



BS-PLCNet 2: Two-stage Band-split Packet Loss Concealment Network with Intra-model Knowledge Distillation

Zihan Zhang^{1,2}, Xianjun Xia², Chuanzeng Huang², Yijian Xiao², Lei Xie^{1*}

¹Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China

²ByteDance, China

zhzhang@mail.nwpu.edu.cn, lxie@nwpu.edu.cn

Abstract

Audio packet loss is an inevitable problem in real-time speech communication. A band-split packet loss concealment network (BS-PLCNet) targeting full-band signals was recently proposed. Although it performs superiorly in the ICASSP 2024 PLC Challenge, BS-PLCNet is a large model with high computational complexity of 8.95G FLOPS. This paper presents its updated version, BS-PLCNet 2, to reduce computational complexity and improve performance further. Specifically, to compensate for the missing future information, in the wide-band module, we design a dual-path encoder structure (with non-causal and causal path) and leverage an intra-model knowledge distillation strategy to distill the future information from the non-causal teacher to the casual student. Moreover, we introduce a lightweight post-processing module after packet loss restoration to recover speech distortions and remove residual noise in the audio signal. With only 40% of original parameters in BS-PLCNet, BS-PLCNet 2 brings 0.18 PLCMOS improvement on the ICASSP 2024 PLC challenge blind set, achieving state-of-the-art performance on this dataset.

Index Terms: packet loss concealment, band-split, two-stage, intra-model knowledge distillation

1. Introduction

In the past few years, the advancement of the internet and technologies coupled with the pandemic have gradually made voice over internet phone (VoIP) technology an essential part of our daily life [1]. However, due to various factors, such as network congestion, delay and jitter, hardware failures, audio data packets may get lost during transmission, which degrades the speech quality [2]. Packet loss concealment (PLC) technology is applied to restore the missing content, thus improving the speech quality and intelligibility.

In contemporary speech codecs, traditional interpolation [3] or Hidden Markov Model (HMM) based approaches [4] are typically employed for PLC. While these approaches only work well in scenarios with short-time lost, and can not handle conditions of high packet loss rate and long-gap lost, leading to a significant degrade in speech quality. In recent years, with the development of deep learning, methods based on deep neural networks (DNN) have shown great potential in PLC. The early deep PLC approach [5] used log-spectra to train a Time-Frequency (T-F) domain DNN model as a nonlinear regression function for PLC. As T-F domain approaches need to recover the phase information to reconstruct the waveform, the performance may be limited [6]. More recent approaches show a preference for using time-domain models [2, 6, 7, 8]. Lin et al. [2]

propose a time-domain convolutional recurrent network (CRN) for online packet loss concealment. Westhausen et al. [7] predict lost frames from a short context buffer in a sequence-to-one (seq2one) fashion by only performing the inference. Valin et al. [8] implement PLC in the feature domain. They use a predictive conditioning model to predict the acoustic features and a generative vocoder to synthesis speech. Li et al. [6] propose a time domain generative adversarial network (GAN) structure, using discriminator and multi-loss to ensure both speech quality and automatic speech recognition (ASR) compatibility.

Recently, audio transmission has transitioned to full-band signals (48KHz). However, due to the latency constraint and high computational cost, the aforementioned approaches work only at wide-band signals (16kHz). The latest ICASSP 2024 PLC Challenge promotes full-band real-time PLC and focuses on more challenging long-gap packet lost. Among all the submissions, BS-PLCNet [9] with a band-split gated convolutional recurrent neural network leads to superior performance on the blind test set, showing the effectiveness of T-F domain models. However, it has a high computational complexity of 8.95G FLOPS. Additionally, the common encoder-decoder structure in BS-PLCNet fails to learn a robust feature structure as future information is missing.

