Diff-DEQ: Differentiable Dynamic Equalization for Studio-Quality Speech Processing

Parakrant Sarkar, Permagnus Lindborg
SoundLab, *School of Creative Media*, *City University of Hong Kong*, Hong Kong SAR, China
parakrant.sarkar@my.cityu.edu.hk, pm.lindborg@cityu.edu.hk

***Abstract*—We present Differentiable Dynamic Equalization (Diff-DEQ), a fully differentiable deep learning framework for speech equalization and enhancement to achieve studio quality for audio post-production tasks. Unlike fixed-rule equalization methods, it adapts spectral components dynamically, responding to input signal variations to attain precise and content-aware spectral shaping. The model combines a FiLM-modulated Temporal Convolutional Network (TCN) and a Bidirectional Gated Recurrent Unit (BiGRU) to predict per-band equalization parameters with audio feature-based conditioning for improved adaptability. We have trained the model in a self-supervised manner that eliminates the need for paired input-target data. We evaluate its performance using objective metrics on Diff-DEQ and parametric equalization (PEQ) across LibriTTS, DAPS, and VCTK datasets and non-intrusive speech quality assessment for subjective evaluation. Our results show that Diff-DEQ enhances speech intelligibility and perceived quality, making it well-suited for audio post-production.**

***Index Terms*—equalization, dynamic equalization, differential digital signal processing, audio production**

I. Introduction

Speech equalization (EQ) is a vital tool in audio post-production, used to shape the tonal balance of speech for improved clarity and intelligibility [1]. It plays a key role in producing studio-quality audio for broadcast, communication systems, and assistive listening devices. With the rise of podcasting [2], online content creation, and remote communication, there is growing demand for automated EQ tools that enhance speech without requiring expert knowledge creating a low barrier of entry for novice users. Traditional approaches like parametric EQ [3] and graphic EQ [1] rely on fixed gain settings across frequency bands and often require manual tuning. However, the highly variable nature of speech across speakers, recording environments, and noise conditions limits the effectiveness of such static methods.

To address these challenges, we introduce Differentiable Dynamic Equalization (Diff-DEQ), a fully differentiable end-to-end deep learning framework that dynamically adapts spectral components in response to input signal characteristics. Diff-DEQ is designed for adaptive speech equalization, enhancement, and audio mastering tasks, ensuring optimal tonal balance in professional audio production and making high-quality speech accessible to a broader range of users, especially novice or non-expert users.

A. Comparison of Parametric and Dynamic Equalization

We conducted a small experiment comparing rule-based parametric equalization (PEQ) and dynamic equalization (DEQ) on a speech sample of approximately three seconds with sampling rate of *24kHz*. The PEQ implementation was used from the `dasp-pytorch toolkit`[1] , while the DEQ implementation was developed as part of this work. A detailed explanation of our DEQ framework and its methodology is provided in Section II. The Fig. 1 compares the frequency response of rule-based DynamicEQ (DEQ) and ParametricEQ (PEQ) applied to a speech signal. Both of them use the same EQ parameters. The DEQ (red line) closely follows the target response (black dashed line), particularly in the mid and high-frequency regions, whereas the PEQ (blue dashed line) deviates significantly. In the *1 kHz – 10 kHz* range, PEQ struggles to match the target response, leading to spectral inconsistencies. DEQ provides a smoother transition with fewer abrupt gain changes, especially around *3 kHz – 6 kHz*, a crucial region for speech intelligibility. Additionally, DEQ maintains a slight boost above *10 kHz*, improving speech clarity and preserving high-frequency details that PEQ fails to capture. These results demonstrate that adaptive, signal-aware equalization in DEQ is more effective than the static adjustments of PEQ, allowing for a more natural and balanced spectral response. This motivated us to develop a Diff-DEQ a fully differentiable deep learning framework for automatic equalization tasks.

