



End-to-End Multi-Loss Training for Low Delay Packet Loss Concealment

Nan Li, Xiguang Zheng, Chen Zhang, Liang Guo, Bing Yu

Kuaishou Technology Co. Beijing, China

linan06@kuaishou.com

Abstract

Real-time teleconferencing has become one of the essential parts in our daily life. While packet loss during real-time data transmission is unavoidable, traditional signal processing based Packet Loss Concealment (PLC) techniques have been developed in recent decades. In recent years, deep learning based approaches have also proposed and achieved state-of-the-art PLC performance. This work presents a low-delay multi-loss based neural PLC system. The multi-loss is consisted by a signal loss, a perceptual loss and an ASR loss ensuring good speech quality and automatic speech recognition compatibility. The proposed system was ranked 1st place in INTERSPEECH 2022's Audio Deep Packet Loss Concealment Challenge.

Index Terms: packet loss concealment, deep neural network, wave-U-Net

1. Introduction

Over the past few decades, real-time communication over internet has become one of the essential part of our daily life. While packet loss is unavoidable for practical applications, traditional interpolation [1] and Hidden Markov Model (HMM) [2] based approaches have been employed in contemporary speech codecs such as opus [3] and amr-wb [4]. These traditional methods can in general work well for moderate packet loss conditions, but the speech quality degrades drastically for high and burst packet loss scenarios [5].

In recent years, methods based on deep neural network (DNN) has outperformed the tradition methods and become the mainstream for various speech applications such as Speech Enhancement (SE) [6], Acoustic Echo Cancellation (AEC) [7]. DNN based Packet Loss Concealment (PLC) approach is less studied but promising [8]. The delay-insensitive audio concealment techniques [9, 10] (also referred as audio inpainting) trade latency with quality. It requires seconds of history and future audio frames to recover the current missing audio frame. This work focuses on the low-delay PLC task that can only look tens of milliseconds future frames.

Most of the early low-delay deep PLC approaches are operating in the Time-Frequency (T-F) domain. [5] employs stacked fully connected layers with 20ms windowed time-frequency log-spectra and phase features and trains the DNN with the MSE losses. While improved Automatic Speech Recognition (ASR) accuracy and speech quality compared to the traditional HMM based methods is demonstrated, it is harder for the time-frequency domain based deep PLC system to accurately predict the phase information comparing to the NS and AEC tasks where the phase of the noisy and echoic input signal is available for the NS and AEC tasks to estimate the phase of the target speech.

More recently, time domain based end-to-end methods are proposed to overcome such limitation [11]. In [12], a Long Short-Term Memory (LSTM) based Recurrent Neural Network (RNN) is proposed with a 10ms frame length. System proposed

in [11] extends the model structure by including the Convolutional neural network (CNN) layers on top of the RNN layers and achieved better performance. The effect of lookahead mechanism up to two 20ms frames is also investigated demonstrating the trade-off between system latency and speech quality. In one of the most recent works [13], a Generative Adversarial Network (GAN) is proposed that incorporates the dilated residual convolution into the encoder-decoder architecture alongside with a multi-resolution time-frequency domain and a time-domain discriminator.

In this work, an end-to-end system is proposed for low-delay PLC task. Compared to [13], the proposed system is also based on the time domain GAN structure except the SEANet[14, 15] is employed for the encoder-decoder structure in the generator with multi-loss to simultaneously ensure good speech quality and ASR compatibility. While employing the multi-loss consisted by a signal loss, a wav2vec [16] based perceptual loss and an ASR loss during the model training stage can provide a good balance between the speech quality and the Word Error Rate (WER) for the SE [17] and AEC [18] tasks, here, an extended version of the multi-loss is proposed as the generator loss. Compared to the single resolution signal losses in [17, 18], the signal loss in this work is formed by a multi-resolution STFT loss [19] and a new Multi-Resolution Optimal Scale-Invariant Signal-to-Noise Ratio (MR-O-SISNR) loss extended from [20]. The ASR loss is formed by an ASR feature embedding loss obtained using the encoder portion of the WeNet system [21]. Comparing to the existing time-domain methods [12, 11], the proposed system exploits the flexibility of the time-domain processing by predicting a 1ms non-overlapping frame (16 samples for 16kHz sample rate) each time using 18ms of lookahead samples and 1ms stride resulting in total of 20ms system delay, which simultaneously ensures minimum system delay whilst maintaining good speech quality.