In this paper, we propose BS-PLCNet 2 to further improve the PLC performance and reduce the computational complexity. There are three contributions in this paper:

- We optimize the network structure of BS-PLCNet for superior feature extraction and replace the original convolution operation with depth-wise separable convolution [10].
- Knowledge distillation strategy is adapted. We upgrade the model with intra-model knowledge distillation [11], which allows us to obtain the final model without training the teacher and student model separately.
- We incorporate a post-processing module and utilize a two-stage training strategy to recover speech distortions and remove residual noise in the audio signal.

Experimental results show the proposed BS-PLCNet 2 outperforms BS-PLCNet with a 0.18 OVRL PLCMOS [12] improvement on the ICASSP 2024 PLC Challenge blind set and achieves state-of-the-art (SOTA) results. Remarkably, the non-causal model only has 38.1% and 40% of original computational cost and number of parameters, respectively.

2. Proposed Method

2.1. Problem formulation

Let $s \in \mathbb{R}^L$ be an original speech, where L is the length of the signal. Considering the potential additive noise $v \in \mathbb{R}^L$, the transmission signal x can be represented as $x = s + v$. For

*: Corresponding author.

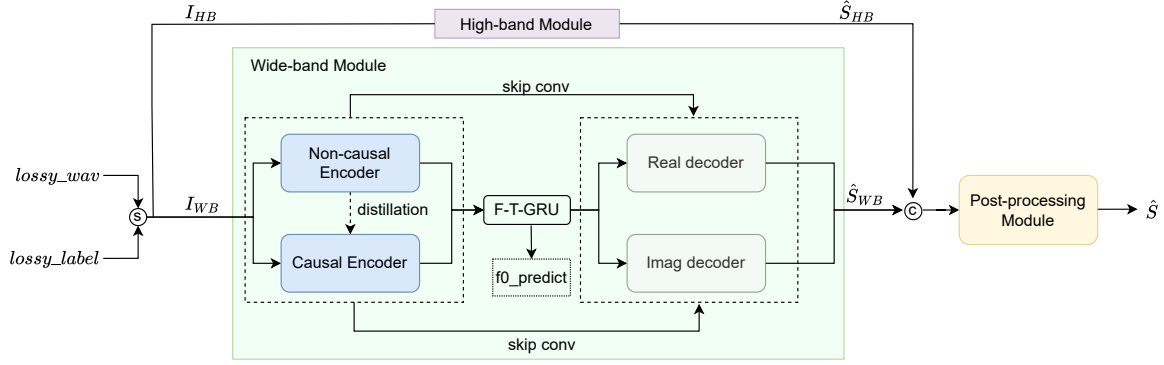


Figure 1: The proposed BS-PLCNet 2.

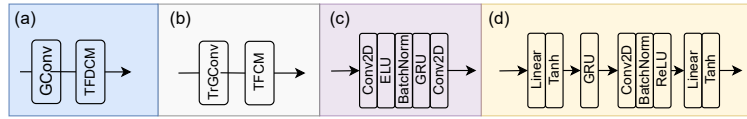


Figure 2: Details of the encoder layer (a), the decoder layer (b), the high-band module (c) and the post-processing module (d).

audio transmission, the continuous audio x is often split into short-time frames with the frame length N . The k th speech frame can be defined as:

$$x_k(n) = x(N \cdot (k - 1) + n), n = 0, 1, \dots, N - 1, \quad (1)$$

where n denotes the time sample index. The packet loss flag is defined as F . If packet loss flag is 1, the frame is lost, and all samples in this frame are set to zero. The received signal can be defined as:

$$\begin{cases} x_k = s_k + v_k, & F_k = 0 \\ x_k = [0, 0, \dots, 0], & F_k = 1 \end{cases} \quad (2)$$

In the frequency domain, Eq. 2 can be formulated as:

$$\begin{cases} X(t, f) = S(t, f) + V(t, f), & F_t = 0 \\ X(t, f) = 0, & F_t = 1 \end{cases} \quad (3)$$

where t and f are the frame index and the frequency index, respectively. Given the 50% overlap used during short-time Fourier transform (STFT), the correspondence between the transmission frame index k and the STFT frame t is: $F_k = 1, F_{t=2k} = F_{t=2k+1} = 1$.

