A Distilled Low-Latency Neural Vocoder with Explicit Amplitude and Phase Prediction

Hui-Peng Du, Yang Ai*, and Zhen-Hua Ling
National Engineering Research Center of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P. R. China
E-mail: redmist@mail.ustc.edu.cn, yangai@ustc.edu.cn, zhling@ustc.edu.cn

***Abstract*—The majority of mainstream neural vocoders primarily focus on speech quality and generation speed, while overlooking latency, which is a critical factor in real-time applications. Excessive latency leads to noticeable delays in user interaction, severely degrading the user experience and rendering such systems impractical for real-time use. Therefore, this paper proposes DLL-APNet, a Distilled Low-Latency neural vocoder which first predicts the Amplitude and Phase spectra explicitly from input mel spectrogram and then reconstructs the speech waveform via inverse short-time Fourier transform (iSTFT). The DLL-APNet vocoder leverages causal convolutions to constrain the utilization of information to current and historical contexts, effectively minimizing latency. To mitigate speech quality degradation caused by causal constraints, a knowledge distillation strategy is proposed, where a pre-trained non-causal teacher vocoder guides intermediate feature generation of the causal student DLL-APNet vocoder. Experimental results demonstrate that the proposed DLL-APNet vocoder produces higher-quality speech than other causal vocoders, while requiring fewer computational resources. Furthermore, the proposed DLL-APNet vocoder achieves speech quality on par with mainstream non-causal neural vocoders, validating its ability to deliver both high perceptual quality and low latency.**

I. Introduction

Neural vocoders convert input acoustic features (e.g., mel spectrogram) into speech waveform by neural networks, thus they directly impact the quality of synthesized speech and have found applications in various domains such as text-to-speech (TTS) [1], [2], speech enhancement (SE) [3], voice conversion (VC) [4], etc.

Early neural vocoders, e.g., WaveNet [5] and SampleRNN [6], employed auto-regressive methods to generate time-domain speech waveforms, which precluded parallel processing. In subsequent work, two categories of neural vocoders emerged: flow-based vocoders [7], [8] and diffusion-based vocoders [9], [10]. While they can produce high-quality speech, their practical deployment, especially on devices with limited computing resources, remains challenging. Recently, generative adversarial networks (GANs) [11] have demonstrated remarkable performance in generative tasks. GAN-based vocoders [1], [2], [12] are among the most prevalent vocoder architectures due to their relatively straightforward algorithmic design, which implicitly incorporates waveform quality improvements through discriminator supervision.

However, beyond the traditional focus on speech quality and generation speed, latency remains a critical yet often overlooked metric in most vocoder research, especially for real-time practical applications. Many existing vocoders rely on standard convolutions, which cause the receptive field to expand as network depth increases when stacking convolutional blocks [13]. This results in considerable model-intrinsic latency, making it infeasible to support frame-by-frame, concurrent transmission and computation—an unacceptable limitation in real-time systems from the receiver's perspective. For example, in communication scenarios involving speech codecs [14], [15] that transmit mel spectrograms, causal vocoders are essential to minimize latency and ensure real-time synthesis at the receiving end.

To address this, we propose DLL-APNet, a Distilled Low-Latency neural vocoder with explicit Amplitude and Phase spectrum prediction. Building on our previous work, APNet2 [16], the proposed vocoder introduces specific improvements to significantly reduce latency. In speech coding fields [17], [18], causal convolutions are commonly employed to reduce model latency and meet communication requirements. Therefore, we replace standard convolutions in APNet2 with causal convolutions, which use asymmetric padding to ensure the convolution kernel does not access future information, thereby minimizing computational latency. To compensate for the speech quality degradation caused by the introduction of causal convolutions, we adopt knowledge distillation training strategy, using a pre-trained APNet2 as the teacher model to guide the learning of the student model DLL-APNet. This allows DLL-APNet to extract maximal knowledge within its limited receptive field, facilitating the generation of high-quality speech waveform. As the results of our experiment, the speech quality synthesized by our proposed DLL-APNet outperforms other causal vocoders and remains comparable to non-causal models, indicating that the knowledge distillation strategy can mitigate the speech quality degradation caused by causal convolutions to some extent.

The rest of the paper is structured as follows. We introduce related works about GAN-based neural vocoders and low-latency speech generation methods in Section II and describe our proposed DLL-APNet in Section III. The experimental setups and results are presented in Section IV and V, respectively.

