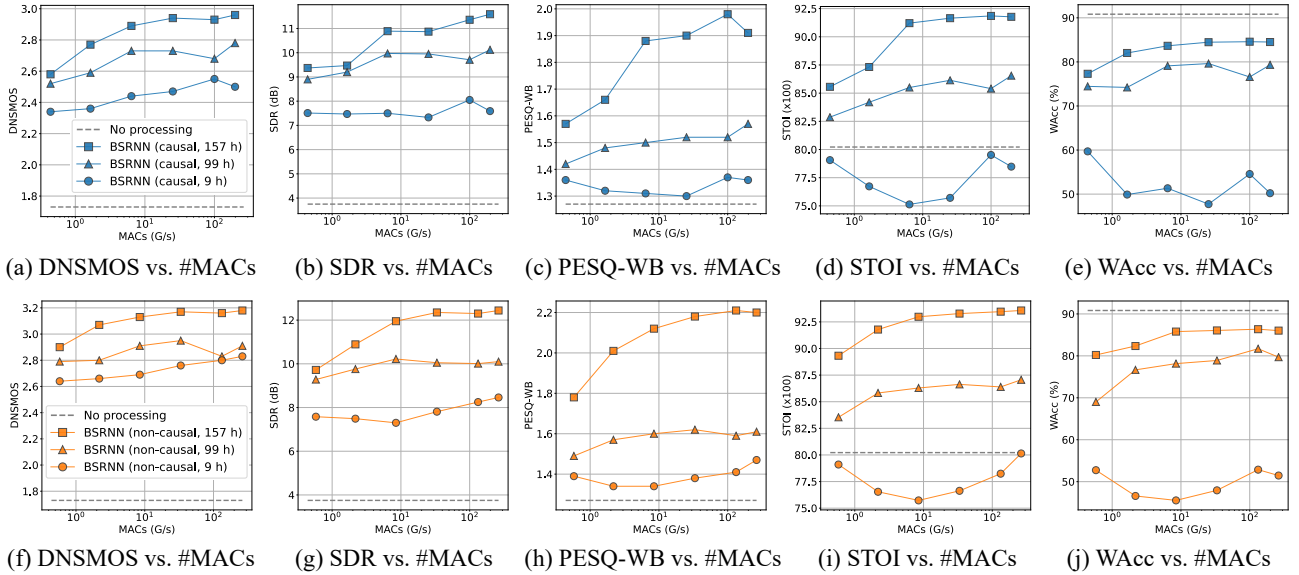
[2]WAcc is equal to $1-$ word error rate (WER).

Figure 1: *Scaling effect of BSRNN with respect to model complexity (#MACs at 48 kHz). Each data point corresponds to an independent model. Causal model setups: (a)–(e). Non-causal model setups: (f)–(j).*

Table 3: *Details of training data.*

No.	Dataset	Condition	Duration training	Duration validation
1	VCTK+DEMAND [9]	Noisy, 48 kHz	8.8 h	0.6 h
2	DNS-2020 [25]	Noisy, 16 kHz	90.0 h	10.0 h
3	WHAMR! [26]	Noisy, reverberant, 16 kHz	58.0 h	14.7 h
Scale 1: No. 1 (~9 h)		**Scale 2**: Nos. 1 + 2 (~99 h)		
Scale 3: Nos. 1 + 2 + 3 (~157 h)				

Table 4: *Details of combined evaluation data.*

No.	Dataset	Condition	Duration
1	VCTK+DEMAND [9]	Noisy, 48 kHz	0.6 h
2	DNS-2020 [25]	Noisy, 16 kHz	0.4 h (no reverb)
3	WHAMR! [26]	Noisy, reverberant, 16 kHz	9.0 h
4	CHiME-4 [27]	Noisy, 16 kHz	2.3 h (Simu) + 2.2 h (Real)
5	REVERB [28]	Reverberant, 16 kHz	4.8 h (Simu) + 0.7 h (Real)

is evaluated using the open-source large-scale pre-trained ASR model — OWSM v3.1 [12]. For all metrics, a higher value indicates better performance.