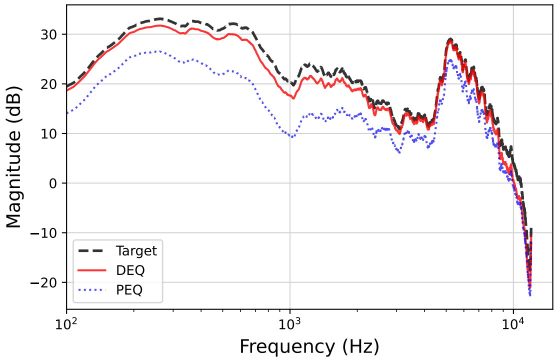

Fig. 1. Frequency Response of comparison of rule-based DynamicEQ and ParametricEQ

[1] `https://github.com/csteinmetz1/dasp-pytorch/`

B. Related Work

Automatic equalization has been a key area of research, with a strong focus on EQ matching, where equalizer parameters are automatically adjusted to match the spectral characteristics of a reference signal. Traditional EQ matching approaches [4] have been widely used in intelligent audio production tasks, either on isolated stems (e.g., vocals, instruments) or on full multi-track mixtures [1], [5]. Early deep learning-based approaches explored end-to-end equalization models. For example, [6] proposed a CNN-based model that approximates target equalization without explicitly estimating filter transfer functions. Later, differentiable signal processing techniques enabled neural parametric equalization (PEQ), where IIR biquadratic (biquad) filters are integrated into deep learning architectures [7]–[9]. These works leveraged the Differentiable Digital Signal Processing (DDSP) framework [10], allowing direct optimization of spectral differences between predicted and target frequency responses. Most of these methods used perceptually meaningful loss functions to improve EQ parameter prediction accuracy.

Recent research has explored neural dynamic range compression (DRC) that adaptively controls loudness and dynamics using neural models. [11] introduced an efficient deep learning model for loudness control, demonstrating the effectiveness of neural-based dynamic range processing. Additionally, [9] proposed *DeepAFX-ST*, which integrates PEQ and DRC in a differentiable signal chain. In NDMP [12], authors presented a six-band neural-driven multi-band processor that applies cascaded parametric equalization followed by band-wise dynamic range compression in a fully differentiable pipeline. This approach supports real-time and offline audio mastering, but the EQ and compression stages remain sequential rather than jointly adaptive.

Despite these advancements, a unified framework for differentiable dynamic equalization (DEQ) remains largely unexplored. Most prior works treat equalization and dynamic range compression as separate processes rather than joint operations. In this work, we propose Diff-DEQ, which integrates per-band parametric equalization with dynamic control, enabling adaptive spectral shaping that adjusts both frequency response and dynamic range in a fully differentiable framework.

II. Methodology

This section describes the design of Diff-DEQ, a fully differentiable dynamic equalization system that integrates parametric equalization, trainable crossover filtering, and differentiable dynamic range control.

A. Dynamic Equalization

Diff-DEQ performs adaptive dynamic equalization by processing the input signal across multiple frequency bands. We adopt a six-band structure, similar to the PEQ framework in [9]. Here, each band undergoes independent equalization and dynamic range processing. The frequency bands are defined during training using a trainable crossover mechanism that helps the model learn optimal band divisions based on the input signal characteristics.

1) Linkwitz-Riley Crossover Filtering: The input signal is split into frequency bands by applying the designed Linkwitz-Riley filters in the time domain, which are implemented as cascaded second-order Butterworth filters by following [13]:

$$y_{\text{low},i}(n) = \text{LPF}_i[x(n)] \tag{1}$$

$$y_{\text{high},i}(n) = \text{HPF}_i[x(n)] \tag{2}$$

where $\text{LPF}_i[\cdot]$ and $\text{HPF}_i[\cdot]$ denote filtering with the i-th low-pass and high-pass Linkwitz-Riley filters, respectively, whose coefficients are derived from the following z-domain transfer functions:

$$H_{\text{LPF}}(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{3}$$

$$H_{\text{HPF}}(z) = \frac{b_0 - b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{4}$$

The filter coefficients b_0, b_1, b_2, a_1, a_2 are computed based on the desired crossover frequency and sample rate, as described earlier. All filtering is performed in the time domain using these coefficients.

