



A Novel Deep Learning Framework for Efficient Multichannel Acoustic Feedback Control

Yuan-Kuei Wu^{1,3*}, Juan Azcarreta², Kashyap Patel³, Buye Xu³, Jung-Suk Lee³, Sanha Lee², Ashutosh Pandey³

¹Graduate Institute of Communication Engineering, National Taiwan University, Taiwan

²Reality Labs Research, Meta, UK

³Reality Labs Research, Meta, USA

ywk991112@gmail.com, {jazcarreta, apandey620}@meta.com

Abstract

This study presents a deep-learning framework for controlling multichannel acoustic feedback in audio devices. Traditional digital signal processing methods struggle with convergence when dealing with highly correlated noise such as feedback. We introduce a Convolutional Recurrent Network that efficiently combines spatial and temporal processing, significantly enhancing speech enhancement capabilities with lower computational demands. Our approach utilizes three training methods: In-a-Loop Training, Teacher Forcing, and a Hybrid strategy with a Multichannel Wiener Filter, optimizing performance in complex acoustic environments. This scalable framework offers a robust solution for real-world applications, making significant advances in Acoustic Feedback Control technology.

Index Terms: acoustic feedback, howling suppression, speech enhancement

1. Introduction

Acoustic feedback[1, 2] is a prevalent challenge in audio devices equipped with multiple microphones and loudspeakers, such as hearing aids and public address systems. This phenomenon occurs when the output sound from the loudspeakers is re-captured by the microphones, leading to various disruptive effects such as howling, screaming, and whistling. This issue not only reduces the audio quality but also limits the maximum achievable amplification, thereby affecting the usability of hearing aids, particularly in environments that require high gain settings.

Traditionally, acoustic feedback control (AFC) has been addressed through various digital signal processing (DSP) techniques aimed at identifying and mitigating the feedback path dynamically. One of the most established methods is the use of adaptive filters, such as the Normalized Least Mean Square (NLMS) algorithm[3], which adjusts its coefficients to minimize the error between the predicted and actual feedback. This method, while robust and simple to implement, often suffers from slow convergence rates and can be biased due to the high correlation between the input and feedback signals. Other classical approaches to reduce feedback include delay insertion[2, 4, 5, 6, 7], frequency shifting[8, 9], and phase modulation[10] to decorrelate signals. Recent advancements like Partitioned Block NLMS[11] and hybrid adaptive filters[12] enhance convergence and stability. Moreover, the Recursive Least Squares (RLS) algorithm[2] offers improved convergence rates for acoustic feedback control, enhancing overall performance in variable acoustic settings.

*Work done during internship at Meta Reality Labs Research.

In recent years, deep learning techniques have begun to make inroads into the domain of AFC, offering promising results particularly in complex acoustic scenarios where traditional methods struggle. Studies[13, 14, 15, 16] have explored the use of recursive neural network training and hybrid models that combine deep learning with Kalman filters to enhance feedback suppression capabilities. These approaches leverage the ability of deep neural networks to model complex nonlinear relationships and handle variations in the feedback path effectively. However, a significant limitation of these studies is their focus on single-channel settings, which do not align with the multi-channel configurations commonly used in practical audio devices. Moreover, the application of such models is often constrained by their computational demands, making them less suitable for edge devices with limited processing power.