3. Results

3.1. Causal vs. non-causal: a case study on BSRNN

We first study the scaling effect of non-causal and causal models based on BSRNN, as it has shown superior performance in [18]. We train BSRNN models with different configurations as listed in Tables 1 and 3, and then evaluate them on all test data shown in Table 4. All results presented below represent averages across all test corpora, unless specified otherwise.

Fig. 1 shows the scaling effect of causal (above) and non-causal (below) models with respect to model complexity which is represented by #MACs at 48 kHz. It is evident that both causal and non-causal models exhibit similar trends across different metrics. In particular, the common observations include:

- *Increasing the model complexity generally improves all metrics when sufficient training data are available.* This aligns with the observation in [18]. One major exception is the SDR, PESQ-WB, and STOI metrics on the smallest data scale (9 h), which degrade initially and then start to improve. The initial degradation can be attributed to overfitting due to limited training data. Combined with the latter, this phenomenon is known as the "double descent" [33, 34], which shows that

modern deep architectures can recover from overfitting by further increasing model parameters. However, it has been rarely reported in the area of speech processing. In our context, this observation implies that the BSRNN with model complexity higher than 10 G/s has more than enough capacities of modeling the 9 h training data. This might also apply to other SE architectures. These results may suggest that SE research should focus more on larger corpora to avoid overkill on small datasets such as VCTK+DEMAND [9] with high-complexity SE models.

- *With more training data, the model performance tends to benefit more from increased model complexity.* The models trained on 157 h data gain more performance improvement than other models.

- *When the model complexity reaches a certain threshold, the initially linear scaling relationship between performance and log-scale #MACs tend to converge.* The initially linear scaling part coincides with the observation in [18], which only investigates the model complexity between 50 M/s and 15 G/s. Our study reveals that this scaling effect does not hold when we further increase the model complexity, implying that more training data is needed when further scaling up the model.

- *More complex models tend to achieve better downstream ASR performance (WAcc), but they still underperform the baseline without SE (no processing).* Similar issues are also reported in prior works [35, 36], where the degradation is attributed to the artifacts introduced by SE models on mismatched data.

Comparing causal and non-causal models in Fig. 1, we can also see that non-causal models consistently outperform causal models in all metrics. The performance difference is especially enlarged when more training data (e.g., 157 h) are available, indicating a much higher model capacity in non-causal models.

Furthermore, we analyze the scaling effect of model performance with respect to dataset sizes. In Fig. 2, we only present the results of non-causal models in terms of SDR and WAcc metrics, and omit the other results as they share a similar trend. We can see that *all metrics keep improving monotonically with more training data*, which is in line with common observations. Moreover, larger models tend to benefit more from increased
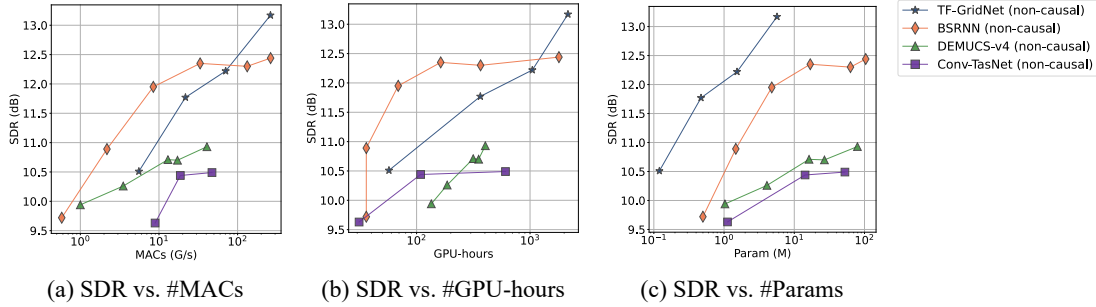
(a) SDR vs. #MACs   (b) SDR vs. #GPU-hours   (c) SDR vs. #Params

Figure 3: *Scaling effect of non-causal SE models with different architectures on 157 h of training data.*