B. Per-Band Equalization and Dynamic Range Processing

Each frequency band in the Diff-DEQ system is processed with its own set of equalization and dynamic range control parameters. Specifically, we apply parametric equalization (PEQ) [7] to adjust the gain in each band based on the input's spectral content, using biquad filters to enable precise and flexible frequency shaping. Alongside this, dynamic range control (DRC) [14] is applied within each band using a differentiable feed-forward compressor implemented via `torchcomp`[2].

We parameterize each band's EQ and DRC independently to allow localized control, and they are jointly predicted by the model to ensure coherent global behaviour. The compressor operates using root mean square (RMS) level estimation, adaptive gain computation, and smooth attack-release dynamics to avoid artefacts. This unified per-band approach enables Diff-DEQ to adaptively shape the spectral balance and control dynamic range across the frequency spectrum, while preserving phase continuity.

III. Model Architecture

In this section, we describe the architecture of the proposed Diff-DEQ model. Our model is inspired by [9] and consists of three main components: a FiLM-Modulated Temporal Convolutional Network (TCN), a Bidirectional Gated Recurrent Unit (BiGRU), and a Multi-Layer Perceptron (MLP), as shown in Figure 2. The model takes two inputs: the raw audio features x, and an auxiliary conditioning vector z that encodes global characteristics of the signal, such as loudness, MFCCs, and spectral centroid. The FiLM conditioning vector z is first passed through a small MLP to produce layer-wise scaling

[2] `https://github.com/DiffAPF/torchcomp`

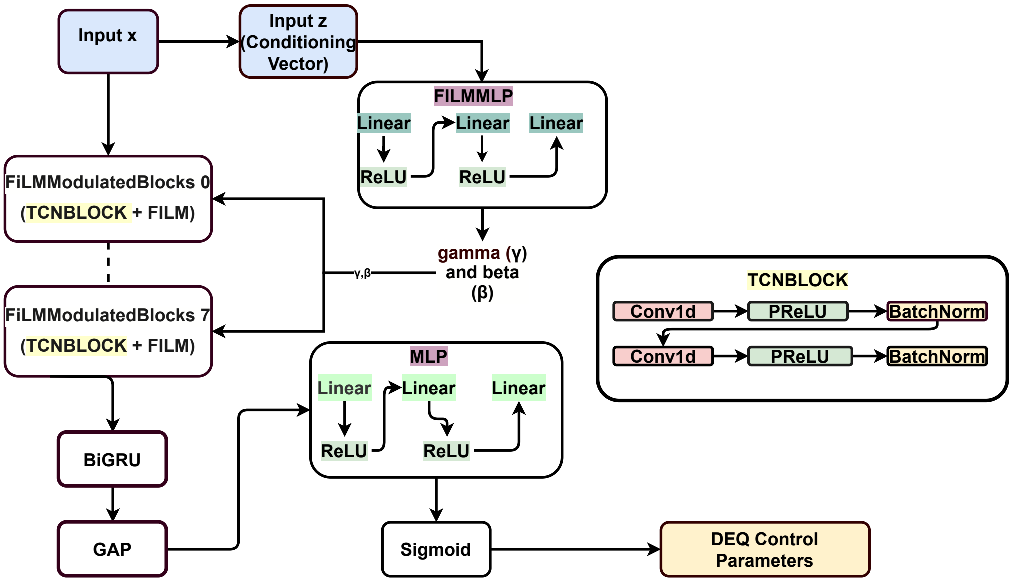

Fig. 2. Overview of the proposed Diff-DEQ model. The model takes in the audio signal x and conditioning vector z, and predicts dynamic equalization control parameters through a FiLM-modulated TCN, BiGRU, and MLP stack.