2.2. Intra-model knowledge distillation

Given that the causal PLC model only utilizes historical information and cannot look ahead to capture future information, which has shown to be effective in [13]. We propose to use dual-path convolution in couples with intra-model knowledge distillation to address this issue. Similar structures have yielded notable results in auto speech recognition (ASR) [14] and voice conversion [11], which also have a significant perception of future information.

As shown in Fig. 1, the dual-path convolution is based on depth-wise separable convolution [10] and contains both causal and non-causal paths. In parallel depth convolution layers, the padding mode of the depth convolution is modified to implement causal and non-causal paths. Meanwhile, the parameters of the parallel depth convolution layers are not shared. Specifically, in gated convolutional module [15] and the time-

frequency dilated convolutional module (TFDCM) [16], all the convolutions in time dimension have been replaced with dual-path convolutions.

Unlike the typical knowledge distillation in teacher-student networks [17, 18, 19], We only design dual paths in the convolution layer of the encoder, while other layers share parameters. Consequently, both causal and non-causal inputs can be simultaneously received by the decoder and the bottleneck layer during training, which reduces the difficulty of knowledge transfer. This approach also allows us to obtain the final model in one single step, avoiding the need first to train the teacher model and then distill the student model. By using the intra-model knowledge distillation strategy, the training cost is reduced significantly.

During the training process, the input data stream is directed through two distinct paths. To maximize the similarity between the causal encoder output and the non-causal encoder output, we calculate the distillation loss between the causal encoder's output and the non-causal encoder's output. Dual-path convolution not only allows for the concurrent training of both causal and non-causal models in one step, but it also employs the non-causal model to guide the causal model during training, thus improving the performance of the causal model by using the future information.

2.3. Network optimization and two-stage post-processing

Similar to BS-PLCNet [9], we retain the Generative Adversarial Network (GAN) architecture. For the generator, we have the following improvements. First, we replace all convolution layers with depth-wise separable convolution layers to reduce the computational complexity of BS-PLCNet. Secondly, the layer number within the TFDCM module is modified. The proposed layer number is dynamically adaptive in accordance with the network depth rather than using fixed 4 layers in [9]. The dynamic way to set the layer number reduces the receptive field and allows a stronger focus on local information, which is achieved as neural network layers deepen and the feature dimensions increase. Finally, we replace the bottleneck layer from three F-T-LSTM [20] layers to a single F-T-GRU

layer [21]. Stacking multiple dual-path layers for the same sequence can lead to limited performance [22].

Motivated by the multi-stage network for speech enhancement (SE), we also introduce a post-processing module to recover the potential distortions. Potential noise is also addressed at this stage. Considering the overall computational complexity increase, the post-processing module only processes the 0-8kHz frequency bands with a simple GRU-based structure, which consists of a Linear layer activated by Tanh, a GRU layer, a 2D convolution (Conv2d) layer, and a batchnorm layer followed by another Linear layer activated by Tanh.

During training, we applied a two-stage training strategy. In the first stage, we utilize multi-frequency discriminators (MFD) [23] and multi-period discriminators (MPD) [24] to train the first stage of the BS-PLCNet 2. During the first training stage, the model focuses on recovering lost audio samples by leveraging the received audio signal and packet loss information. In the second stage, we use MetricGAN [25] to train the post-processing module. The parameters of the first stage module are frozen during the second stage of training. The second stage input is a mixture of the first stage output and additive noise.