The proposed system has also participated the INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge [8]. With in total of 20ms system latency, 2.36M model parameters, 0.27 real-time factor on Intel Core i5 (2.4GHz) CPU, the proposed method is ranked 1st place in this challenge.

2. Signal Model

For an input speech stream $s(t)$, it can be divided into chunks of frames where each frame contains L time-domain samples. Thus the n^{th} frame \mathbf{S}_n can be represented by:

$$\mathbf{S}_n = [s(t_n), \dots, s(t_n + L - 1)] \quad (1)$$

If frame n is not received (i.e. $\mathbf{S}_n = \mathbf{0}$), the recovered signal $\hat{\mathbf{S}}_n$ can be estimated from the adjacent frames by:

$$\hat{\mathbf{S}}_n = \Theta(\mathbf{S}_I) \quad (2)$$

where $\Theta(\cdot)$ is the PLC algorithm, $I = [n - l, \dots, n + m]$ represents the l history and m lookahead frames relative to the cur-

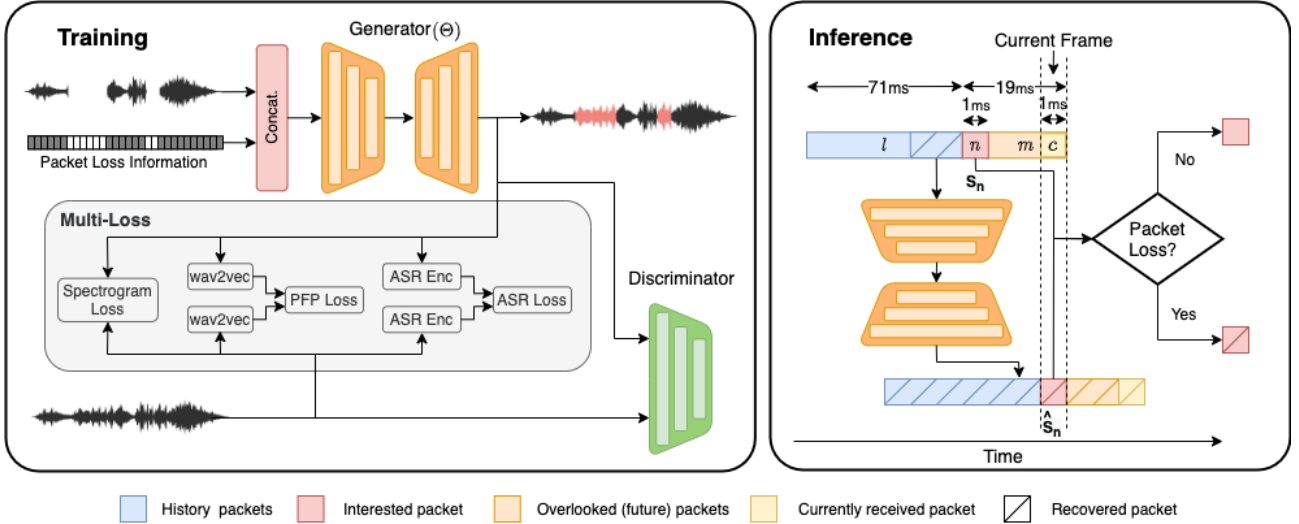


Figure 1: Overview of the proposed system

rent frame n . For each input frame \mathbf{S}_i in \mathbf{S}_I at current frame n , it is either the original received frames, the previously recovered frames or the lost frames:

$$\mathbf{S}_i = \begin{cases} \mathbf{S}_i, & \text{frame } i \text{ is received} \\ \hat{\mathbf{S}}_i, & \text{frame } i \text{ is lost, } i < n \\ \mathbf{0}, & \text{frame } i \text{ is lost, } i \geq n \end{cases} \quad (3)$$

3. Proposed system

3.1. Overview

Figure 1 presents the architecture of the proposed system. During the training stage, the time-domain audio signal with packet loss is fed to the encoder-decoder based generator to recover the missing frames. Multi-loss mechanism is employed as the generator loss to simultaneously ensuring good speech quality and automatic speech recognition compatibility. A discriminator is jointly trained with the aim to distinguish the recovered speech signal from the ground truth.

