# PALGAN: A Joint Optimization-Based Preprocessing method for Speech Restoration in Parametric Array Loudspeakers

Wenyao Ma*† and Jun Yang*†
*Laboratory of Noise and Audio Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
†University of Chinese Academy of Sciences, Beijing, China
E-mail: jyang@mail.ioa.ac.cn

***Abstract*—The parametric array loudspeaker (PAL) inherently introduces baseband distortions in directional sound applications due to the nonlinear process in air. Recently, a two-stage DNN-based framework has been used to generate preprocessed signals for distortion-free speech restoration, where a preprocessing network is cascaded with a frozen forward model that estimates the PAL system. However, the preprocessing network can exploit weaknesses in the fixed model, resulting in adversarial outputs. To address this, we propose a joint optimization strategy based on a PAL-driven generative adversarial network (PALGAN), which simultaneously learns the PAL process and the preprocessing module. A process discrimination mechanism is incorporated into the discriminator, enabling it to determine whether the input signal pair conform to the PAL process while acquiring a forward model. The generator uses a causal TF-GridNet to estimate the real and imaginary parts of the preprocessed signal via a time–frequency band-split mechanism. Simulated and real-world experiments demonstrate that the proposed method achieves competitive reconstruction performance when compared with existing methods across various objective metrics[1].**

## I. Introduction

The parametric array loudspeaker (PAL) is a directional sound technique that produces a narrow sound beam with a significantly smaller aperture than conventional loudspeakers. It leverages the self-demodulation (SD) property of high-intensity ultrasonic waves in air to generate audible difference-frequency components. Recently, the PAL has been applied to private audio projection in smart devices such as computers and televisions [1], reigniting interest in distortion-free speech reproduction through preprocessing techniques.

The goal of the preprocessing modulator is to proactively compensate for baseband distortions inherent in the SD process [2]. In the past few decades, forward modeling and preprocessing techniques have developed in an alternating manner. The Berktay's far-field solution is given as a simple expression in the time domain, predicting the baseband output $y\,(t)$ after SD process by:

$$y(t) \propto \frac{\partial^2 E^2(t)}{\partial t^2}, \tag{1}$$

where $E(t)$ is the envelope signal. The double-sideband (DSB) amplitude modulation was first used to generate wideband primary ultrasonic waves by modulating audio signals onto an ultrasonic carrier [3]. So the envelope function can be written as $1 + mx(t)$ in the Berktay's far-field solution, where $x(t)$ is the input audio and $m$ is the modulating index. The square operation in Eq. (1) introduces nonlinear distortion that contains harmonic and inter-modulated components. For speech signals, harmonics introduced by nonlinearity often coincide with the original spectral components, resulting in spectral overlap and reduced speech quality. The inter-modulated terms usually manifest as non-stationary noise [4]. The second derivative in Eq. (1) results in a 12 dB gain per octave when compared with clean speech.

Several modulators have been developed to suppress specific types of distortion. The square-root (SRT) amplitude modulation [5], [6] is designed to theoretically eliminate square-law distortion in the far-field model. However, this requires infinite bandwidth from both the system and the transducer. For practical implementation, a truncated SRT modulation method was proposed by taking the truncated Taylor series expansion of the ideal SRT signal [7], [8]. Furthermore, the $N$th-order equalization (EQ) method reduces the error introduced by single-sideband (SSB) modulation through iterative refinement [9], [10]. Each recursion includes an error correction circuit, where the difference between current demodulated output and target signal is fed back to generate the next preprocessed input. When applied to speech signals, this method has been reported to introduce slight higher-order distortion components at each recursive step, which degrades speech quality [4], [11].