* Corresponding author. This work was funded by the Anhui Province Major Science and Technology Research Project under Grant S2023Z20004, the National Nature Science Foundation of China under Grant 62301521 and the Anhui Provincial Natural Science Foundation under Grant 2308085QF200.

979-8-3315-7206-8/25/$31.00

Finally, we give conclusion in Section VI.

II. Related Work

Since the benchmark employed in this study utilizes GAN-based neural vocoders and prioritizes causal modeling to reduce latency, this section offers a concise review of GAN-based neural vocoders and recent approaches to low-latency speech generation.

A. GAN-based Neural Vocoders

GAN-based neural vocoders represent one of the most dominant frameworks in modern vocoding technique, typically composed of a generator and a discriminator. We can classify GAN-based neural vocoders into non-all-frame-level and all-frame-level types based on whether upsampling operations are adopted. For non-all-frame-level neural vocoders (e.g., BigVGAN [19], HiFi-GAN [1], and iSTFTNet [20]), mel spectrograms are either left unprocessed or first upsampled to a higher temporal resolution before being directly converted into time-domain waveforms through transposed convolutional layers. However, upsampling operations pose challenges for deployment on devices that lack parallel computing capabilities, as they introduce sequential dependencies and computational overhead. In contrast, all-frame-level neural vocoders (e.g., APNet [12], APNet2 [16], and Vocos [2]) predict amplitude and phase spectrum at the same frame rate as the input mel spectrogram through convolutional operations. The speech waveform is finally reconstructed via the inverse short-time Fourier transform (iSTFT), leveraging the spectral information to synthesize natural-sounding speech. Our previously proposed APNet vocoder [12] achieved explicit amplitude and phase spectrum prediction with small frame shifts for speech waveform reconstruction, especially leveraging parallel estimation architecture and anti-wrapping loss function for accurate phase estimation. Our previously proposed APNet2 vocoder [16] advanced APNet by integrating ConvNeXt v2 blocks [21] and improved discriminators, enabling high-sampling-rate spectral prediction with large frame shifts. However, these approaches all overlooked latency issues, resulting in substantial model delays.

B. Low-Latency Speech Generation Methods

Latency, defined as the minimum amount of input time required to initiate the model, is a critical factor in many speech generation systems. It determines whether such models can be applied in real-time scenarios, such as speech communication. Causal models exhibit extremely low model latency, with relevant research conducted in several low-latency speech generation methods, e.g., speech coding [17], [18], [22], VC [23], SE [24] and speech phase prediction [13]. One approach [22] is to replace standard convolutions with linear layers and employ knowledge distillation strategies to enhance model performance. However, linear layers only allow the current output to access the current input rather than past inputs, presenting certain limitations. Another approach [17], [18] is to use causal convolutions instead of standard convolutions and employ knowledge distillation strategies, avoiding the issue where standard convolutions with symmetric padding can access future information. However, low-latency neural vocoders have not yet been thoroughly investigated.

III. Proposed Method

An overview of the proposed DLL-APNet vocoder is shown in Figure 1. Building upon our previously proposed APNet2 vocoder [16], the input mel spectrogram is processed by a amplitude spectrum prediction branch and a phase spectrum prediction branch in parallel to explicit predict the amplitude and phase spectra, respectively. The time-domain waveform is then reconstructed via iSTFT. To support low latency, all convolutional layers in the DLL-APNet vocoder are causal. At the training stage, a pre-trained APNet2 vocoder is used as a teacher model to generate intermediate features, which guide the distillation-based training of the proposed DLL-APNet vocoder. Details of the model architecture and training procedures will be presented below.

A. Causal Convolution

In scenarios requiring extremely high real-time performance, the latency introduced by models must be strictly constrained. However, for ordinary convolutions, the receptive field of the convolution kernel can access future information, which introduces model latency. For example, for a standard convolution with kernel size k and dilation d , the number of future input frames required is

$$\zeta(k,d) = \left\lfloor \frac{(k-1)d}{2} \right\rfloor, \tag{1}$$

where $\lfloor \cdot \rfloor$denotes flooring. Most GAN-based neural vocoders, structured as stacks of non-causal convolutional layers, exhibit increasing model latency with deeper architectures, which poses certain challenges for real-time applications.