This study proposes a novel deep-learning-based framework to control multichannel acoustic feedback in audio devices, overcoming the limitations of single-channel models and computational inefficiencies prevalent in existing deep learning solutions. We leverage three innovative training methods, In-a-Loop Training, Teacher Forcing, and a Hybrid strategy incorporating a Multichannel Wiener Filter, to optimize our model for complex multichannel environments. Central to our approach is the use of a Convolutional Recurrent Network (CRN)[17], a model that uniquely combines spatial and temporal processing to address multichannel speech enhancement challenges. The CRN model is specifically designed for resource efficiency, characterized by low latency, lightweight architecture, and minimal computational demands, making it ideal for deployment on edge devices such as smart glasses. Here, the minimal delay in processing and the high correlation between feedback and input sources in such devices emphasize the need for robust multichannel processing to effectively mitigate feedback while maintaining audio quality. This model utilizes trainable filters for spatial processing and Long Short-Term Memory (LSTM) networks for temporal dynamics, achieving significant performance improvements over robust baselines with fewer parameters and reduced computational load. Our contributions provide a scalable and effective multi-channel AFC system that is uniquely adapted for real-world applications, offering a substantial advancement in the management of acoustic feedback in audio devices.

2. Methods

We propose a deep-learning-based framework for controlling multichannel acoustic feedback in audio devices with multiple microphones and loudspeakers. Although these devices can take many forms, this section focuses on the major elements of acoustic feedback and on three training strategies for our model.

2.1. Acoustic Feedback System

In a device with multiple loudspeakers (indexed by j) and multiple microphones (indexed by i), acoustic feedback arises when loudspeaker signals re-enter the microphones. Let $s(t)$ be the desired external source (e.g., a user's speech). Each microphone $m_i(t)$ contains the desired source, and the signals returning from all loudspeakers via their respective feedback paths:

$$m_i(t) = s(t) + \sum_j (h_{ij} * y_j(t)), \quad (1)$$

where h_{ij} denotes the impulse response of the feedback path from loudspeaker j to microphone i . The convolution $*$ reflects acoustic propagation and enclosure-specific effects. Whenever the loop gain at certain frequencies exceeds unity, howling or whistling emerges.

2.1.1. Default System

In the simplest default system, each loudspeaker j outputs the same signal derived from a pre-selected reference microphone $m_{\text{ref}}(t)$. Let Δt be the total system delay, G the amplifier gain, and $\sigma(\cdot)$ the loudspeaker's nonlinear response. Then,

$$y_j^{(D)}(t) = \sigma(m_{\text{ref}}(t - \Delta t) \cdot G), \quad \forall j. \quad (2)$$

Because $m_{\text{ref}}(t)$ itself includes feedback from previous time steps, this setup can cause significant howling if not controlled.

2.1.2. Feedback-Controlled System (MISO Model)

To suppress feedback, we replace $m_{\text{ref}}(t)$ with a model-based estimate $\hat{s}(t)$ of the desired source. The model accepts all microphone signals as inputs (MISO: multichannel input, single-channel output) and produces one estimated signal:

$$y_j^{(F)}(t) = \sigma(\hat{s}(t - \Delta t) \cdot G), \quad \forall j. \quad (3)$$

Since $\hat{s}(t)$ ideally omits the loudspeaker component from past frames, the feedback loop is significantly weakened, preventing re-amplification of loudspeaker signals.

2.2. Training Methods

Below, we describe three strategies for training the proposed MISO model, as depicted in Fig. 1. Although each aims to ensure that $\hat{s}(t)$ approximates $s(t)$ while minimizing feedback artifacts, they differ in how they handle the feedback loop during data generation and training. Regardless of the strategy, we define a training loss that compares the estimated source $\hat{s}(t)$ with the true source $s(t)$. A common choice is the Signal-to-Noise Ratio (SNR) loss[18].

2.2.1. In-a-Loop Training

This method explicitly simulates the entire feedback process at every time step. The network output $\hat{s}(t)$ is looped back into the system to produce future loudspeaker signals $y_j^{(F)}(t)$, which then influence subsequent microphone inputs.