Although well-developed, Berktay's far-field solution and the corresponding preprocessing modulator are only valid within the axial far-field range, which significantly exceeds the distances typically encountered in most practical appli-

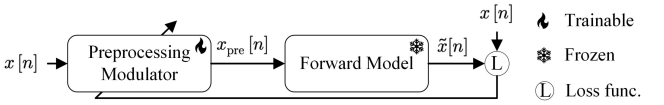
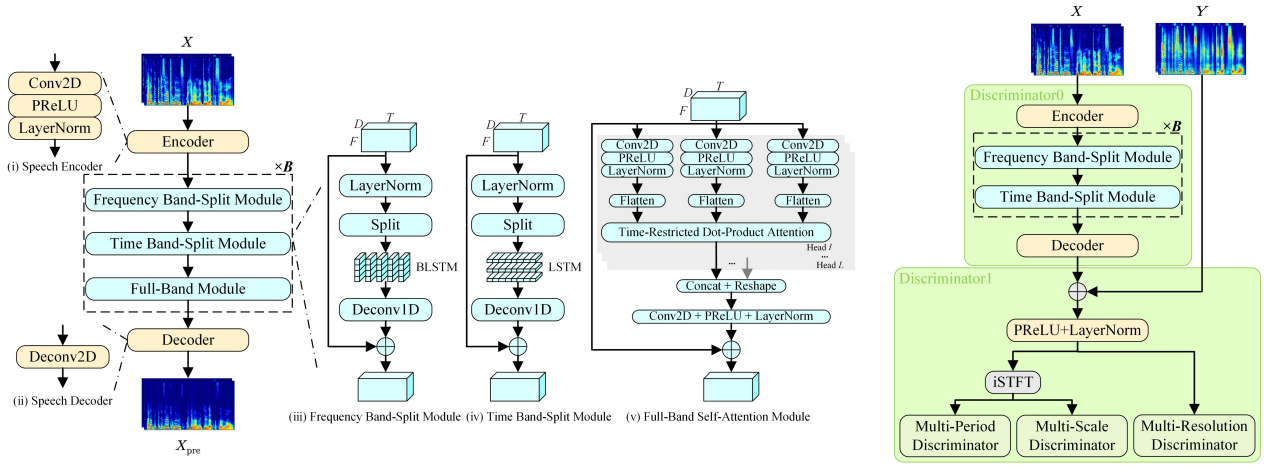



Fig. 1. A two-stage framework for training the preprocessing modulator, based on a frozen and pre-trained forward model.

[1]For access to the source code and speech samples, please visit: https://github.com/MWY0615/PALGAN.

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




(a) Generator architecture

(b) Process discriminator architecture

Fig. 2. An overview of the proposed PALGAN architecture.

cations [12]. To address this limitation, the Volterra filter (VF) was subsequently introduced for SD modeling, offering flexibility across various acoustic fields, particularly in the near field [13], [14]. The one-dimensional Volterra filter (ODVF) enhances the VF by extracting diagonal kernel, thereby reducing computational complexity and improving convergence behavior. However, the $p$th-order inverse filter employed to compensate for the ODVF is effective only for single-frequency signals [15], [16]. Unlike these VF-based models, which are heuristically introduced into the PAL system, our previous work proposed an interpretable model called the Differential Volterra Filter (DiffVF), derived via a two-stage decoupling method based on the Westervelt equation [17]. A recursive equalizer, specifically tailored to the DiffVF model, offers an effective solution for speech reconstruction in the near field [4]. Despite improvements in objective metrics, the method remains limited by high-order distortions and the computational cost inherent to recursive operations.

DNN-based approaches have been applied to the identification of nonlinear systems due to their powerful function approximation capabilities and hierarchical nonlinear structures [18], [19]. A two-stage framework tailored for speech scenarios was first proposed in [4], consisting of a forward model for PAL system modeling and a preprocessing modulator, both sharing the same network architecture. Fig. 1 shows the detailed training strategy. A similar framework is later adopted for single-frequency scenarios, with WaveNet serving as the core network [20]. Although a shared network architecture facilitates the training of the cascaded preprocessing module, limited diversity in recorded data may cause the preprocessing modulator to exploit weaknesses in the frozen forward model, producing adversarial outputs that are well restored within the model but fail in a real PAL system.

To this end, we proposed a joint optimization strategy implemented by a PAL-driven generative adversarial network (PALGAN). A process discrimination mechanism is introduced into the discriminator to distinguish between real and estimated input signal pairs, where each pair consists of an input and its corresponding PAL output. In addition, the intermediate output of the discriminator contributes explicit supervision to the forward model. Both the explicit non-adversarial loss and the implicit adversarial loss jointly guide the generator to produce effective preprocessed signals for speech restoration. The generator adopts a causal TF-GridNet based on [21], using a T-F band-split mechanism to estimate real and imaginary (RI) components. Objective comparisons with baselines demonstrated the effectiveness of the proposed method in real-world experiments.