Causal convolutions layers, as illustrated in Figure 2, avoid using future information by asymmetrically padding the input sequence $\boldsymbol{x}_T = [x_1, x_2, ..., x_T]^\top$. Due to asymmetric padding, the last element of the convolution kernel at the current time step t corresponds to the t-th element of the input sequence, which ensures that the output value $\boldsymbol{y}_T = [y_1, y_2, ..., y_T]^\top$ at time step t depends only on the input $\boldsymbol{x}_{\leq t}$, forming a fully causal operation.

B. Model Structure

As illustrated in Figure 1, for both amplitude and phase prediction in the DLL-APNet vocoder, causal convolutions are first employed for input mel spectrogram, followed by K causal ConvNeXt v2 [21] blocks for deep feature extraction. The original ConvNeXt v2 block consists of a depth-wise convolution with a large kernel, two point-wise convolutions, along with normalization layers and activation functions inserted between them. Since the point-wise convolutional layers are linear and do not introduce latency, we only convert the original depth-wise convolution into a causal convolution with the same kernel size to minimize model latency, and named

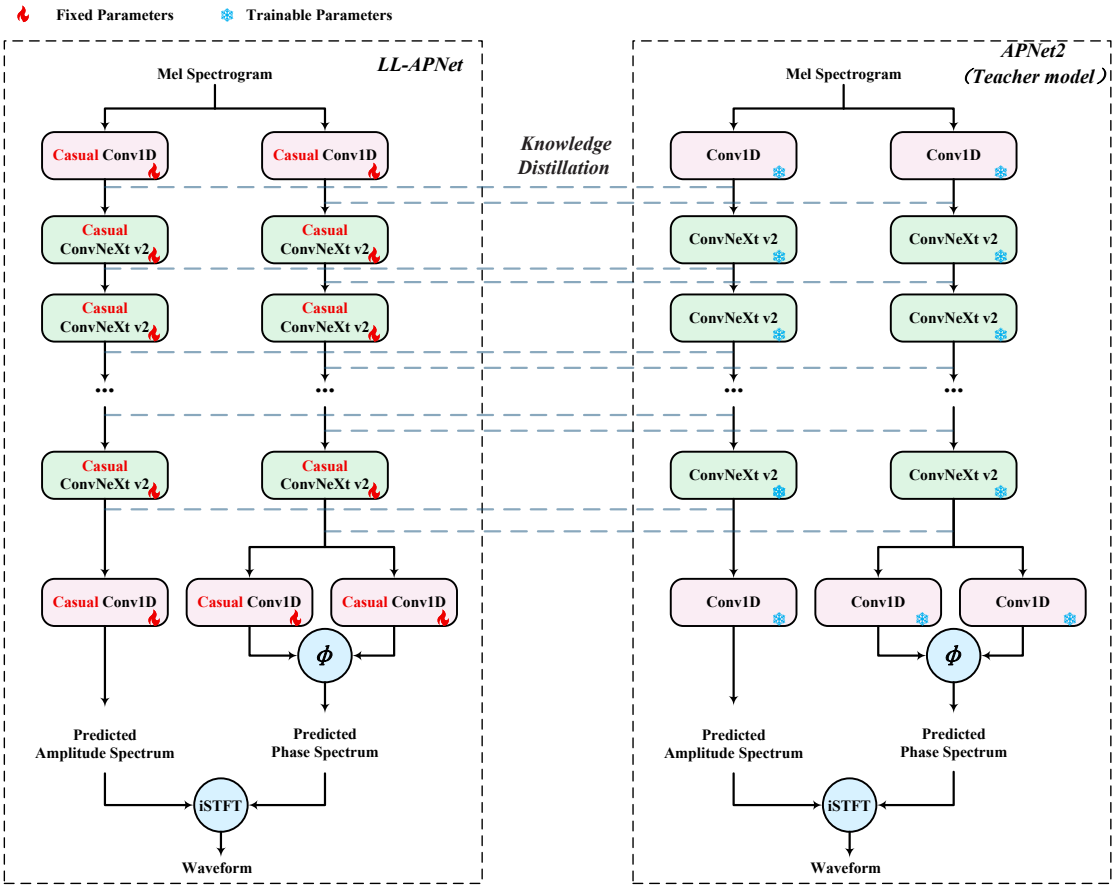


Fig. 1: Model structure and knowledge distillation strategy of the proposed DLL-APNet vocoder, where ϕ denotes the phase calculation formula. During training, part of the parameters of DLL-APNet are trained through knowledge distillation from teacher model APNet2.

this module as causal ConvNeXt v2 block. At the output of the amplitude prediction branch, the output of the last causal ConvNeXt v2 block undergoes post-processing via causal convolutions to yield the predicted amplitude spectrum. In contrast to the amplitude prediction branch, the phase prediction branch employs a causal parallel estimation architecture (PEA) [13] to enable explicit phase spectrum prediction. In causal PEA, two parallel causal convolutional layers are first used to estimate the pseudo-real and pseudo-imaginary components, and then they are activated using the two-argument arctangent function to estimate the wrapped phase directly.