In this approach, the model generates $\hat{s}(t)$ at each time step using the current microphone signals. These microphone signals include feedback from the loudspeaker outputs at previous time steps. We then produce the loudspeaker outputs $y_j^{(F)}(t)$ by applying the delay, gain, and nonlinearity to $\hat{s}(t)$. These outputs feed back into the microphones for time $t + 1$. After the

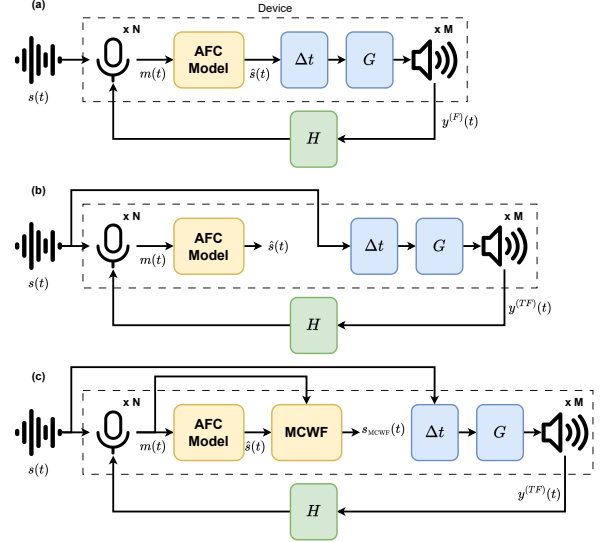


Figure 1: Block diagrams illustrating the three training approaches for acoustic feedback control: (a) In-a-Loop Training, (b) Teacher Forcing Training, and (c) Hybrid with Multichannel Wiener Filter (MCWF).

entire sequence is processed, we concatenate all $\hat{s}(t)$ estimates and compare them with $s(t)$ using the training loss. This loop simulation closely matches real-time behavior but is computationally expensive per sample.

2.2.2. Teacher Forcing Training

Teacher forcing assumes that the model completely suppresses any feedback in the loudspeaker signal. In other words, it presumes that the loudspeakers play only the pure desired source $s(t)$ (plus any deliberate processing such as delay, gain, and nonlinearity), and ignores the possibility that the model's output $\hat{s}(t)$ might still contain residual loudspeaker components that could loop back into the microphones.

Under this assumption, we take Equation (3) and replace $\hat{s}(t)$ with the ground-truth source $s(t)$. Thus, instead of relying on the model's estimate, the loudspeaker output is simplified to

$$y_j^{(TF)}(t) = \sigma(s(t - \Delta t) \cdot G). \quad (4)$$

Because $y_j^{(TF)}(t)$ no longer depends on $\hat{s}(t)$, the model's input and output are decoupled, and the microphone signals can be generated offline without iteratively simulating the model's feedback. Specifically, we form the microphone input $m_i(t)$ by

$$m_i^{(TF)}(t) = s(t) + \sum_j (h_{ij} * y_j^{(TF)}(t)). \quad (5)$$

During training, the network learns to map these offline-generated inputs $\{m_i(t)\}$ to the clean source $s(t)$, and the training loss is computed once $\hat{s}(t)$ is produced. This procedure significantly reduces computational overhead, since there is no need to simulate real-time feedback from $\hat{s}(t)$ at each step. However, it creates a potential train-test mismatch. In practice, $\hat{s}(t)$ may contain small errors that re-enter the system when the model is deployed, a scenario never observed during teacher-forced training. Consequently, while teacher forcing enables efficient offline data generation and can yield strong performance

on synthetic data, it may reduce robustness once the model faces real feedback conditions.

2.2.3. Hybrid with Multichannel Wiener Filter

This strategy augments our deep network with a classical Multichannel Wiener Filter (MCWF). The core idea is to leverage both the nonlinear modeling ability of deep learning and the spatial filtering capabilities of Wiener-based beamforming in a multichannel setting.