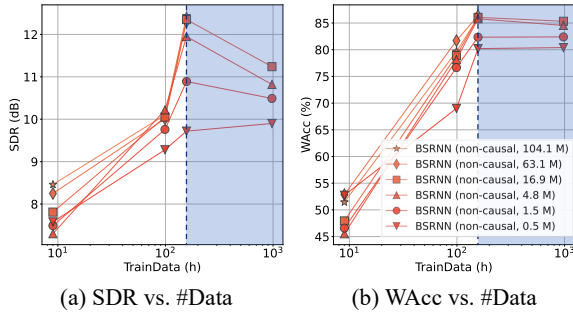(a) SDR vs. #Data   (b) WAcc vs. #Data

Figure 2: *Scaling effect of non-causal BSRNN with respect to dataset sizes (#Data). Main results are shown in white regions, while shaded regions illustrate the degradation caused by improper data scaling (detailed in § 3.1).*

data, indicating that a higher model capacity allows for better utilization of the data. *Note that this does not imply that simply increasing data is always helpful.* In fact, when we attempted to further increase the dataset size by simply simulating more DNS-2020 data on top of "scale 3" in Table 3 (from 157 h to ~900h), the overall performance was even degraded. This is shown in the shaded regions in Fig. 2. The major cause roots in the highly imbalanced data distribution, with 93% DNS-2020, 1% VCTK+DEMAND, and 6% WHAMR! data. The trained SE model is thus biased towards DNS-2020 data and suffers from mismatch during evaluation. A similar observation has also been reported in [37], where increasing mismatched pre-training data does not always improve the downstream finetuning performance. This implies the importance of well-balanced and multi-domain data when scaling up the training dataset, which has often been overlooked in prior SE studies.

3.2. Comparison of different SE architectures

Next, we investigate the scaling effect of different SE architectures. Given the findings from BSRNN in the previous section, we mainly focus on the 157 h data scale and non-causal model setup for the best performance. In Fig. 3, we explore three different scaling factors, i.e., model complexity (#MACs at 48 kHz), model parameters (#Params), and compute budget (#GPU-hours).

In Fig. 3 (a), we observe that BSRNN outperforms others in the low model complexity region (<100 G/s), while TF-GridNet dominates the higher model complexity region. They are both T-F domain dual-path models, the recently developed SOTA architecture. The U-Net-based DEMUCS-v4 models performs better than TCN-based Conv-TasNet, but still lags behind BSRNN. This implies that T-F domain dual-path models have great potential in both low and high-complexity configurations, where BSRNN and TF-GridNet can be chosen depending on the complexity budget. However, *no single architecture in*

our investigation can dominate all complexity conditions.

In Fig. 3 (b), we present the scaling effect of SDR performance with respect to the compute budget. Here, we use RTX A5000 GPUs for training TF-GridNet and Tesla V100 GPUs for other models[3]. It is shown that BSRNN initially has very high compute efficiency, corresponding to the low model complexity region in Fig. 3 (a). In contrast, TF-GridNet's high performance comes at a high cost, requiring more computing than other architectures even for its smallest configuration (with only 0.5 M parameters). Although these RNN-based architectures show excellent performance, they are highly inefficient in scaling up due to the limited parallelization. *This calls for more explorations on efficiently scalable architectures that can be an SE counterpart to transformers in ASR and LLMs.*

Finally in Fig. 3 (c), we show that the usually reported model parameters do not have a clear correlation with the SE performance when comparing different model architectures. This is attributed to different SE architectures having highly divergent parameter efficiency. Therefore, we strongly recommend that *SE studies should provide model complexity as an informative indicator instead of solely model parameters.* We also note that Fig. 3 lacks sufficient explorations in the large parameter region (≥ 100 M), which we leave for future work.

4. Conclusion

In this paper, we have investigated the scalability of SE models in terms of model architectures, model setups (causal or non-causal), model complexity and sizes, compute budget, and dataset sizes, where we consumed >25k GPU hours in total. Our experiments on combined public corpora verified the great potential of two T-F domain dual-path models (BSRNN and TF-GridNet) respectively in both low and high model complexity conditions. However, these RNN-based architectures are highly inefficient in scaling up due to the limited parallelization. This suggests that more exploration in the SE architecture is needed to achieve better efficiency and scalability. In addition, the "double descent" phenomenon on VCTK+DEMAND implies that a small corpus is not suitable to benchmark high-capacity SE models. Our findings in § 3.1 further show the necessity of gathering large-scale multi-domain SE corpora to facilitate the scaling of SE models while mitigating mismatches. We expect the new insights gained from our investigation can inspire more research towards scalable SE models.