2.4. Loss function

For the first training stage, we utilize speech loss $\mathcal{L}_{\text{speech}}$ and GAN loss. The speech loss is composed of a power-law compressed phase aware (PLCPA) loss $\mathcal{L}_{\text{PLCPA}}$ [15] and time-domain MAE loss \mathcal{L}_{MAE} .

$$\mathcal{L}_{\text{PLCPA}} = \mathcal{L}_{\text{mag}} + \mathcal{L}_{\text{pha}}, \quad (4)$$

$$\mathcal{L}_{\text{MAE}} = \sum_{i=0}^N |s_i - \hat{s}_i|, \quad (5)$$

$$\mathcal{L}_{\text{speech}} = \mathcal{L}_{\text{PLCPA}} + \mathcal{L}_{\text{MAE}}. \quad (6)$$

For the GAN loss, we use the least-square GAN (LSGAN) [26] style training objectives, which can be defined as:

$$\mathcal{L}_D = (1 - D(s))^2 + D(G(x))^2, \quad (7)$$

$$\mathcal{L}_G = (1 - D(G(x)))^2. \quad (8)$$

The MAE based f0 prediction loss \mathcal{L}_{f0} is still retained as:

$$\mathcal{L}_{f0} = \sum_{i=0}^T |f_i - \hat{f}_i| \quad (9)$$

where T represents the frame index. The overall multi-loss for the first training stage is:

$$\mathcal{L}_{\text{first}} = \mathcal{L}_{\text{PLCPA}} + \mathcal{L}_{\text{MAE}} + \alpha \cdot \mathcal{L}_{f0} + \mathcal{L}_G, \quad (10)$$

where α equals $1e^{-1}$, and for the second training stage, we only use speech loss to train the post-processing module.

3. Experiments

3.1. Datasets

For the training data, we use train-clean-100 and train-clean-360 from LibriSpeech [27] and the 5th DNS Challenge dataset¹ as the original clean speech, from which the training lossy sig-

¹<https://github.com/microsoft/DNS-Challenge>

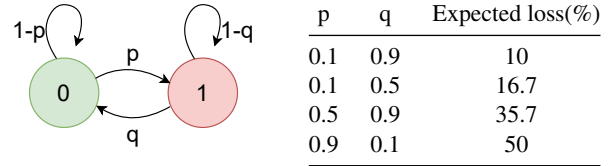


Figure 3: *The Gilbert-Elliott Model.*

nals are simulated. The training set contains nearly 800 hours data in total.

3.2. Pack loss rate configuration

We use the Gilbert-Elliott model [28] to model the conditional consecutive packet loss. As shown in Fig. 3, the Gilbert-Elliott model uses a two-state first-order Markov chain to describe the packet loss status of the current frame and the next frame, where 0 indicates no packet loss, and 1 denotes the existence of packet loss. If the current frame transmits normally, The next frame has a p probability of being lost. If the current frame is lost, the probability of the next frame not being lost is q . By setting p and q , the expected packet loss rate r can be controlled as:

$$r = p/(p + q). \quad (11)$$

The packet loss rate is set to not exceed 50%. During training, we first randomly select two transition probabilities p and q , ranging from 10% to 90%. Then we check whether the expected packet loss rate is below 50% according to Eq. 11. If the packet loss rate condition is not satisfied, then p and q are reset. Note that r is the expected packet loss rate, actual packet loss rate for individual sentence may vary.

For the first training stage, we only simulate packet loss. Noise is introduced in the second training stage. We use the 5th DNS Challenge noise dataset to simulate the noisy mixtures. The training data are generated and segmented into 8s chunks in each batch with signal-to-noise ratio (SNR) ranging from 10 to 20dB.

3.3. Experimental setup

The proposed BS-PLCNet 2 works in T-F domain, we apply the Hanning window for each frame, the frame shift and FFT points are set to 20ms, 10ms and 960, respectively. We train the model with the Adam optimizer [31]. The initial learning rate is set to $1e^{-3}$ and $1e^{-4}$ in the first and second training stages, respectively. The learning rate is halved as the validation loss does not decrease in two epochs. We apply a maximum L2 norm of 5 for gradient clipping. Each stage is trained for 30 epochs. For the wide-band module, both the encoder and decoder have 4 layers, each with 80 channels, except for the first layer of the encoder and the last layer of the decoder. The kernel size and the strides are set to (2, 3) and (1, 2) in time and frequency axes, respectively.