In a multichannel system with N microphones, let $\mathbf{m}(t)$ denote the stacked microphone signal:

$$\mathbf{m}(t) = \begin{bmatrix} m_1(t) \\ m_2(t) \\ \vdots \\ m_N(t) \end{bmatrix}. \quad (6)$$

In the short-time Fourier transform (STFT) domain, we can write $\mathbf{m}(k, \ell)$ for each frequency bin k and time frame ℓ . The goal of the MCWF is to estimate the desired source $s(k, \ell)$ by forming a linear combination of the microphone channels:

$$s_{\text{MCWF}}(k, \ell) = \mathbf{w}^H(k, \ell) \mathbf{m}(k, \ell), \quad (7)$$

where $\mathbf{w}(k, \ell) \in \mathbb{C}^N$ is the complex-valued Wiener filter for each frequency bin and time frame, and \cdot^H denotes the Hermitian operation.

The optimal Wiener filter $\mathbf{w}(k, \ell)$ is typically derived by minimizing the mean-squared error (MSE) between the filter output $s_{\text{MCWF}}(k, \ell)$ and the desired source $s(k, \ell)$. In the frequency domain, this involves correlation and cross-correlation matrices. Denoting the *autocorrelation matrix* of the microphone signals by

$$\Phi_{\mathbf{m}\mathbf{m}}(k, \ell) = \mathbb{E}[\mathbf{m}(k, \ell) \mathbf{m}^H(k, \ell)], \quad (8)$$

and the *cross-correlation vector* between the microphone signals and the desired source by

$$\Phi_{\mathbf{m}s}(k, \ell) = \mathbb{E}[\mathbf{m}(k, \ell) s^*(k, \ell)], \quad (9)$$

the Wiener solution in matrix form is:

$$\mathbf{w}(k, \ell) = \Phi_{\mathbf{m}\mathbf{m}}(k, \ell)^{-1} \Phi_{\mathbf{m}s}(k, \ell), \quad (10)$$

assuming $\Phi_{\mathbf{m}\mathbf{m}}(k, \ell)$ is invertible. Here \cdot^* denotes complex conjugation.

In our hybrid approach, the deep model first outputs a raw estimate $\hat{s}(t)$ of the desired source, using either in-a-loop or teacher forcing method. Once $\hat{s}(t)$ is available, we convert it to the STFT domain to obtain $\hat{s}(k, \ell)$, which serves as a reference for the MCWF. Concretely, the MCWF now aims to exploit the spatial information from the multichannel microphone signals $\mathbf{m}(k, \ell)$ and refine $\hat{s}(k, \ell)$. This procedure can be conceptualized as:

$$s_{\text{MCWF}}(k, \ell) = \mathbf{w}^H(k, \ell) \mathbf{m}(k, \ell), \quad (11)$$

$$\mathbf{w}(k, \ell) = \Phi_{\mathbf{m}\mathbf{m}}(k, \ell)^{-1} \Phi_{\mathbf{m}\hat{s}}(k, \ell), \quad (12)$$

where $\Phi_{\mathbf{m}\hat{s}}(k, \ell)$ is the cross-correlation between the microphones and $\hat{s}(k, \ell)$. Here, $\hat{s}(k, \ell)$ replaces $s(k, \ell)$ as the desired target in the filter design. While this may not perfectly match the ground-truth $s(k, \ell)$, it provides a guiding signal for the Wiener filter that steers the spatial beamformer toward suppressing undesired feedback paths.

3. Experimental Setup

In this study, we investigate the acoustic feedback problem in multichannel speech data using the Rayban Meta smart glasses microphone array. The input speech is reverberant, sourced from the DNS Challenge clean dataset[19], which provides distinct training, validation, and testing sets. To simulate realistic environments, Room Impulse Responses (RIRs) were generated for 4000 virtual rooms for the training set and 400 rooms for both the validation and testing sets. The microphone array was positioned in 10 different locations within these rooms. The simulation of RIRs utilized the image method with an order of six, considering room sizes ranging from a minimum of 3x3x2 meters to a maximum of 10x10x5 meters. Absorption coefficients varied between 0.1 and 0.4 to mimic diverse acoustic conditions. The source-array distance was varied from 0.5 meters to 2.5 meters. Early reverberation lengths were set at 25, 50, 75, and 100 milliseconds. Each dataset, training, validation, and testing, contains 72000, 3600, and 3600 samples, respectively, with each audio sample having a duration of 2 seconds. The multichannel signals were synthesized using these parameters to study their effects on acoustic feedback in a controlled yet realistic setting.