and bias parameters (γ, β). The TCN is composed of 8 FiLM-modulated convolutional blocks with exponentially increasing dilation factors $\{1, 3, 9, 27\}$, allowing it to capture both short and long-term spectral dependencies. Each block applies a 1D convolution (kernel size 5), followed by PReLU activation and batch normalisation. The FiLM mechanism applies adaptive modulation to each layer's activations, enabling content-aware feature extraction tailored to the conditioning vector z. The output of the final TCN block is passed through a BiGRU, which models temporal dependencies and ensures smooth parameter transitions across frames. The bidirectional structure supports context-aware estimation and accommodates variable-length inputs. A global average pooling (GAP) layer is used to summarize the temporal dynamics into a fixed-size vector. The MLP maps this aggregated representation to dynamic equalization parameters: gain, Q-factor, threshold, ratio, attack, release, and makeup gain. A sigmoid activation ensures these outputs stay within valid operational ranges.

Trainable Crossover Mechanism: Diff-DEQ employs trainable Linkwitz-Riley crossover filters to adaptively split the input audio into frequency bands using learned cutoff frequencies. Both the crossover points and per-band DEQ parameters are predicted jointly, enabling signal-dependent spectral partitioning and precise dynamic control for each band. This architectural combination was chosen for its ability to balance real-time performance, temporal consistency, and interpretability. TCNs provide a scalable and parallelizable structure for local-to-global feature modelling, FiLM enables task-aware conditioning, and BiGRU stabilizes time-dependent outputs.

IV. Implementation

In this section, we describe the training and evaluation of our proposed model. We train two separate models: one for *Dynamic Equalization (DEQ)* and another for *Parametric Equalization (PEQ)*. Both models follow the same training pipeline.

A. Dataset and Preprocessing

We train Diff-DEQ using the `train-clean-360` subset of LibriTTS [15], consisting of 360 hours of speech sampled at $f_s = 24$ kHz. We use a 90/5/5 split for training, validation, and testing. To assess generalization, we evaluate our models on additional datasets, including DAPS [16] and VCTK [17], which contain diverse speakers, recording environments, and microphone types.

Each input signal is paired with a reference signal, but instead of relying on predefined input-target pairs, we adopt a *self-supervised training strategy*. The input is generated by applying randomly sampled DEQ parameters to the reference signal, simulating real-world variations in automatic equalization. This approach allows the model to learn DEQ parameter estimation without requiring ground-truth labels. Additionally, we introduce random gain adjustments within a range of ± 24 dB to enhance robustness. The model is trained on fixed-length audio segments of 70972 samples ($\approx$ 2.95s at 24 kHz). Frames with silent or low-energy content are excluded, where silence is defined as an amplitude threshold of $1e-4$ and low-energy frames as having energy below 0.01. All input signals are peak-normalized to ensure uniform dynamic range across training.

B. Training Procedure

We use a hybrid loss function consisting of a *Multi-Resolution Short-Time Fourier Transform (STFT) loss* [18] and an *L1 loss*. The STFT loss captures spectral differences between the predicted and reference signals, ensuring spectral fidelity, while the L1 loss enforces waveform-level accuracy. The model is trained using the Adam optimizer with a learning rate of $1e{-}4$, which decays using a cosine annealing scheduler over 200 epochs. We use a batch size of 16, and training is performed on randomly sampled segments from the dataset. We used `RTX4090` GPU with `24GBVRAM` with intel `i9-14900KF` cpu in training.

C. Inference and Evaluation

In the inference, we apply the same preprocessing pipeline to generate the input signal as training. We evaluate both DEQ and PEQ models using objective and subjective metrics. Objective metrics include PESQ, STOI, LUFS, and Mel spectral distance measures etc. We provide the training code and audio samples (for both DEQ and PEQ) at GitHub for reproducibility.