C. Knowledge-Distillation-based Training Criteria

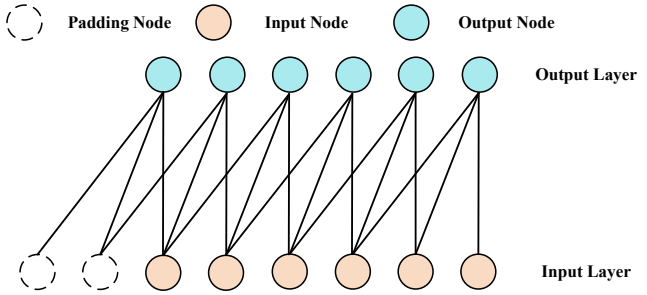
Knowledge distillation strategy is commonly employed to leverage the superior learning capabilities of a teacher model to guide the generation of intermediate features in a weaker student model, thereby enhancing the latter's performance. Specifically, we first trained a non-causal APNet2 vocoder, which serves as the teacher model. Given that causal convolutions can only access current and past information, we leverage features extracted by non-causal modules in APNet2, which are capable of utilizing future context, to guide the causal modules in DLL-APNet student model and enable it to implicitly learn temporal dependencies within causal constraints. As illustrated in Figure 1, knowledge distillation is performed after both the input convolution and each ConvNeXt v2 block. The L1 distance between the intermediate features of the teacher and student models, serves as the knowledge distillation loss. Specifically, we define the output of the input convolutional layer and the k-th ConvNeXt v2 block ($k = 1, \ldots, K$) of teacher model (i.e., APNet2) as $\hat{\boldsymbol{O}}$ and $\hat{\boldsymbol{O}}_k^{CNX}$. The outputs of the student model (i.e., DLL-APNet) at corresponding positions are respectively denoted as $\tilde{\boldsymbol{O}}$ and $\tilde{\boldsymbol{O}}_k^{CNX}$, then the knowledge distillation loss can be defined as:

$$\mathcal{L}_{KD} = \mathbb{E}_{(\hat{\boldsymbol{O}},\tilde{\boldsymbol{O}})} \left\| \hat{\boldsymbol{O}} - \tilde{\boldsymbol{O}} \right\|_1 + \sum_{k=1}^{K} \mathbb{E}_{(\hat{\boldsymbol{O}}_k^{CNX},\tilde{\boldsymbol{O}}_k^{CNX})} \left\| \hat{\boldsymbol{O}}_k^{CNX} - \tilde{\boldsymbol{O}}_k^{CNX} \right\|_1 . \quad (2)$$

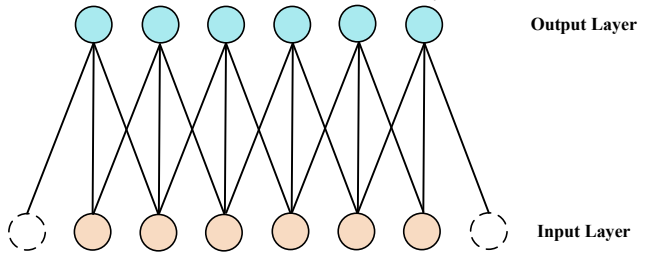
We employed multi-resolution discriminator (MRD) [25] and multi-period discriminator (MPD) [1] to supervise the generated waveforms, ensuring the quality of the synthesized waveforms across various scales and frequency bands. For the traning loss, we also introduce the amplitude loss $\mathcal{L}_A$, phase loss $\mathcal{L}_P$, reconstructed STFT loss $\mathcal{L}_S$, and final waveform loss

TABLE I: Objective evaluation results of the proposed DLL-APNet vocoder and baselines on the VCTK test set.