The latency of the proposed system is illustrated in the right side of Figure 1. As shown, each frame contains 1ms of non-overlapping time-domain audio samples (16 points for 16kHz sample rate). The yellow frame c (first frame from the right side) indicates the currently received frame. If packet loss occurs for the red frame n (18ms earlier than the current frame c), the proposed deep PLC system takes $m = 18$ frames of the orange look-ahead (future) frames alongside with $l = 71$ frames of blue history samples relative to the red frame n as the input to $\Theta(\cdot)$ to recover the lost red frame. Otherwise, the output frame at time instant c is the received red frame n . Combined with 1ms stride frame, the proposed system thus introduces in total of 18ms (lookahead frames) + 1ms (current frame) + 1ms (stride frame) = 20ms latency, which satisfies the total system delay (20ms) of the INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge [8].

3.2. Generator

The generator is based on the encoder-decoder structure in [15]. The encoder consists a 1D convolution layer followed by 4 encoder blocks. As listed in Table 1, each of the encoder blocks consist of 3 residual units (RU) with kernel of 7 and dilation rate of 1, 3 and 9, respectively. The last residual unit is followed by

Table 1: Encoder configurations

Block	Layer	In	Out	Kernel	Stride	Dilation
	Conv 1	2	16	7	1	1
Encoder 1	RU 1	16	16	7	1	1
	RU 2	16	16	7	1	3
	RU 3	16	16	7	1	9
	DS	16	32	4	2	1
Encoder 2	RU 1	32	32	7	1	1
	RU 2	32	32	7	1	3
	RU 3	32	32	7	1	9
	DS	32	64	4	2	1
Encoder 3	RU 1	64	64	7	1	1
	RU 2	64	64	7	1	3
	RU 3	64	64	7	1	9
	DS	64	128	4	2	1
Encoder 4	RU 1	128	128	7	1	1
	RU 2	128	128	7	1	3
	RU 3	128	128	7	1	9
	DS	128	256	4	2	1
	Conv 2	256	320	3	1	1

a convolution with stride as a down-sampler (DS). The channel number of the first encoder block down-sampler is 16 and each following down-samplers double the channel number when applying down-sample. A final 1D convolution layer is used to smooth the down-sampling feature of the encoder blocks. The architecture of the decoder is approximately the same as the encoder. A 1D convolution layer is followed by 4 decoder blocks. Each of the decoder blocks consist 1 transposed convolution as up-sampler. The strides and channel numbers of the transposed convolutions mirrors the configurations of down-samplers in the 4 encoder blocks. The up-sampler is followed by the same 3 residual units in encoder blocks. A final 1D convolution layer is used to smooth the output of the whole model. An ELU activation function is apply after each 1D convolution.

3.3. Discriminator

A time domain and a time-frequency domain convolutional discriminator from [15] are employed. The multi-resolution con-

volitional discriminator is employed where three structurally identical models are used to the input audio at original, 2-times down-sampled and 4-times down-sampled resolutions. The time-frequency domain discriminator operates on a signal scale with a 1024 samples window size and a hop length of 256 samples. The input features are fed into a 2D-convolution layer and followed by several sequentially arranged residual blocks. The training target of the discriminator is to classify original vs. recovered audio frames. The implementation details of the discriminator can be found in [15].

3.4. Multi-Loss

As shown in the gray box inside the generator of Figure 1, the proposed multi-loss contains three loss functions to measure the estimation error from the signal, perceptual, and ASR aspects.