## II. Methodology

In this section, we first formulate the physical process. Then, the generator and the process discriminator are illustrated respectively. Finally, we describe the loss functions used in PALGAN.

### A. Formulation of Preprocessing

Our goal is to design a preprocessing modulator, making the obtained preprocessed signal reproduce the clean speech after the SD process. For convenience, we define a general self-demodulation (GSD) process, as a baseband mapping from speech input to the speech output that incorporates the effects of DSB modulator, ultrasonic transducer, and SD in air. Thus the restored speech $\hat{x}$ can be formulated as:

$$\hat{x}[n] = \mathcal{M}(\mathcal{G}(x[n]))\,, \tag{2}$$

where $\mathcal{M}(\cdot)$ denotes the GSD process, $\mathcal{G}(\cdot)$ denotes the preprocessing modulator that generates the preprocessed signal $x_{\text{pre}}$ to be modulated and emitted into the air by the transducer, and $x$ denotes the target speech. During the design of the preprocessing modulator, the estimated GSD process serves as a substitute for the real process and is termed the forward model.

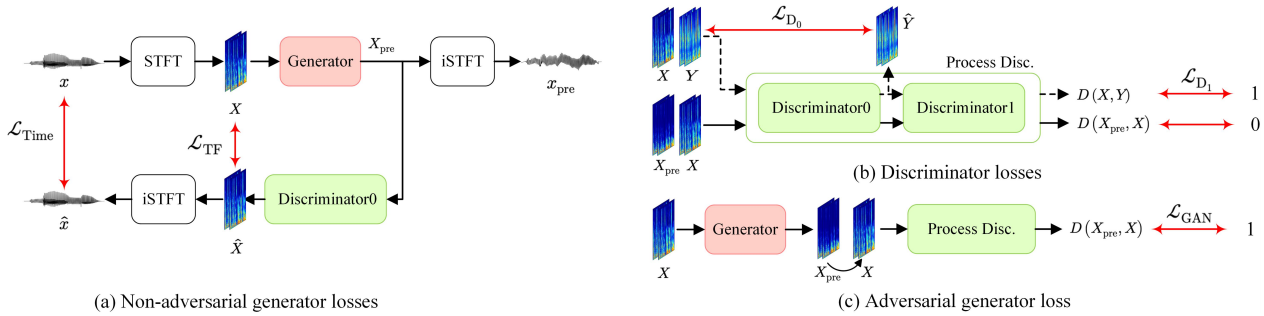



Fig. 3. An illustration of the propagated loss functions in the PALGAN architecture. For simplicity, $X$ and $Y$ denote the complex spectrograms of the clean input signal and the recorded PAL output, respectively, while $X_{\text{pre}}$ represents the spectrogram of the preprocessed signal produced by the generator.

### B. Generator architecture

The preprocessing modulator aims to map the T-F representation of the clean speech to the preprocessed signal under implicit supervision of the forward model, described as: $X_{\text{pre}} = \mathcal{G}(X;\Phi)$, where $X$ and $X_{\text{pre}}$ are the complex spectra of the target speech and the preprocessed signal respectively, the frame and frequency indices $\{t, f\}$ are omitted for clarity, and $\Phi$ denotes the trainable parameters. Fig. 2(a) depicts the internal structure of the preprocessing modulator, designed as a causal variant of TF-GridNet [21], [22], which also functions as the generator in PALGAN. It features an encoder module, causal TF-GridNet blocks and a decoder module. The encoder applies a two-dimensional convolution (Conv2D) layer to the speech spectra and then uses a PReLU and a layer normalization (LayerNorm) to produce a $D$-dimensional feature vector for each T-F bin. Correspondingly, the decoder reconstructs the preprocessed speech in the T-F domain through a two-dimensional deconvolution (Deconv2D).