	Causality	SNR (dB) ↑	LAS-RMSE (dB)↓	MCD (dB)↓	F0-RMSE (cent)↓	V/UV error (%)↓	UTMOS ↑	GFLOPS ↓
Natural	-	-	-	-	-	-	4.04	-
BigVGAN	✗	6.42	3.63	0.90	21.04	3.22	3.97	230.51
HiFi-GAN	✗	4.15	4.51	1.58	31.61	3.97	3.93	25.65
iSTFTNet	✗	4.15	5.24	1.87	32.87	4.13	3.93	19.22
APNet2	✗	6.56	4.23	0.99	17.38	2.88	4.00	6.30
Vocos	✗	6.05	3.70	0.80	25.17	3.47	3.91	2.70
causal HiFi-GAN	✓	3.00	5.24	2.32	58.06	5.96	3.88	25.66
causal iSTFTNet	✓	2.22	5.84	2.27	54.43	6.37	3.75	19.23
causal APNet2	✓	3.63	4.23	1.63	26.06	4.10	3.90	6.30
causal Vocos	✓	5.32	4.33	0.89	32.87	4.32	3.87	2.70
DLL-APNet	✓	6.07	4.29	1.04	20.53	3.16	3.98	6.30



(a) causal convolutional layer.



(b) Non-causal convolutional layer.

Fig. 2: Illustration of the mapping relationship of causal and non-causal convolutions, taking a 3×1 convolution kernel with stride = dilation = 1 as an example.

$\mathcal{L}_W$ used in APNet2 [16], and combine them for adversarially training DLL-APNet, i.e.,

$$\mathcal{L} = \lambda_A \mathcal{L}_A + \lambda_P \mathcal{L}_P + \lambda_S \mathcal{L}_S + \lambda_W \mathcal{L}_W + \lambda_{KD} \mathcal{L}_{KD}, \quad (3)$$

where $\lambda_P, \lambda_A, \lambda_S$, λ_W, and λ_{KD} are hyperparameters.

IV. Experiments Setups

A. Dataset

For our experimental setup, the VCTK-0.92 dataset [26] was utilized, with all speech utterances downsampled to a 16 kHz sampling rate. This corpus comprises recordings from 108 English-speaking individuals, totaling approximately 44 hours of speech material. Regarding data partitioning, we first extracted samples from 100 speakers. Of these, 90% were randomly allocated to the training subset, while the remaining 10% constituted the validation set. For the test set, 2,937 utterances were specifically chosen from the remaining 8 unseen speakers, ensuring a disjoint evaluation corpus.

B. Implementation

For our proposed DLL-APNet vocoder[1], the amplitude and phase spectra were computed using the STFT with a frame length, frame shift, and FFT size of 320, 80, and 1024, respectively. The mel spectrogram was extracted with the same configuration, with a dimensionality of 80. We set the hyperparameters as $K = 8$, $\lambda_P = 100$, $\lambda_A = 45$, and $\lambda_S = 1$ as the configuration in APNet2 [16]. λ_{KD} was set to 5 in the main experiment and we will discuss its selection in Section V-B1. The model was trained using the AdamW optimizer for up to 0.5 million steps.

C. Baselines

We compared DLL-APNet with BigVGAN [19], HiFi-GAN [1], iSTFTNet [20], APNet2 [16], and Vocos [2]. We reproduced the experimental results using the methods described in their original paper under our experimental implementation. For fair comparison, We also reproduced their causal versions by replace their origin non-causal convolutional layers with causal convolutional layers.

D. Evaluation Metrics

In the present research, we utilized five objective metrics for evaluating the quality of synthesized speech. These metrics consist of the signal-to-noise ratio (SNR), root mean square error (RMSE) of log amplitude spectra (LAS-RMSE), mel-cepstrum distortion (MCD), root mean square error of fundamental frequency (F0-RMSE), and voiced/unvoiced (V/UV) error. And we introduced the neural evaluation metric UTMOS [27] as an objective auditory perception metric. Moreover, the computational complexity of each model was gauged by the floating-point operations (FLOPs) needed for generating 1-second speech.

[1]Examples of generated speech can be found at our demo page https://redmist328.github.io/DLL-APNet/.