In the development of our acoustic feedback pipeline, we employ a Convolutional Recurrent Network (CRN) tailored for spatial and temporal processing, based on the frequency domain version of model presented in [17]. The CRN model incorporates spatial convolution layers to enhance spatial feature extraction, followed by LSTM layers for temporal dynamics analysis. The process involves channel expansion through spatial convolution, and the resulting features are then processed by LSTM layers for deeper temporal insights. The combined output involves element-wise multiplication of LSTM and spatial convolution outputs to generate the final result. The architecture consists of three CRN layers with channel configurations of 10/20, 20/20, and 20/1, respectively. The model comprises 684K parameters and operates at a processing speed of 82 million MACs per second.

Training settings for the model vary between fixed and variable configurations. Specifically, the amplifier gain is set between 40-75 dB for variable training, and fixed at 75 dB. Similarly, the time delay employed varies between 5-30 ms or is fixed at 8 ms. To emulate the physical constraints of the loudspeaker device, we use a nonlinearity setting of clipping between -1000 to 1000, which effectively mimics the loudspeaker's limitations without inducing howling. During inference, the settings are standardized with a fixed gain and delay. Unless otherwise specified, the gain and delay default to 40 dB and 8 ms, respectively.

As suitable datasets for acoustic feedback transfer functions are not publicly available, we collected our own data using the Rayban Meta smart glasses, an on-the-shelf product. We designed eight specific scenarios to simulate different user interactions with the glasses, including Normal Glassware, Phone Pickup (R/L), Button Press (R/L), Adjust Glasses (R&L), Cover Ears (R/L), and Grabbing Nose (R). In each scenario, we played back white noise through a loudspeaker and calculated the linear transfer function between the emitted noise and the signal captured by the microphones, capturing the nuances of each interaction scenario. It is critical to note that we assumed a time-invariant system for our experiments, meaning that the feedback transfer function was considered stable and unchanged throughout the time. In this context, 'R' refers to the right hand, and 'L' denotes the left hand.

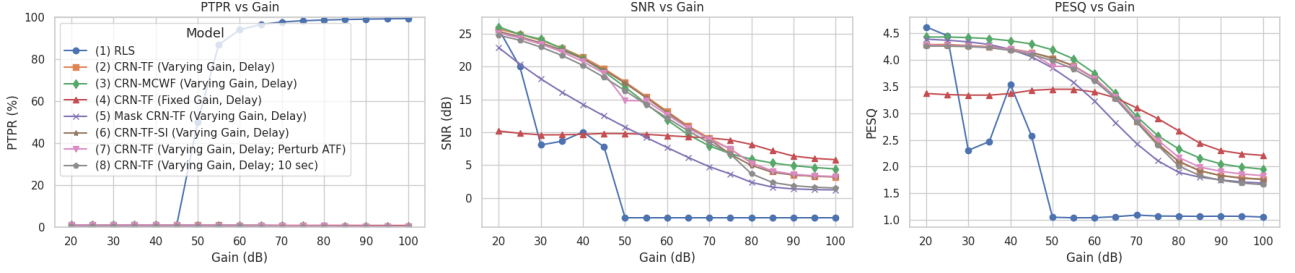


Figure 2: Comprehensive evaluation of acoustic feedback control models showing PTPR, SNR, and PESQ across varying gain levels. Demonstrate the models’ capabilities in managing howling, enhancing signal clarity, and maintaining speech quality under different operational conditions.