For the signal loss \mathcal{L}_{sig} , a multi-resolution time-frequency domain signal loss $\mathcal{L}_{\text{T-F}}$ [15] with different window sizes are employed:

$$L_{\text{T-F}} = \sum_{K \in \{2^6, \dots, 2^{11}\}} \left(\sum_n \sum_k \left\| S(n, k) - \hat{S}(n, k) \right\|_1 + \alpha_{\mathcal{L}_{\text{T-F}}} \sum_n \sum_k \left\| \ln S(n, k) - \ln \hat{S}(n, k) \right\|_2 \right) \quad (4)$$

where $S(n, k)$ denotes the clean speech magnitude for the k^{th} frequency bin ($1 \leq k \leq K$) at the n^{th} frame obtained from a K point fft with $K/4$ hop length. $\hat{S}(n, k)$ is the estimated corresponding speech signal. $\alpha_{\mathcal{L}_{\text{T-F}}}$ is set to $\alpha_{\mathcal{L}_{\text{T-F}}} = \sqrt{K/2}$ as in [19]. In addition, a time-frequency domain Multi-Resolution Optimal Scale-Invariant Signal-to-Noise Ratio (MR-O-SISNR) loss is proposed to extend the original time-domain O-SISNR in [20]:

$$S_{\text{target}}(n, k) = \frac{|\hat{S}(n, k)|^2 S(n, k)}{\langle S(n, k), \hat{S}(n, k) \rangle} \quad (5)$$

$$E_{\text{noise}}(n, k) = \hat{S}(n, k) - S_{\text{target}}(n, k) \quad (6)$$

$$\mathcal{L}_{\text{MR-O-SISNR}} = \sum_K \sum_n \sum_k 10 \log_{10} \frac{\|S_{\text{target}}(n, k)\|^2}{\|E_{\text{noise}}(n, k)\|^2} \quad (7)$$

The final signal loss is given by:

$$\mathcal{L}_{\text{SIG}} = \mathcal{L}_{\text{T-F}} + \mathcal{L}_{\text{MR-O-SISNR}} \quad (8)$$

The Phone-Fortified Perceptual Loss (PFPL) proposed in [22] is applied to take phonetic information into account for training the proposed deep PLC network. The $\mathcal{L}_{\text{PFPL}}$ is calculated by Wasserstein distance between the latent representations of wav2vec [16] model for original and recovered speech.

Moreover, an ASR-oriented loss \mathcal{L}_{ASR} is used to reduce the speech distortion of enhanced speech and reduce the WER for ASR. More specifically, Wasserstein distance is calculated between the embeddings of the ASR encoder for clean and enhanced speech. A pre-trained ASR encoder with LibriSpeech dataset [23] by WeNet toolkit [21] is used to extract the embeddings. The overall multi-loss \mathcal{L}_G for the generator is given by:

$$\mathcal{L}_G = \alpha_{\mathcal{L}_{\text{SIG}}} \mathcal{L}_{\text{SIG}} + \mathcal{L}_{\text{PFPL}} + \mathcal{L}_{\text{ASR}} \quad (9)$$

where the value of the signal weight $\alpha_{\mathcal{L}_{\text{SIG}}}$ is set to 0.1 in order to ensure roughly equal contribution among the signal, perceptual and the ASR losses in \mathcal{L}_G .

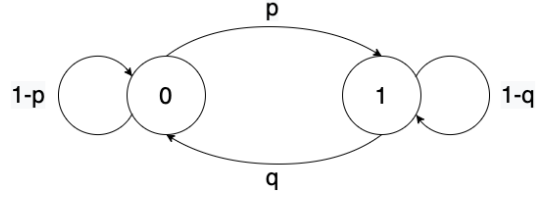


Figure 2: The Gilbert-Elliott Model

4. Experiments

4.1. Packet loss simulator

In addition to the real loss pattern provide by the Challenge, we also used the packet loss traces simulated by the Gilbert-Elliott Channel Model [24] during training as used in [11] to further increase the diversity of the packet loss patterns. As shown in Figure 2, it is a two state first order Markov chain, which could be used to model the conditional consecutive packet loss. The state 0 indicates successfully receiving a packet while state 1 represents packet loss. The probabilities for the transmission from state 0 to 1 and from state 1 to 0 are p and q , respectively. The packet loss rates r are set randomly from 20% to 50% corresponding to p and q randomly changing from 0.2 to 0.8 while satisfying $r = p/(p + q)$.