The $B$-layer causal TF-GridNet blocks, adapted for real-time causal processing, leverage both full- and sub-band spectral-temporal information to predict the embedding of preprocessed speech. It comprises three submodules. For the frequency band-split module, the input tensor $R_b$ with shape $D \times T \times F$ in the $b$th block is viewed as $T$ separate sequences, such that each $D \times F$ unit can be processed with a sequence model, where $T$ and $F$ are the number of frames and frequency bins respectively. To aggregate neighboring features, the band-split mechanism is applied by LayerNorm and unfolding the sequences along the frequency axis using a kernel size of $I$ and stride of $J$. This allows the succeeding bidirectional LSTM (BLSTM) to capture the cross-band correlations within flattened embeddings. The formulation can be written as:

$$\dot{R}_b = \Big[\texttt{BLSTM}\left(\texttt{Unfold}\left(\texttt{LN}\left(R_b\left[:,t,:\right]\right)\right)\right), \quad \text{for } t = 1,\ldots,T\Big] \in \mathbb{R}^{2M \times T \times \lceil \frac{F-I}{J}+1 \rceil}, \tag{3}$$

where $M$ denotes the hidden size in each direction of BLSTM. Next, a one-dimensional deconvolution (Deconv1D) layer with output channel of $D$ is performed on $R_b$, whose result is added to $R_b$ via a residual connection to produce the final output.

The time band-split module mirrors the above procedure but differs in its band-split mechanism. The input tensor is viewed as $F$ separate sequences of length $T$, stacking neighboring features along the time axis. An LSTM is then used to capture the temporal information within each frequency band, while maintaining causality.

In the full-band self-attention module, an $L$-head self-attention mechanism is employed to model long-range global information within each frame [23]. The frame-level query, key, and value vectors are computed and flattened to construct the attention matrix based on the dot-product attention mechanism.

### C. Process Discriminator

Instead of fully relying on explicitly minimizing the losses from regression approaches, which may cause overfitting problems, the discriminator provides a high-level abstract measure of realness. Here we propose a process discriminator for joint optimization, consisting of two modules: Discriminator0 and Discriminator1. Fig. 2(b) shows the internal structure of the discriminator. Discriminator0 resembles a forward model, because the difference between estimated output $\hat{Y}$ and the real PAL output $Y$ constitutes part of the discriminator loss, as shown in Fig. 4(b). However, it differs by being fully trainable. Moreover, a process discrimination mechanism is incorporated to Discriminator1 to judge whether each input signal pair is real or fake PAL process, thereby offering the preprocessing modulator an auxiliary form of supervision. For example, when the $X$ and its corresponding real PAL output $Y$ serve as the input pairs, the discriminator identifies this pair as a true PAL process.

Discriminator0 comprises an encoder, $B$-layer blocks each containing a frequency band-split module and a time band-split module, and a decoder. Each module is identical to those used in the generator. The output of Discriminator0, along with $Y$ is passed to Discriminator1. The process discrimination mechanism is implemented by adding the output to $Y$ to form a residual connection, followed by a PReLU and LayerNorm, and then fed into three sub-discriminators. We adopt a multi-period (MPD) and a multi-scale (MSD) waveform discriminators, which improve audio fidelity [25], [26]. In addition, a complex STFT discriminator at multiple resolution (MRD) [27], [28] is employed because it leads to improved magnitude and phase modeling.

TABLE I
THE HYPERPARAMETER CONFIGURATION FOR PALGAN.

| Model | $B$ | $L$ | $I$ | $J$ | $M$ | $D$ | Param. | MACs |
|---|---|---|---|---|---|---|---|---|
| Generator | 4 | 4 | 8/8 | 8/8 | 32 | 256 | 4.87 M | 19.53 G/s |
| Discriminator0 | 4 | – | 6/24 | 6/24 | 16 | 192 | 14.83M | 8.55G/s |