V. Results and Analysis

In this section, we briefly discuss the objective metrics to evaluate our proposed DEQ with the PEQ model. Later, a non-intrusive subjective test is carried out on the DEQ model.

A. Objective Evaluation

We have followed the objective metrics used in [9]. Table I compares Diff-DEQ (deq) with Parametric EQ (peq) across the LIBRI, DAPS, and VCTK datasets using various objective metrics. PESQ and STOI scores indicate that PEQ achieves slightly higher intelligibility and perceptual quality across datasets. However, Diff-DEQ demonstrates a lower Mean Squared Error (MSE) and Mel-Spectral Distance (MSD) on most datasets, suggesting a better spectral match to the reference signal. Notably, Diff-DEQ achieves lower LUFS differences, meaning it preserves loudness characteristics more accurately than PEQ. In terms of phase distortion, PEQ generally outperforms Diff-DEQ, exhibiting lower values across datasets, indicating more phase coherence. However, Diff-DEQ excels in transient preservation, across all the datasets, where it retains sharper attacks compared to PEQ. Similarly, the crest factor difference suggests that Diff-DEQ provides more balanced dynamic range control. While PEQ achieves slightly better spectral centroid accuracy, Diff-DEQ maintains competitive performance while integrating adaptive spectral shaping and dynamic compression within a single differentiable framework. These results highlight the trade-offs between static PEQ and adaptive DEQ, showcasing Diff-DEQ's potential in content-aware equalization while preserving dynamic nuances.

B. Subjective Evaluation

To assess the perceptual quality of the enhanced speech, we conducted a non-intrusive speech quality assessment (NISQA v2.0) using the `TorchMetrics` framework [19]. We evaluated the predicted and target audio samples across the LibriTTS, DAPS, and VCTK datasets, computing speech quality metrics to compare Diff-DEQ's performance against ground truth recordings. The Fig. 3, Fig. 4, and Fig. 5 illustrate the distribution of quality scores for the predicted and target speech across all three datasets. The results indicate that Diff-DEQ consistently produces high-quality outputs, closely aligning with the ground truth reference signals.

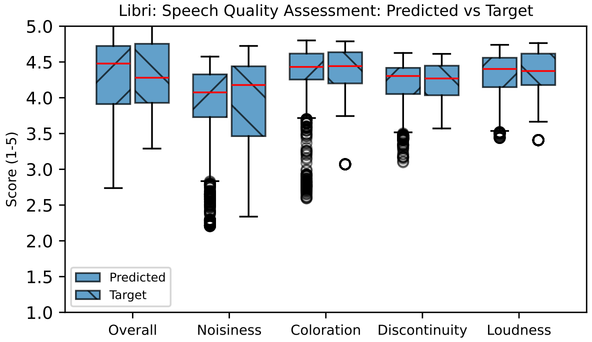

Fig. 3. Speech quality assessment results for the LibriTTS dataset.

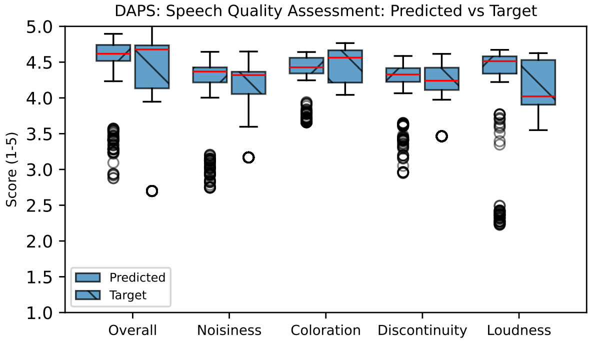

Fig. 4. Speech quality assessment results for the DAPS dataset.