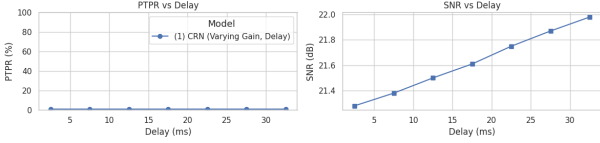


Figure 3: Performance evaluation of the CRN model under varying time delays.

In the evaluation of our acoustic feedback control system, effective suppression of howling and preservation of speech quality are paramount. We assess the speech quality of the model’s output signals using the Signal-to-Noise Ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ) [20], with higher values indicating better quality. To detect howling, we employ the Peak-to-Threshold Power Ratio (PTPR) [2], defined as $PTPR(\hat{\omega}_i, t)[dB] = 10 \log_{10} \left(\frac{|Y(\hat{\omega}_i, t)|^2}{P_0} \right)$. Here, $|Y(\hat{\omega}_i, t)|^2$ represents the power of the candidate howling component, and P_0 is a fixed absolute power threshold set such that $10 \log_{10} P_0 = 35$ dB. The incidence of howling is quantified by the percentage of frames where $PTPR > 0$ dB, with lower values indicating better suppression of howling and improved system stability.

4. Results

In our baseline model, model (1) uses the RLS algorithm, which acts as an adaptive filter and is designed to quickly converge and adapt to rapidly changing system parameters. As depicted in Fig. 2, the RLS algorithm underperforms, particularly in higher gain settings where its propensity to howl is evidenced by poor PTPR scores. In contrast, our proposed model (2), CRN-TF (where TF stands for teacher-forcing), maintains high speech quality with a PESQ score above 4 even at a 50 dB gain. Model (3), which incorporates a MCWF, exhibits superior speech quality, achieving higher PESQ scores. Both models (2) and (3) effectively suppress howling across all gain settings, including gains higher than those in the training set. Although we experimented with In-a-loop training to align training with on-the-fly inference simulation, its unstable training regimen resulted in suboptimal performance across all settings, not shown in the figure.

Subsequent findings demonstrate the performance of various ablation studies on howling suppression and speech quality evaluation using SNR and PESQ scores, as depicted in Fig. 2. When comparing models (2) and (4), where model (2) was

trained with gains varying between 40-75 dB and model (4) at a fixed gain of 75 dB, it is evident that the varying gain approach outperforms the fixed gain at lower gains (below 75 dB). However, at higher gains (equal to or above 75 dB), the fixed gain approach yields better performance. This suggests that training with a varying gain enhances model generalization within a broader gain range but is less effective at the maximum gain setting. For models (2) and (5), where model (2) generates the spectrogram and model (5) produces a mask, the spectrogram-based approach shows superior performance in this task. Model (6), implementing system identification (CRN-TF-SI) by using the loudspeaker output as a reference signal, did not outperform the multichannel microphone input method, indicating the challenge of modeling the relationship between feedback signals and loudspeaker output directly. In comparing models (2) and (7), where model (7) included perturbations in the feedback transfer functions by adding Gaussian noise during training, we found that despite introducing noise during training, which created misalignment in the transfer functions between training and testing, model (7) still maintains robust performance in SNR values. This resilience highlights the model’s effectiveness in managing SNR despite discrepancies in the acoustic conditions. Finally, comparing models (2) and (8), which tested the effect of extending the inference signal from 2 seconds to 10 seconds, revealed no occurrence of howling and only minor degradation in speech quality. This finding suggests that even with extended inference, the model robustly suppresses howling without significantly compromising speech quality.

Figure 3 illustrates the performance of our model across various delay settings from 2.5 ms to 32.5 ms. Remarkably, the model effectively suppresses howling at all tested delays, demonstrating robustness beyond the trained delay range. Additionally, we observe an improvement in SNR as the delay increases, suggesting that the model performs more effectively when the feedback signal has less correlation with the input source signal. This trend indicates an enhanced ability of the model to handle increased delays by effectively distinguishing between the source and feedback signals.