### D. Loss Functions

Inspired by [29], we use a linear combination of magnitude loss $\mathcal{L}_{\text{Mag.}}$ and complex loss $\mathcal{L}_{\text{RI}}$ in the T-F domain:

$$\mathcal{L}_{\text{TF}} = \alpha \mathcal{L}_{\text{Mag.}} + (1-\alpha)\,\mathcal{L}_{\text{RI}}, \tag{4}$$

$$\mathcal{L}_{\text{Mag.}} = \mathbb{E}_{X_m,\hat{X}_m}\left[\left\|X_m - \hat{X}_m\right\|_2\right], \tag{5}$$

$$\mathcal{L}_{\text{RI}} = \mathbb{E}_{X_i,\hat{X}_i}\left[\left\|X_i - \hat{X}_i\right\|_2\right] + \mathbb{E}_{X_r,\hat{X}_r}\left[\left\|X_r - \hat{X}_r\right\|_2\right], \tag{6}$$

where subscripts $\{m, r, i\}$ denote the magnitude and RI components of the complex spectrograms, and $\alpha$ is a weighting factor typically set to 0.5 [30]. Moreover, an additional penalization in the waveform $\mathcal{L}_{\text{Time}}$ is proven to improve the restored speech quality [31]

$$\mathcal{L}_{\text{Time}} = \mathbb{E}_{x,\hat{x}}\left[\|x - \hat{x}\|_1\right]. \tag{7}$$

In this context, $\|\cdot\|_1$ denotes the $\ell_1$ distance and $\|\cdot\|_2$ denotes the $\ell_2$ distance. $\mathcal{L}_{\text{TF}}$ and $\mathcal{L}_{\text{Time}}$ together constitute the non-adversarial loss for the generator. Explicit supervision for Discriminator0 is implemented by minimizing the loss $\mathcal{L}_{\text{D}_0}$:

$$\mathcal{L}_{\text{D}_0} = \mathbb{E}_{Y,\hat{Y}}\left[\left\|Y - \hat{Y}\right\|_2\right]. \tag{8}$$

Similar to least-square GANs [32], which emplying binary coding (1 for real, 0 for fake), the adversarial training is following a min-min optimization task over the adversarial generator loss $\mathcal{L}_{\text{GAN}}$ and the adversarial discriminator loss $\mathcal{L}_{\text{D}_1}$:

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{X_{\text{pre}},X}\left[\|D\left(X_{\text{pre}}, X\right) - 1\|_2\right], \tag{9}$$

$$\mathcal{L}_{\text{D}_1} = \mathbb{E}_{X,Y}\left[\|D\left(X, Y\right) - 1\|_2\right] + E_{X_{\text{pre}},X}\left[\|D\left(X_{\text{pre}}, X\right)\|_2\right], \tag{10}$$

where $D$ refers to the progress discriminator. The final generator loss and discriminator loss are formulated as follows, respectively:

$$\mathcal{L}_{\text{G}} = \gamma_1 \mathcal{L}_{\text{TF}} + \gamma_2 \mathcal{L}_{\text{Time}} + \gamma_3 \mathcal{L}_{\text{GAN}}, \tag{11}$$

$$\mathcal{L}_{\text{D}} = \beta \mathcal{L}_{\text{D}_0} + (1-\beta)\,\mathcal{L}_{\text{D}_1}, \tag{12}$$

where $\gamma_1, \gamma_2, \gamma_3, \beta$ are the weights of the corresponding losses and they are chosen to reflect equal importance.

## III. EXPERIMENTAL SETUP

We performed both simulated and real-world experiments. The PALGAN was trained on paired clean-recorded speech dataset. In simulated inference, restored signals were obtained by feeding the preprocessed signals into the Discriminator0. In the real-world experiment, the preprocessed signals were transmitted through the PAL, and the real restored speech was recorded in the near acoustic field.

### A. Dataset

We used a randomly assembled 67-hour clean speech from the VCTK dataset, resampled to 16 kHz, for training and validation, with clean speech serving as both input and target signals. To prepare the clean-recorded speech pairs for the process discriminator, we modulated the clean speech onto a 40 kHz ultrasonic carrier using DSB modulation and emitted it through a transducer. The resulting SD audible sound was recorded by a B&K type 4189 microphone placed 1.5 m away from the transducer in a full anechoic room. In addition to the VCTK test set, we also employed the read speech from the DNS Challenge dataset [33] for unseen testing. They together comprise a 2-hour test set.