4.2. Datasets and experiments setup

60 hours of speech signals from INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge are used with additional 60 hours speech signals randomly chosen from AISHELL-1 [25], AISHELL-3 [26], VCTK [27] and LibriSpeech [28] to improve the generalization ability of the proposed system. 10% of the speech signals are mixed with noise from MUSAN [29] at a random SNR from 6dB to 18dB. During the training stage, the packet loss traces are randomly selected for each epoch. For each of the selected packet loss traces, the starting point of a given trace is also randomly selected to allow sufficient diversity of the packet loss patterns.

All training data are sampled to 16kHz. We use the Adam optimizer to train the model and the learning rate for training the generator and discriminator are 0.0001 and 0.00005 respectively. The generator is firstly trained for 60 epochs before introducing discriminator. It can be found in the experiment results that including the discriminator can further improve the PESQ [30], PLC-MOS [8] and WER scores.

4.3. Conditions

The detailed experiment conditions are summarized as follows:

- G1+ \mathcal{L}_{SIG} : the generator network in Section 3.2 with 1ms frame length, 1ms stride, 18ms lookahead (20ms system latency in total), trained using the signal loss \mathcal{L}_{SIG} consisted by $\mathcal{L}_{\text{T-F}}$ and $\mathcal{L}_{\text{MR-O-SISNR}}$ in (8);
- G1+ $\mathcal{L}_{\text{T-F}}$ + $\mathcal{L}_{\text{SR-O-SISNR}}$: condition G1+ \mathcal{L}_{SIG} with the MR-O-SISNR replace by a Single Resolution O-SISNR (SR-O-SISNR) with 512-point FFT and 128-point stride;
- G1+ \mathcal{L}_G : the generator network in Section 3.2 with the full generator loss \mathcal{L}_G in (9);
- G1+ \mathcal{L}_G +D: condition G1+ \mathcal{L}_G with the discriminator in Section 3.3;

Table 2: The objective scores and WER of different models on the 966 testing clips

	Packet Loss Rate %	G1+L _{TF} +L _{SR-O-SISNR}	G1+L _{SIG}	G1+L _G	G1+L _G +D (proposed)	G5+L _G +D	G10+L _G +D	T-F GAN [31]	Zero filling baseline
PLC-MOS	0 to 10	4.44	4.46	4.46	4.49	4.45	4.40	4.28	3.79
	10 to 20	4.15	4.23	4.26	4.36	4.11	3.94	3.63	2.51
	20 to 30	3.81	3.96	3.98	4.16	3.68	3.42	3.09	1.93
	30 to 50	3.45	3.61	3.70	3.99	3.22	2.85	2.44	1.58
	50 to 100	2.52	2.72	2.88	3.39	2.43	2.05	1.81	1.78
	Overall	4.02	4.10	4.13	4.27	3.97	3.80	3.57	2.88
PESQ	0 to 10	3.93	3.95	3.95	3.96	3.81	3.72	3.44	3.11
	10 to 20	2.74	2.76	2.76	2.80	2.54	2.35	2.17	1.63
	20 to 30	2.14	2.17	2.17	2.20	1.98	1.82	1.69	1.31
	30 to 50	1.68	1.68	1.69	1.70	1.54	1.44	1.36	1.14
	50 to 100	1.25	1.25	1.26	1.27	1.20	1.18	1.14	1.06
	Overall	2.97	2.99	2.99	3.01	2.84	2.72	2.56	2.19
WER (%)		12.46	12.46	11.05	10.97	11.88	12.91	12.15	11.24

Table 3: INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge evaluation on blind test and rank

Rank	Team	PLC-MOS	DNS-MOS	CMOS	Word accuracy	Final score
1	Kuaishou (proposed)	4.282	3.797	-0.552	0.875	0.845
	Zero filling baseline	2.904	3.444	-1.231	0.861	0.725

- G5+L_G+D: condition G1+L_G+D with 5ms frame length, 5ms stride, 10ms lookahead (20ms system latency in total);
- G10+L_G+D: condition G1+L_G+D with 10ms frame length, 10ms stride, 5ms lookahead (25ms system latency in total);
- T-F GAN: the temporal-spectral Gan system proposed in [31] and open sourced here¹. The total algorithmic latency is 42ms.
- Zero filling baseline: the baseline system with the lost frames filled by zeros.