V. Results and Analysis

A. Main Experimental Results

Table I presents the objective experimental results of the proposed DLL-APNet and baseline vocoders evaluated on the test set. Comparative analysis between mainstream vocoders and their causal variants reveals that replacing non-causal convolutions with causal counterparts leads to varying degrees of degradation across all speech quality metrics, indicating an inevitable trade-off between low latency and synthesis fidelity. Particularly in F0-related metrics (i.e., F0-RMSE and V/UV error), causal models exhibit substantial performance drops compared to their non-causal counterparts, indicating that relying solely on past and current information for prediction may lead to insufficient information capture, thereby affecting pronunciation accuracy. However, FLOPs measurements remain nearly unchanged after causal transformation of the models, suggesting that this process imposes minimal impact on computational efficiency. Notably, the causal variant of BigVGAN failed to converge during training, rendering it incapable of generating intelligible speech, thus we didn't present its results in Table I. This highlights that causal adaptation is not universally applicable to all vocoder architectures.

A comparison between DLL-APNet and causal vocoders reveals that among all causal vocoders, the proposed DLL-APNet significantly outperformed others in most speech quality metrics, validating the effectiveness of our proposed method. In Table I, the causal APNet2 represents the result of training DLL-APNet without using the knowledge distillation strategy (i.e., $\mathcal{L}_{KD} = 0$). It can be seen that the SNR, F0-RMSE, V/UV error, and UTMOS metrics of the causal APNet2 all showed significant decreases compared to original APNet2. After introducing the knowledge distillation training strategy, these metrics all improve to a certain extent and approach the level of the original APNet2, indicating the effectiveness of the knowledge distillation strategy. Although DLL-APNet ranks second only to causal Vocos in terms of FLOPS, it surpasses causal Vocos by 0.09 in UTMOS while maintaining significantly lower FLOPS than other vocoders. These results provide theoretical justification for deploying DLL-APNet in real-time and resource-constrained scenarios.

When comparing the proposed DLL-APNet with non-causal neural vocoders, its objective speech quality metrics are comparable to those of other vocoders. Additionally, while its UTMOS score is second only to APNet2 and higher than those of other vocoders, it maintains relatively low computational complexity. This demonstrates that the proposed DLL-APNet delivers speech quality on par with mainstream vocoders, while also providing low latency and computational efficiency.

B. Analysis and Discussion

1) Discussion on Distillation Weight Selection: To investigate the impact of the distillation weight λ_{KD} on model performance, we conducted a series of controlled-variable experiments on the hyperparameter of knowledge distillation.

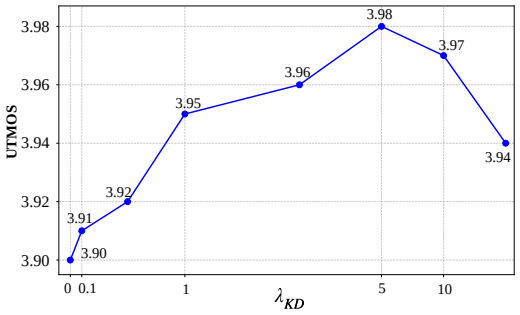


Fig. 3: The UTMOS curve of DLL-APNet as the distillation weight λ_{KD} varies.

TABLE II: UTMOS results of DLL-APNet with different distilled numbers of ConvNeXt v2 blocks.

Distilled Numbers	0	2	4	6	8
UTMOS	3.73	3.92	3.95	3.96	3.98

Specifically, we set λ_{KD} to 0 (meaning no knowledge distillation strategy is used), 0.1, 0.5, 1, 2, 5, 10, and 20, and explored the UTMOS performance of the DLL-APNet. The results are shown in Figure 3. It can be seen that when the hyperparameter value is small, UTMOS increases as the hyperparameter value increases, indicating the effectiveness of the knowledge distillation strategy we used. When the hyperparameter reaches 5, UTMOS reaches the maximum value, and then decreases as the hyperparameter increases. This suggests that λ_{KD} should be appropriately set during training. Too small values would prevent the knowledge distillation strategy from functioning effectively, while excessively large values could cause the model to overly focus on the distillation loss, neglecting other important objectives such as direct supervision from natural speech, ultimately degrading speech quality.

2) Discussion on Distillation Position Selection: To investigate the impact of distillation positions on model performance of DLL-APNet, we designed analytical experiments by varying the number of distilled layers in the backbone ConvNeXt v2 blocks. Specifically, in our setup with 8 ConvNeXt v2 blocks, we trained models by varying the number of blocks involved in knowledge distillation to 0, 2, 4, 6, and 8, respectively. The UTMOS results are shown in Table II. It can be observed that speech quality increased with the number of involved blocks, indicating that the effect of knowledge distillation enhanced as the model's participation degree increased. These findings also provide useful guidance for the design of knowledge distillation strategies in other speech generation tasks.