### B. Configuration

We set the window size to 20 ms, hop size to 10 ms and FFT number to 320. Our configurations were primarily in line with those described in [21], except for the settings listed in Table I. The kernel sizes and strides used in the frequency band-split module and the time band-split module are separated by a slash (/). The multiply-accumulate operations (MACs) required for processing a 4-second mixture, reported in giga-operations per second (G/s), and the number of trainable parameters in million (M) are also presented. During the training phase, each utterance was segmented into non-overlapping 4-second segments, while the entire utterance was used during inference. We trained the network for 50 epochs using the Adam optimizer [34] with an initial learning rate of 0.0002.

We introduced two baseline teqniques: DSB and recursive EQ. The forward model used in recursive EQ was the identified DiffVF model based on the recorded dataset. The mudulation index $m$ was set to 0.8. The scaling parameter $c_0$ in recursive EQ was set to 0.065. The EQ order was roughly determined by the point at which the spectral MSE stop decreasing.

## IV. RESULT AND DISCUSSION

To conduct meaningful comparisons, we employed three non-intrusive and four intrusive objective metrics: non-intrusive speech quality and naturalness assessment (NISQA) [35], where quality dimensions of overall quality and coloration were included, deep noise suppression mean opinion score (DNSMOS) [36], where speech quality (SIG) and overall quality (OVRL) were used, UTokyo-SaruLab MOS (UTMOS) [37], narrow-band perceptual evaluation of speech quality (NB-PESQ) [38], extended short-time objective intelligibility (eSTOI) [39], signal-to-distortion ratio (SDR) [40], and Mel cepstral distortion (MCD) [41]. Higher values in these metrics indicate superior performance except for MCD.

TABLE II
QUANTITATIVE COMPARISONS WITH BASELINES ON THE MIXED TEST SET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND THE SECOND BEST RESULTS ARE UNDERLINED.

| Condition | Method | EQ order | NISQA | | DNSMOS | | UTMOS↑ | NB-PESQ ↑ | eSTOI (%) ↑ | SDR (dB) ↑ | MCD↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MOS↑ | COL↑ | OVRL↑ | SIG↑ | | | | | |
| Simulated | DSB [3] | – | 3.93 | 3.91 | **3.39** | **3.63** | 3.90 | **4.38** | **99.06** | 7.88 | 12.67 |
| | Recursive EQ [4] | 20 | 3.42 | 3.29 | 3.39 | 3.62 | 3.08 | 2.83 | 93.25 | 4.93 | 11.85 |
| | Recursive EQ [4] | 40 | 2.99 | 2.90 | 3.29 | 3.55 | 2.73 | 2.38 | 90.23 | 3.33 | 11.50 |
| | PALGAN | – | **4.42** | **4.14** | 3.35 | 3.58 | **4.05** | 4.30 | 93.98 | **25.43** | **5.19** |
| Real-world | DSB [3] | – | 1.94 | 2.04 | 2.80 | 3.11 | 2.25 | 2.11 | 57.39 | -4.76 | 15.05 |
| | Recursive EQ [4] | 20 | 1.56 | 1.90 | 2.96 | 3.25 | 1.47 | 1.72 | 58.76 | 5.74 | 17.46 |
| | Recursive EQ [4] | 40 | 1.48 | 1.96 | 2.78 | 3.05 | 1.42 | 1.69 | 54.69 | 4.89 | 16.47 |
| | PALGAN | – | **4.62** | **4.25** | **3.46** | **3.65** | **3.98** | **3.69** | **83.16** | **17.55** | **6.88** |