4.4. Results

The conditions in Table 2 is evaluated using PESQ [30], PLC-MOS [8] and WER metrics on 966 testing clips provided by INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge. The WER is calculated based on a pretrained ASR model² released by WeNet [21]. The results are listed in Table 2. The scores are further categorized by different packet loss rates with 457, 202, 103, 116 and 88 samples for the packet loss rate classes from 0% to 100% in Table 2.

For the G1 systems, the benefit of employing the proposed multi-loss training is demonstrated progressively. First, condition G1+L_{SIG} achieved higher PLC-MOS and PESQ scores compared to condition G1+L_{TF}+L_{SR-O-SISNR} indicating the effectiveness of the proposed Multi-Resolution-O-SISNR signal loss. The ASR accuracy is significantly improved while maintaining the speech quality compared to condition G1+L_{SIG} when applying the perpetual loss L_{PFPL} and the ASR loss L_{ASR} in additional to the signal loss L_{SIG} as in G1+L_G. Finally, the speech quality and the ASR accuracy is further improved when the discriminator is introduced after 60 epochs of the generator training indicating the contribution of the adversarial training strategy as in condition G1+L_G+D.

¹<https://github.com/guanyuansheng/TFGAN-PLC>

²<https://github.com/wenet-e2e/wenet/blob/main/examples/gigaspeech>

Conditions with different frame lengths are also compared. The G5 and G10 system are designed to ensure similar system delays for fair comparison with the G1 systems. As shown, the objective scores are decreased when increasing the frame length. This is caused by less available lookahead time when the frame length increases under the same system latency requirement. The T-F GAN system proposed in [31] and the Zero filling baseline systems are finally compared with the proposed system. As shown, the proposed system also outperforms these systems demonstrating the benefit of using the proposed G1+L_{SIG}+D system.

4.5. Blind test results from Interspeech deep PLC challenge

The proposed system participated the INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge [8]. The challenge evaluated the proposed system (condition G1+L_{SIG}+D) using objective and subjective metrics. PLC-MOS [8] and DNS-MOS [32] are employed as the objective metrics for speech quality. CMOS based on ITU. P.808 [33] is employed as the subjective metric for speech quality. Word accuracy is used to evaluate the ASR accuracy. The final score is given by averaging the normalized score (between 0 and 1) from the Word accuracy and CMOS scores. As listed in Table 3, with 20ms system latency, 2.36M model parameters, 0.27 real-time factor on Intel Core i5 (2.4GHz) CPU, the proposed method is ranked 1st place in this challenge.

5. Conclusions

In this paper, a low-delay multi-loss based deep PLC system is proposed for real-time applications. With the proposed multi-loss with 1ms input frame length, the proposed system outperforms other experiment conditions on PLC-MOS, PESQ and WER metrics. Besides, the proposed system is ranked the 1st place of INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge.