In the TFDCM module, the stacked n convolutional layers are adopted with a dilation rate of $[2^0, 2^1, \dots, 2^{n-1}]$ in both the time and frequency axis. From the first layer to the fourth layer of the encoder, n is respectively set as [4,4,3,2]. The stride and kernel size of DConv in TFDCM are set to (1, 1) and (3, 3). The TFCM module is the same as that in [32]. The F-T-GRU layer is a single GRU layer with 128 hidden states. For the high-band module, the first Conv2D has an output channel of 128, and the GRU layer has a hidden state of 128.

Table 1: The objective scores on the 2024 blind test set. Here, the “Structure optimization”, the “Dual conv” and the “Two-stage” strategies are for optimizing the complexity, utilization of future information and post-processing, respectively.

System	Model	Para.(M)	FLOPs(G)	Structure optimization	Dual conv	Two stage	Causal	PLCMOS	SIG	BAK	OVRL
S0	lossy	-	-					2.69	3.17	3.92	2.89
S1	Team 1024K [29]	-	-				✓	4.16	-	-	-
S2	BS-PLCNet [9]	6.50	8.95	×	×	×	✓	4.02	3.62	4.09	3.35
S3	BS-PLCNet 2	2.01	3.30	✓	×	×	✓	4.05	3.63	4.09	3.35
S4	BS-PLCNet 2	2.01	3.30	✓	✓	×	✓	4.14	3.64	4.10	3.37
S5	BS-PLCNet 2	2.62	3.41	✓	✓	✓	✓	4.20	3.61	4.15	3.37
S6	BS-PLCNet 2-NC	2.62	3.41	✓	✓	✓	×	4.30	3.66	4.14	3.40

Table 2: The objective scores on the 2022 blind test set.

Model	PLCMOS	OVRL	PESQ	STOI
lossy	2.90	2.56	2.37	0.84
Team Kuaishou [6]	4.27	-	3.01	-
LPCNet [8]	3.74	3.09	2.93	0.90
PLCNet [30]	3.83	-	2.37	0.87
BS-PLCNet [9]	4.29	3.20	3.04	0.90
BS-PLCNet 2	4.36	3.25	3.22	0.92

3.4. Metrics

This paper evaluates different systems on the blind test sets of ICASSP 2024 and Interspeech 2022 PLC Challenges. Note that labels for the former are unavailable, while labels for the latter are publicly available. On the ICASSP 2024 PLC Challenge blind test, PLCMOS [12] and DNSMOS [33] are used, which are both neural networks trained to estimate the ratings human raters would assign to an audio file. Specifically, DNSMOS consists of speech quality (SIG), background noise quality (BAK), and the overall quality (OVRL) of the audio. On Interspeech 2022 PLC Challenge blind test, PLCMOS, DNNMOS, PESQ [34] and STOI [35] are used to evaluate different systems. It is worth noting that the released PLCMOS version from Microsoft is different for these two challenges.

3.5. Results and analysis

In Table 1, we perform ablation experiments for the proposed method on the blind test set of ICASSP 2024 PLC Challenge and compare it with two other systems, Team 1024K [29] and BS-PLCNet [9], which tied for the 1st place in the challenge.

To verify the effectiveness of optimizing the network structure in Section 2.3, we perform the system comparison between BS-PLCNet (S2) and BS-PLCNet 2 (S3) as shown in Table 1. The new network structure achieves comparable PLCMOS and DNSMOS with a substantial reduction in both computational complexity and the number of parameters.

To show that utilization of future information through knowledge distillation in a casual way is beneficial to the PLC system, BS-PLCNet 2 with dual-conv (S4) is compared with S3. As can be seen from Table 1, the use of intra-model knowledge distillation significantly improves PLCMOS from 4.05 to 4.14. This observation demonstrates that under the guidance of the non-causal model, the causal model achieves superior performance without using the feature information. It is worth noting

that since the dual-path convolution only executes one branch during inference, the computational complexity and parameters remain the same.