VI. Conclusion

This paper presents a novel causal low-latency neural vocoder, DLL-APNet, which explicitly predicts amplitude and phase spectra from input mel spectrogram using causal-convolution-based predictors and reconstructs speech waveform via iSTFT. To enhance speech quality under causality constraints, we employ a pre-trained non-causal APNet2

vocoder as the teacher model to guide the intermediate feature generation of the proposed DLL-APNet vocoder. Experimental results demonstrate that DLL-APNet synthesizes speech with higher quality than other causal vocoders and comparable to mainstream non-causal vocoders while requiring fewer computations. Future work will focus on reducing model size and computational overhead to better suit practical vocoder applications.

REFERENCES

[1] 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.

[2] H. Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," in *Proc. ICLR*, 2024.

[3] R. Mira et al., "LA-VocE: Low-SNR audio-visual speech enhancement using neural vocoders," in *Proc. ICASSP*, 2023, pp. 1–5.

[4] D. Cao, Z. Zhang, and J. Zhang, "NeuralVC: Any-to-any voice conversion using neural networks decoder for real-time voice conversion," *IEEE Signal Processing Letters*, 2024.

[5] A. v. d. Oord et al., "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[6] S. Mehri et al., "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. ICLR*, 2017.

[7] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, 2019, pp. 3617–3621.

[8] W. Ping, K. Peng, K. Zhao, and Z. Song, "WaveFlow: A compact flow-based model for raw audio," in *Proc. ICML*, 2020, pp. 7706–7716.

[9] T. D. Nguyen, J.-H. Kim, Y. Jang, J. Kim, and J. S. Chung, "Fregrad: Lightweight and fast frequency-aware diffusion vocoder," in *Proc. ICASSP*, 2024, pp. 10 736–10 740.

[10] Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, "Specgrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping," in *Proc. Interspeech*, 2022, pp. 803–807.

[11] I. Goodfellow et al., "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[12] Y. Ai and Z.-H. Ling, "APNet: An all-frame-level neural vocoder incorporating direct prediction of amplitude and phase spectra," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2145–2157, 2023.

[13] Y. Ai and Z.-H. Ling, "Low-latency neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses for speech generation tasks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2283–2296, 2024.

[14] H. Li et al., "Single-Codec: Single-codebook speech codec towards high-performance speech generation," in *Proc. Interspeech*, 2024, pp. 3390–3394.

[15] R. Langman, A. Jukić, K. Dhawan, N. R. Koluguri, and B. Ginsburg, "Spectral codecs: Spectrogram-based audio codecs for high quality speech synthesis," *arXiv preprint arXiv:2406.05298*, 2024.

[16] H.-P. Du, Y.-X. Lu, Y. Ai, and Z.-H. Ling, "APNet2: High-quality and high-efficiency neural vocoder with direct prediction of amplitude and phase spectra," in *Proc. NCMMSC, 2023*, pp. 66–80.

[17] Y.-C. Wu, I. D. Gebru, D. Marković, and A. Richard, "AudioDec: An open-source streaming high-fidelity neural audio codec," in *Proc. ICASSP*, 2023, pp. 1–5.

[18] 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.

[19] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *Proc. ICLR*, 2023.

[20] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform," in *Proc. ICASSP*, 2022, pp. 6207–6211.

[21] S. Woo et al., "ConvNeXt v2: Co-designing and scaling convnets with masked autoencoders," in *Proc. CVPR*, 2023, pp. 16 133–16 142.

[22] Y. Ai, X.-H. Jiang, Y.-X. Lu, H.-P. Du, and Z.-H. Ling, "APCodec: A neural audio codec with parallel amplitude and phase spectrum encoding and decoding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3256–3269, 2024.

[23] Z. Ning et al., "DualVC: Dual-mode voice conversion using intra-model knowledge distillation and hybrid predictive coding," in *Proc. Interspeech*, 2023, pp. 2063–2067.

[24] E. Tsunoo, Y. Saito, W. Nakata, and H. Saruwatari, "Causal speech enhancement with predicting semantics based on quantized self-supervised learning features," in *in Proc. ICASSP*, 2025, pp. 1–5.

[25] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," in *Proc. Interspeech*, 2021, pp. 2207–2211.

[26] C. Veaux, J. Yamagishi, K. MacDonald, et al., "Superseded-CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[27] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.