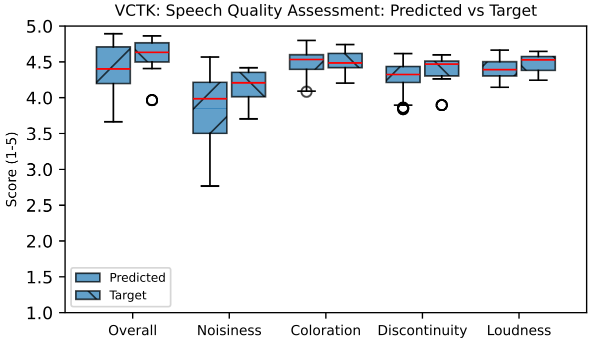

Fig. 5. Speech quality assessment results for the VCTK dataset.

These results indicate that across all datasets, Diff-DEQ achieves scores that are highly comparable to the target references, demonstrating its effectiveness in maintaining speech intelligibility, naturalness and minimal perceptual quality implying studio-quality speech. As shown in Table I and Figures 3–5, Diff-DEQ consistently achieves these benchmarks

TABLE I

Comparison of DEQ and PEQ across LIBRI, DAPS, and VCTK datasets. *PESQ*, *STOI*, and *Transient Preservation* favor higher values, while lower values are preferred for other metrics. **Boldface** highlights the best-performing model for each metric and dataset..

	LIBRI		DAPS		VCTK	
	deq	peq	deq	peq	deq	peq
PESQ	4.2897	**4.4237**	4.2898	**4.5269**	4.3294	**4.5180**
STOI	0.9877	**0.9988**	0.9851	**0.9979**	0.9730	**0.9983**
MSE	0.0672	**0.0414**	0.1877	**0.1222**	0.0737	**0.0348**
MSD	**1.1790**	1.3244	**2.1177**	2.2746	**1.0258**	1.0666
RMS Error	0.0340	**0.0337**	0.1328	**0.1150**	**0.0188**	0.0231
LUFS diff	**-5.8266**	-6.6102	**-11.3706**	-13.2795	**-0.0243**	-3.2614
STFT Loss	**1.2352**	1.5126	**2.1006**	2.1257	1.0817	**1.0718**
Spectral Centroid Error	**339.30**	370.22	238.07	**121.12**	423.08	**242.86**
Spectral Bandwidth Diff.	**0.0104**	0.0158	0.0168	**0.0117**	**0.0095**	0.0100
Spectral Flatness Diff.	**0.0362**	0.0395	0.0255	**0.0207**	0.0459	**0.0238**
Phase Distortion	0.4715	**0.1161**	0.4930	**0.0883**	0.4793	**0.1029**
Transient Preservation	**5.8366**	2.8849	**4.5725**	2.2156	**11.0627**	1.1047
Crest Factor Diff.	**0.2306**	0.2489	**0.6851**	0.7792	**0.1661**	0.2276

across multiple datasets. These findings demonstrate the potential of Diff-DEQ to deliver studio-quality, professional-grade speech enhancement for automated audio post-production.

VI. Conclusion

In this work, we introduced Diff-DEQ, a fully differentiable deep learning framework for adaptive dynamic equalization. It integrates adaptive spectral processing and dynamic range control into a unified model through a FiLM-modulated TCN, BiGRU-based temporal modelling, and an MLP-driven parameter estimation that helps the model to predict content-aware equalization settings. The experimental results across LIBRI, DAPS, and VCTK datasets demonstrate that Diff-DEQ effectively preserves transient details, maintains loudness consistency, and achieves better spectral matching, while offering a fully differentiable and trainable approach to speech enhancement achieving studio-quality for post-production workflows. In future work, we will explore real-time deployment and finetuning with perceptual loss functions further to enhance its effectiveness and use this in a signal chain with other audio effects. Also, we need to carry out subjective listening tests with real users to get the feedback on Diff-DEQ model.