5. Conclusion

This study introduced a pioneering deep learning-based framework for controlling multichannel acoustic feedback in audio devices. The framework successfully suppresses howling and maintains high speech quality under various conditions. Notably, our CRN efficiently combines spatial and temporal processing, making it ideal for edge devices with limited capabilities and offering a robust solution for real-world AFC.

6. References

- [1] R. V. Waterhouse, "Theory of howlback in reverberant rooms," *The Journal of the Acoustical Society of America*, vol. 37, no. 5, pp. 921–923, 1965.
- [2] T. Van Waterschoot and M. Moonen, "Fifty years of acoustic feedback control: State of the art and future challenges," *Proceedings of the IEEE*, vol. 99, no. 2, pp. 288–327, 2010.
- [3] S. C. Douglas, "A family of normalized lms algorithms," *IEEE signal processing letters*, vol. 1, no. 3, pp. 49–51, 1994.
- [4] M. G. Siqueira and A. Alwan, "Steady-state analysis of continuous adaptation in acoustic feedback reduction systems for hearing-aids," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 443–453, 2000.
- [5] A. Spriet, S. Doclo, M. Moonen, and J. Wouters, "Feedback control in hearing aids," *Springer Handbook of Speech Processing*, pp. 979–1000, 2008.
- [6] J. Hellgren and F. Urban, "Bias of feedback cancellation algorithms in hearing aids based on direct closed loop identification," *IEEE transactions on speech and audio processing*, vol. 9, no. 8, pp. 906–913, 2001.
- [7] S. Laugesen, K. Hansen, and J. Hellgren, "Acceptable delays in hearing aids and implications for feedback cancellation," *The Journal of the Acoustical Society of America*, vol. 105, no. 2-Supplement, pp. 1211–1212, 1999.
- [8] M. R. Schroeder, "Improvement of acoustic-feedback stability by frequency shifting," *The Journal of the Acoustical Society of America*, vol. 36, no. 9, pp. 1718–1724, 1964.
- [9] F. Strasser and H. Puder, "Adaptive feedback cancellation for realistic hearing aid applications," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2322–2333, 2015.
- [10] M. Guo, S. H. Jensen, J. Jensen, and S. L. Grant, "On the use of a phase modulation method for decorrelation in acoustic feedback cancellation," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 2000–2004.
- [11] Vasundhara, B. K. Mohanty, G. Panda, and N. B. Puhan, "Hardware design for vlsi implementation of acoustic feedback canceller in hearing aids," *Circuits, Systems, and Signal Processing*, vol. 37, no. 4, pp. 1383–1406, 2018.
- [12] S. Nordholm, H. Schepker, L. T. Tran, and S. Doclo, "Stability-controlled hybrid adaptive feedback cancellation scheme for hearing aids," *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. 150–166, 2018.
- [13] H. Zhang, M. Yu, and D. Yu, "Deep ahs: A deep learning approach to acoustic howling suppression," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [14] H. Zhang, M. Yu, Y. Wu, T. Yu, and D. Yu, "Hybrid ahs: A hybrid of kalman filter and deep learning for acoustic howling suppression," *INTERSPEECH*, 2023.
- [15] H. Zhang, Y. Zhang, M. Yu, and D. Yu, "Advancing acoustic howling suppression through recursive training of neural networks," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 711–715.
- [16] —, "Enhanced acoustic howling suppression via hybrid kalman filter and deep learning models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2828–2840, 2024.
- [17] A. Pandey and B. Xu, "Decoupled spatial and temporal processing for resource efficient multichannel speech enhancement," *arXiv preprint arXiv: 2401.07879*, 2024.
- [18] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.
- [19] C. K. A. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Inter-speech 2021 deep noise suppression challenge," *arXiv preprint arXiv: 2101.01902*, 2021.
- [20] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.