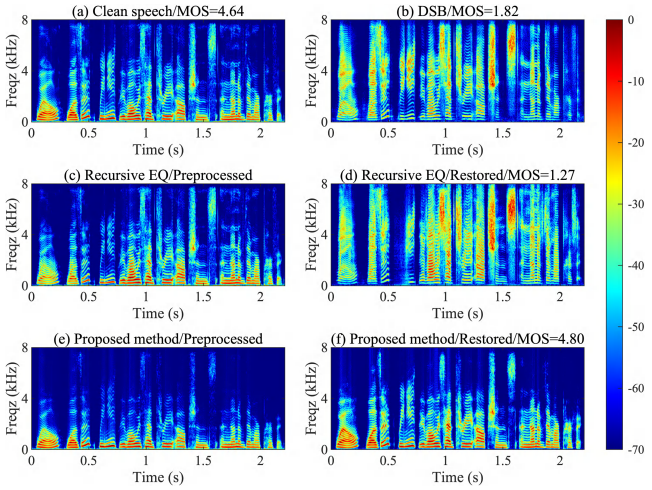



Fig. 4. Spectrograms of preprocessed (left) and restored (right) speech examples in real-world condition, with relevant NISQA-MOS indicated.

Table II presents the quantitative results of the proposed method compared with two baselines under simulated and real-world experimental conditions. It can be seen that the baseline methods generally achieve higher objective scores under simulated conditions, whereas the results of the proposed method remain more consistent across different conditions. This consistency reinforces the accuracy of the jointly learned Discriminator0, and facilitates the efficient development of preprocessing algorithms without requiring a pre-trained forward model.

To investigate the influence of EQ order on performance metrics, several orders were tested. Despite improvements in eSTOI and SDR under real test, recursive EQ yields lower scores than DSB across perceptual metrics, including NISQA, DNSMOS, UTMOS, and NB-PESQ, as its point-wise error compensation provides little perceptual benefit. In contrast, the proposed method operates in the T-F domain, enabling more refined control over spectral restoration.

Moreover, the DNSMOS, UTMOS, NB-PESQ, eSTOI, and SDR scores in the refined EQ method do not increase monotonically with order. This fluctuation implies its instability across different speeches, making the method sensitive to order selection. Moreover, higher order compromises the feasibility of real-time deployment. The proposed method overcomes these issues and demonstrates notable performance improvements across all metrics under both simulated and real-world conditions. We visualized the spectrograms of preprocessed and restored speech in the real-world condition, as shown in Fig. 4. Our method produces a more sophisticated preprocessed signal, resulting in restoration that closely approximates clean speech.

## V. CONCLUSIONS

In this paper, we proposed a joint optimization strategy implemented by a PALGAN, which leveraged adversarial training to simultaneously obtain the preprocessing network and a forward model that is more consistent with the real PAL system. A process discrimination mechanism was incorporated into the discriminator, providing a high-level abstraction of realness to offer auxiliary supervision to the preprocessing network. Quantitative results under both simulated and real-world experiments demonstrated that the proposed method achieved competitive performance when compared with state-of-the-art methods.

## REFERENCES


[1] Z. Kuang, J. Mao, and Y. Hu, "Efficient monaural directional sounding ultrasonic screen and manufacturing process thereof," Patent CN115 348 514A, 2022.

[2] K. Aoki, T. Kamakura, and Y. Kumamoto, "Parametric loudspeaker—characteristics of acoustic field and suitable modulation of carrier ultrasound," *Electronics and Communications in Japan*, vol. 74, pp. 76–82, 1991.

[3] M. Yoneyama, J.-i. Fujimoto, Y. Kawamo, and S. Sasabe, "The audio spotlight: An application of nonlinear interaction of sound waves to a new type of loudspeaker design," *The Journal of the Acoustical Society of America*, vol. 73, pp. 1532–1536, 1983.

[4] W. Ma, Y. Zhu, F. Hao, L. Qin, F. Fan, and J. Yang, "Deepprenet: A deep learning pre-processing method for speech distortion correction in parametric array loudspeaker," in *Proc. ICASSP*. IEEE, 2025, pp. 1–5.

[5] T. Kamakura, "Development of parametric loudspeaker for practical use," in *Proc. 10th Int. Symp. Nonlinear Acoustics*, 1984, pp. 147–150.

[6] T. D. Kite, J. T. Post, and M. F. Hamilton, "Parametric array in air: Distortion reduction by preprocessing," *The Journal of the Acoustical Society of America*, vol. 103, p. 2871, 1998.

[7] C. Shi and Y. Kajikawa, "Effect of the ultrasonic emitter on the distortion performance of the parametric array loudspeaker," *Applied Acoustics*, vol. 112, pp. 108–115, 2016.