To investigate the importance of our proposed post-processing module and two-stage training strategy, BS-PLCNet 2 combining dual-conv and two-stage post-processing strategy (S5) is compared with system S4. The post-processing module we added has improved both PLCMOS and DNSMOS, especially the BAK scores, while only leading to an increase of 0.11 GFLOPs in computational complexity. The use of the post-processing module and two-stage training strategy enables the model to handle potential distortions and noise, thus improving the quality of the generated speech.

To show the system upper bound by using the future information in a non-causal way, BS-PLCNet 2-NC (S6) shows the result of the non-causal branch in the dual-path convolution. Due to its ability to access future information, the non-causal model still has an advantage in PLC.

In Table 2, we also use the Interspeech 2022 blind test set to compare our system with the top three systems from the challenge, which are Team Kuaishou [6], LPCNet [8], and PLCNet [30]. The results show that the proposed BS-PLCNet 2 continues to retain its superiority when compared with more systems. The two-stage band-split PLCNet with intra-model knowledge distillation outperforms other methods in all metrics. Especially in PESQ, compared to the top 1 team Kuaishou, the proposed method brings 0.21 PESQ performance improvement. The results on both the ICASSP 2024 blind test set and the Interspeech 2022 blind test set show that the proposed BS-PLCNet 2 achieves SOTA. Our samples are available ².

4. Conclusion

In this paper, we propose a novel two-stage band-split packet loss concealment network – BS-PLCNet 2, which is an upgrade of BS-PLCNet. Benefiting from the intra-model knowledge distillation, BS-PLCNet 2 achieves the state-of-the-art PLCMOS performance in the ICASSP 2024 PLC Challenge blind test set with only 38.1% FLOPs of BS-PLCNet in computation complexity. Additionally, both causal and non-causal models can be obtained simultaneously within a single training process. In the future, we will particularly consider the efficient use of temporal information, also a combination in the time-frequency domain to further improve speech quality.