References

[1] V. Välimäki and J. D. Reiss, "All About AudioEqualization: Solutions and Frontiers," *Applied Sciences*, vol. 6, no. 5, p. 129, May 2016, number: 5 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-3417/6/5/129

[2] N. Schmidt, J. Pons, and M. Miron, "PodcastMix: A dataset for separating music and speech in podcasts," Jul. 2022, arXiv:2207.07403 [cs]. [Online]. Available: http://arxiv.org/abs/2207.07403

[3] H. Behrends, A. von dem Knesebeck, W. Bradinal, P. Neumann, and U. Zölzer, "Automatic Equalization Using Parametric IIR Filters," *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 59, pp. 102–109, Mar. 2011.

[4] F. G. Germain, G. J. Mysore, and T. Fujioka, "Equalization matching of speech recordings in real-world environments," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 609–613.

[5] B. De Man, J. Reiss, and R. Stables, "Ten Years of Automatic Mixing," Salford, UK, Sep. 2017.

[6] M. A. M. Ramírez and J. D. Reiss, "End-to-end equalization with convolutional neural networks," 2018, iSSN: 2413-6689. [Online]. Available: https://www.dafx.de/paper-archive/details/nBQy7u-y-TRD67IzjekyOA

[7] S. Nercessian, "Neural Parametric Equalizer Matching Using Differentiable Biquads," in *International Conference on Digital Audio Effects (DAFx)*. eDAFx: DAFx, 2020, iSSN: 2413-6689.

[8] F. Mockenhaupt, J. S. Rieber, and S. Nercessian, "Automatic Equalization for Individual Instrument Tracks Using Convolutional Neural Networks," in *27th International Conference on Digital Audio Effects (DAFx24)*, Guildford, Surrey, UK, Sep. 2024. [Online]. Available: https://www.dafx.de/paper-archive/2024/papers/DAFx24_paper_27.pdf

[9] C. J. Steinmetz, N. J. Bryan, and J. D. Reiss, "Style Transfer of Audio Effects with Differentiable Signal Processing," *Journal of the Audio Engineering Society*, vol. 70, no. 9, pp. 708–721, Sep. 2022. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=21883

[10] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," Sep. 2019. [Online]. Available: https://openreview.net/forum?id=B1x1ma4tDr

[11] C. J. Steinmetz and J. D. Reiss, "Efficient neural networks for real-time modeling of analog dynamic range compression," in *152nd Audio Engineering Society Convention*, Apr. 2022. [Online]. Available: http://arxiv.org/abs/2102.06200

[12] P. Sarkar and P. Lindborg, "Neural-Driven Multi-Band Processing for Automatic Equalization and Style Transfer," in *28th International Conference on Digital Audio Effects (DAFx25)*, Ancona, Italy, Sep. 2025.

[13] J. D'appolito, "Active Realization of Multiway All-Pass Crossover Systems," *Journal of The Audio Engineering Society*, Apr. 1987.

[14] C.-Y. Yu, C. Mitcheltree, A. Carson, S. Bilbao, J. D. Reiss, and G. Fazekas, "Differentiable All-pole Filters for Time-varying Audio Systems." DAFx, Apr. 2024, arXiv:2404.07970 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2404.07970

[15] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 1526–1530. [Online]. Available: https://www.isca-archive.org/interspeech_2019/zen19_interspeech.html

[16] G. J. Mysore, "Can we Automatically Transform Speech Recorded on Common Consumer Devices in Real-World Environments into Professional Production Quality Speech?—A Dataset, Insights, and Challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, Aug. 2015. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6981922

[17] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive: (http://web.ku.edu/˜idea/readings/rainbow.htm).*, Nov. 2019. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443

[18] C. Steinmetz and J. D. Reiss, "auraloss: Audio-focused loss functions in PyTorch," in *Digital Music Research Network One-Day Workshop (DMRN*, London, UK, 2020.

[19] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Interspeech 2021*, Aug. 2021, pp. 2127–2131, arXiv:2104.09494 [eess]. [Online]. Available: http://arxiv.org/abs/2104.09494