[8] E.-L. Tan, P. Ji, and W.-S. Gan, "On preprocessing techniques for bandlimited parametric loudspeakers," *Applied Acoustics*, vol. 71, pp. 486–492, 2010.

[9] K. C.-M. Lee and W.-S. Gan, "Bandwidth-efficient recursive pth-order equalization for correcting baseband distortion in parametric loudspeakers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 706–710, 2006.

[10] J. J. Croft, M. E. Spencer, and J. O. Norris, "Modulator processing for a parametric speaker system," Patent WO2 001 015 491A1, 2001.

[11] E. R. Geddes and L. W. Lee, "Auditory perception of nonlinear distortion-theory," in *Audio Engineering Society Convention 115*. Audio Engineering Society, 2003.

[12] J. Zhong, R. Kirby, and X. Qiu, "The near field, westervelt far field, and inverse-law far field of the audio sound generated by parametric array loudspeakers," *The Journal of the Acoustical Society of America*, vol. 149, pp. 1524–1535, 2021.

[13] W. Ji and W.-S. Gan, "Identification of a parametric loudspeaker system using an adaptive volterra filter," *Applied Acoustics*, vol. 73, pp. 1251–1262, 2012.

[14] C. Shi and Y. Kajikawa, "Ultrasound-to-ultrasound volterra filter identification of the parametric array loudspeaker," in *Proc. DSP*. IEEE, 2015, pp. 1–4.

[15] M. Schetzen, "Theory of pth-order inverses of nonlinear systems," *IEEE Transactions on Circuits and Systems*, vol. 23, no. 5, pp. 285–291, 1976.

[16] Y. Mu, P. Ji, W. Ji, M. Wu, and J. Yang, "Modeling and compensation for the distortion of parametric loudspeakers using a one-dimension volterra filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 2169–2181, 2014.

[17] W. Ma, Y. Zhu, P. Ji, Z. Kuang, M. Wu, and J. Yang, "Differential volterra filter: A two-stage decoupling method for audible sounds generated by parametric array loudspeakers based on westervelt equation," *The Journal of the Acoustical Society of America*, vol. 157, no. 2, pp. 1057–1071, 2025.

[18] S. Nercessian, A. Sarroff, and K. J. Werner, "Lightweight and interpretable neural modeling of an audio distortion effect using hyperconditioned differentiable biquads," in *Proc. ICASSP*. IEEE, 2021, pp. 890–894.

[19] H. Zhang, K. Tan, and D. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions." in *Proc. Interspeech*, 2019, pp. 4255–4259.

[20] M. Li, T. Zhuang, K. Chen, J.-X. Zhong, and J. Lu, "Deep learning-based approach for identification and compensation of nonlinear distortions in parametric array loudspeakers," *IEEE Signal Processing Letters*, vol. 32, pp. 1455–1459, 2025.

[21] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[22] F. Hao, A. Li, X. Li, and C. Zheng, "Dsinet: Towards real-time target speaker extraction with dynamic speaker information fusion," in *Proc. ICASSP*, 2025, pp. 1–5.

[23] Y. Liu, B. Thoshkahna, A. Milani, and T. Kristjansson, "Voice and accompaniment separation in music using self-attention convolutional neural network," *arXiv preprint arXiv:2003.08954*, 2020.

[24] D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khudanpur, "A time-restricted self-attention layer for asr," in *Proc. ICASSP*. IEEE, 2018, pp. 5874–5878.

[25] 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.

[26] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[27] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*. IEEE, 2020, pp. 6199–6203.

[28] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[29] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2021, pp. 72–76.

[30] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.

[31] S. Abdulatif, K. Armanious, J. T. Sajeev, K. Guirguis, and B. Yang, "Investigating cross-domain losses for speech enhancement," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 411–415.

[32] 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.

[33] 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," *arXiv preprint arXiv:2101.01902*, 2021.

[34] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv:2104.09494*, 2021.

[36] 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 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 886–890.

[37] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.

[38] 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. ICASSP*, vol. 2. IEEE, 2001, pp. 749–752.

[39] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[40] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[41] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1. IEEE, 1993, pp. 125–128.