²<https://zzhdzd.github.io/BS-PLCNet2>

5. References

- [1] V.-A. Nguyen, A. H. Nguyen, and A. W. Khong, "Improving performance of real-time full-band blind packet-loss concealment with predictive network," in *ICASSP*. IEEE, 2023, pp. 1–5.
- [2] J. Lin, Y. Wang, K. Kalgaonkar, G. Keren, D. Zhang, and C. Fuegen, "A time-domain convolutional recurrent network for packet loss concealment," in *ICASSP*. IEEE, 2021, pp. 7148–7152.
- [3] C. Perkins, O. Hodson, and V. Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE network*, vol. 12, no. 5, pp. 40–48, 1998.
- [4] B. J. Borgström, P. H. Borgström, and A. Alwan, "Efficient hmm-based estimation of missing features, with applications to packet loss concealment," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
- [5] B.-K. Lee and J.-H. Chang, "Packet loss concealment based on deep neural networks for digital speech transmission," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 378–387, 2015.
- [6] N. Li, X. Zheng, C. Zhang, L. Guo, and B. Yu, "End-to-end multi-loss training for low delay packet loss concealment," in *Interspeech*, 2022, pp. 585–589.
- [7] N. L. Westhausen and B. T. Meyer, "tplcnet: Real-time deep packet loss concealment in the time domain using a short temporal context," in *Interspeech*, 2022.
- [8] J.-M. Valin, A. Mustafa, C. Montgomery, T. B. Terriberry, M. Klingbeil, P. Smaragdis, and A. Krishnaswamy, "Real-time packet loss concealment with mixed generative and predictive model," in *Interspeech*, 2022.
- [9] Z. Zhang, J. Sun, X. Xia, C. Huang, Y. Xiao, and L. Xie, "Bsplcnet: Band-split packet loss concealment network with multi-task learning framework and multi-discriminators," in *ICASSP*. IEEE, 2024.
- [10] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.
- [11] Z. Ning, Y. Jiang, P. Zhu, J. Yao, S. Wang, L. Xie, and M. Bi, "Dualvc: Dual-mode voice conversion using intra-model knowledge distillation and hybrid predictive coding," in *ICASSP*. IEEE, 2023.
- [12] L. Diener, M. Purin, S. Sootla, A. Saabas, R. Aichner, and R. Cutler, "Plcmos – a data-driven non-intrusive metric for the evaluation of packet loss concealment algorithms," in *Interspeech*, 2023.
- [13] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "Interspeech 2022 audio deep packet loss concealment challenge," in *Interspeech*, 2022.
- [14] C. Liang, X.-L. Zhang, B. Zhang, D. Wu, S. Li, X. Song, Z. Peng, and F. Pan, "Fast-u2++: Fast and accurate end-to-end speech recognition in joint ctc/attention frames," in *ICASSP*. IEEE, 2023, pp. 1–5.
- [15] S. Zhang, Z. Wang, J. Sun, Y. Fu, B. Tian, Q. Fu, and L. Xie, "Multi-task deep residual echo suppression with echo-aware loss," in *ICASSP*. IEEE, 2022, pp. 9127–9131.
- [16] J. Sun, D. Luo, Z. Li, J. Li, Y. Ju, and Y. Li, "Multi-task sub-band network for deep residual echo suppression," in *ICASSP*. IEEE, 2023, pp. 1–2.
- [17] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size dnn with output-distribution-based criteria," in *Fifteenth annual conference of the international speech communication association*, 2014.
- [18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [19] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition," in *Interspeech*, 2016, pp. 3439–3443.
- [20] Z. Zhang, S. Zhang, M. Liu, Y. Leng, Z. Han, L. Chen, and L. Xie, "Two-step band-split neural network approach for full-band residual echo suppression," in *ICASSP*. IEEE, 2023, pp. 1–2.
- [21] Z. Chen, X. Xia, S. Sun, Z. Wang, C. Chen, G. Xie, P. Zhang, and Y. Xiao, "A progressive neural network for acoustic echo cancellation," in *ICASSP*. IEEE, 2023, pp. 1–2.
- [22] W. Tong, J. Zhu, J. Chen, S. Kang, T. Jiang, Y. Li, Z. Wu, and H. Meng, "Scnet: Sparse compression network for music source separation," in *ICASSP*. IEEE, 2024.
- [23] Q. Tian, Y. Chen, Z. Zhang, H. Lu, L. Chen, L. Xie, and S. Liu, "Tfgan: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis," *arXiv preprint arXiv:2011.12206*, 2020.
- [24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.
- [25] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," in *Interspeech*, 2021.
- [26] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.
- [27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.
- [28] M. Mushkin and I. Bar-David, "Capacity and coding for the gilbert-elliott channels," *IEEE Transactions on Information Theory*, 1989.
- [29] N. Li, G. Yu, C. Zhang, C. Zhou, Q. Huang, and B. Yu, "Multi-stage training for cross-domain full-band audio packet loss concealment," in *ICASSP*, 2024.
- [30] B. Liu, Q. Song, M. Yang, W. Yuan, and T. Wang, "Plcnet: Real-time packet loss concealment with semi-supervised generative adversarial network," in *Interspeech*, 2022, pp. 575–579.
- [31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.
- [32] Y. Ju, S. Zhang, W. Rao, Y. Wang, T. Yu, L. Xie, and S. Shang, "Tea-pse 2.0: Sub-band network for real-time personalized speech enhancement," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 472–479.
- [33] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*. IEEE, 2022, pp. 886–890.
- [34] 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 *ICASSP*, vol. 2. IEEE, 2001, pp. 749–752.
- [35] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.