---

[3]The equivalent GPU hours using Tesla V100 for TF-GridNet should be even larger since RTX A5000 has faster computing speed.

5. Acknowledgment

6. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[2] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 334–340.

[3] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proc. IEEE ICASSP*. IEEE, 2023, pp. 1–5.

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

[5] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE ICASSP*, 2020, pp. 46–50.

[6] 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.

[7] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE ICASSP*, 2021, pp. 21–25.

[8] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim *et al.*, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Trans. ASLP.*, vol. 31, pp. 3221–3236, 2023.

[9] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Proc. Interspeech*, 2016, pp. 352–356.

[10] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[12] Y. Peng, J. Tian, W. Chen, S. Arora, B. Yan *et al.*, "OWSM v3. 1: Better and faster open Whisper-style speech models based on E-Branchformer," *arXiv preprint arXiv:2401.16658*, 2024.

[13] H. Wu, X. Chen, Y.-C. Lin, K. wei Chang, H.-L. Chung *et al.*, "Towards audio language modeling – an overview," *arXiv preprint arXiv:2402.13236*, 2024.

[14] J. Droppo and O. Elibol, "Scaling laws for acoustic models," in *Proc. Interspeech*, 2021, pp. 2576–2580.

[15] Y. Gu, P. Gurunath Shivakumar, J. Kolehmainen, A. Gandhe, A. Rastrow *et al.*, "Scaling laws for discriminative speech recognition rescoring models," in *Proc. Interspeech*, 2023, pp. 471–475.

[16] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess *et al.*, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[17] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 104–12 113.

[18] H. Chen, J. Yu, and C. Weng, "Complexity scaling for speech denoising," in *Proc. IEEE ICASSP*, 2024, pp. 12 276–12 280.

[19] J. Yu and Y. Luo, "Efficient monaural speech enhancement with universal sample rate band-split RNN," in *Proc. IEEE ICASSP*, 2023.

[20] J. Yu, H. Chen, Y. Luo, R. Gu, and C. Weng, "High fidelity speech enhancement with band-split RNN," in *Proc. Interspeech*, 2023, pp. 2483–2487.

[21] W. Zhang, K. Saijo, Z.-Q. Wang, S. Watanabe, and Y. Qian, "Toward universal speech enhancement for diverse input conditions," in *Proc. IEEE ASRU*, 2023.

[22] Y.-J. Lu, S. Cornell, X. Chang, W. Zhang, C. Li *et al.*, "Towards low-distortion multi-channel speech enhancement: The ESPnet-SE submission to the L3DAS22 challenge," in *Proc. IEEE ICASSP*, 2022, pp. 9201–9205.

[23] C. Li, J. Shi, W. Zhang, A. S. Subramanian, X. Chang *et al.*, "ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration," in *Proc. IEEE SLT*, 2021, pp. 785–792.

[24] Y.-J. Lu, X. Chang, C. Li, W. Zhang, S. Cornell *et al.*, "ESPnet-SE++: Speech enhancement for robust speech recognition, translation, and understanding," in *Proc. Interspeech*, 2022, pp. 5458–5462.

[25] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020, pp. 2492–2496.

[26] M. Maciejewski, G. Wichern, E. McQuinn, and J. Le Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proc. IEEE ICASSP*, 2019, pp. 696–700.

[27] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[28] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE WASPAA*, 2013, pp. 1–4.

[29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, vol. 2, 2001, pp. 749–752.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. ASLP.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[31] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[32] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE ICASSP*, 2022, pp. 886–890.

[33] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[34] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak *et al.*, "Deep double descent: Where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, 2021.

[35] S. E. Eskimez, X. Wang, M. Tang, H. Yang, Z. Zhu *et al.*, "Human listening and live captioning: Multi-task training for speech enhancement," in *Proc. Interspeech*, 2021, pp. 2686–2690.

[36] K. Iwamoto, T. Ochiai, M. Delcroix, R. Ikeshita, H. Sato *et al.*, "How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR," in *Proc. Interspeech*, 2022, pp. 5418–5422.

[37] B. Isik, N. Ponomareva, H. Hazimeh, D. Paparas, S. Vassilvitskii *et al.*, "Scaling laws for downstream task performance of large language models," *arXiv preprint arXiv:2402.04177*, 2024.