6. References

- [1] C. Perkins, O. Hodson, and V. Hardman, *A Survey of Packet Loss Recovery Techniques for Streaming Audio*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, p. 607–615.
- [2] B. Borgstrom, P. Borgström, and A. Alwan, “Efficient hmm-based estimation of missing features, with applications to packet loss concealment,” 2010, pp. 2394–2397.
- [3] K. Vos, K. Sørensen, S. Jensen, and J.-M. Valin, “Voice coding with opus,” *135th Audio Engineering Society Convention 2013*, pp. 722–731, 01 2013.
- [4] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, “The adaptive multirate wideband speech codec (amr-wb),” *Speech and Audio Processing, IEEE Transactions on*, vol. 10, pp. 620 – 636, 12 2002.
- [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, Feb. 2016.
- [6] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, “Interspeech 2021 deep noise suppression challenge,” in *INTERSPEECH*, 2021.
- [7] R. Cutler, A. Saabas, T. Parnamaa, M. Loide, S. Sootla, M. Purin, H. Gamper, S. Braun, K. Sorensen, R. Aichner, and S. Srinivasan, “Interspeech 2021 acoustic echo cancellation challenge,” in *INTERSPEECH*, 2021.
- [8] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, “Interspeech 2022 audio deep packet loss concealment challenge,” in *INTERSPEECH 2022 - 23rd Annual Conference of the International Speech Communication Association*, 2022 (submitted).
- [9] A. Marafioti, N. Perraudin, N. Holighaus, and P. Majdak, “A context encoder for audio inpainting,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 12, p. 2362–2372, dec 2019.
- [10] H. Zhou, X. Xu, P. Luo, and X. Wang, “Vision-infused deep audio inpainting,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 283–292.
- [11] 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 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 7148–7152.
- [12] R. Lotfidereshgi and P. Gourmay, “Speech Prediction Using an Adaptive Recurrent Neural Network with Application to Packet Loss Concealment,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5394–5398.
- [13] J. Wang, Y. Guan, C. Zheng, R. Peng, and X. Li, “A temporal-spectral generative adversarial network based end-to-end packet loss concealment for wideband speech transmission,” *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2577–2588, Oct. 2021.
- [14] Y. Li, M. Tagliasacchi, O. Rybakov, V. Ungureanu, and D. Roblek, “Real-time speech frequency bandwidth extension,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 06 2021, pp. 691–695.
- [15] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, “Soundstream: An end-to-end neural audio codec,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.
- [16] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in *Proc. of Interspeech*, 2019.
- [17] L. Chen, C. Xu, X. Zhang, X. Ren, X. Zheng, C. Zhang, L. Guo, and B. Yu, “Multi-stage and multi-loss training for fullband non-personalized and personalized speech enhancement,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9296–9300.
- [18] H. Zhao, N. Li, R. Han, L. Chen, X. Zheng, C. Zhang, L. Guo, and B. Yu, “A deep hierarchical fusion network for fullband acoustic echo cancellation,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9112–9116.
- [19] A. A. Gritsenko, T. Salimans, R. van den Berg, J. Snoek, and N. Kalchbrenner, “A spectral energy distance for parallel speech synthesis,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.
- [20] C. Ma, D. Li, and X. Jia, “Optimal scale-invariant signal-to-noise ratio and curriculum learning for monaural multi-speaker speech separation in noisy environment,” in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2020, pp. 711–715.
- [21] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, “WeNet: Production Oriented Streaming and Non-Streaming End-to-End Speech Recognition Toolkit,” in *Proc. of Interspeech*, 2021, pp. 4054–4058.
- [22] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, “Improving perceptual quality by phone-fortified perceptual loss using wasserstein distance for speech enhancement,” in *Proc. of Interspeech*, 2020.
- [23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *Proc. of ICASSP*. IEEE, 2015, pp. 5206–5210.
- [24] M. Mushkin and I. Bar-David, “Capacity and coding for the gilbert-elliott channels,” *IEEE Trans. Inf. Theory*, vol. 35, pp. 1277–1290, 1989.
- [25] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline,” in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.
- [26] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, “Aishell-3: A multi-speaker mandarin its corpus and the baselines,” *arXiv preprint arXiv:2010.11567*, 2020.
- [27] C. Veaux, J. Yamagishi, and K. Macdonald, “CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” 2017.
- [28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [29] D. Snyder, G. Chen, and D. Povey, “MUSAN: A music, speech, and noise corpus,” *CoRR*, vol. abs/1510.08484, 2015.
- [30] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.
- [31] J. Wang, Y. Guan, C. Zheng, R. Peng, and X. Li, “A temporal-spectral generative adversarial network based end-to-end packet loss concealment for wideband speech transmission,” *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 2577–2588, 2021.
- [32] C. K. A. Reddy, V. Gopal, and R. Cutler, “Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *2020 International Conference on Acoustics, Speech, and Signal Processing*. IEEE, October 2020, pp. 6493–6497.
- [33] B. Naderi and R. Cutler, “An open source implementation of itu-t recommendation p.808 with validation,” *Proc. Interspeech*, 2020.