DPDFNet: Boosting DeepFilterNet2 via Dual-Path RNN

Daniel Rika[*]*, Nino Sapir*[*]*, Ido Gus*

Ceva Inc.

Daniel.Rika@ceva-ip.com, Nino.Sapir@ceva-ip.com, Ido.Gus@ceva-ip.com

***Abstract*—We present DPDFNet, a causal single-channel speech enhancement model that extends DeepFilterNet2 architecture with dual-path blocks in the encoder, strengthening long-range temporal and cross-band modeling while preserving the original enhancement framework. In addition, we demonstrate that adding a loss component to mitigate over-attenuation in the enhanced speech, combined with a fine-tuning phase tailored for "always-on" applications, leads to substantial improvements in overall model performance. To compare our proposed architecture with a variety of causal open-source models, we created a new evaluation set comprising long, low-SNR recordings in 12 languages across everyday noise scenarios, better reflecting real-world conditions than commonly used benchmarks. On this evaluation set, DPDFNet delivers superior performance to other causal open-source models, including some that are substantially larger and more computationally demanding. We also propose an holistic metric named PRISM, a composite, scale-normalized aggregate of intrusive and non-intrusive metrics, which demonstrates clear scalability with the number of dual-path blocks. We further demonstrate on-device feasibility by deploying DPDFNet on Ceva-NeuPro™-Nano edge NPUs. Results indicate that DPDFNet-4, our second-largest model, achieves real-time performance on NPN32 and runs even faster on NPN64, confirming that state-of-the-art quality can be sustained within strict embedded power and latency constraints.**

***Index Terms*—Speech enhancement; Noise reduction; Dual-Path RNN; Real-time systems; Embedded systems.**

I. Introduction

SINGLE-channel speech enhancement (SE) aims to recover clear, natural speech from a noisy recording when only one microphone is available, a common setting in telephony, conferencing, mobile devices, and assistive products. Unlike multi-microphone arrays, the single-channel case has no spatial cues from additional microphones and must rely on time and frequency information, which makes reliable operation at low SNRs especially hard. For interactive use, models must run *causally* with low delay to avoid conversational lag and audio-video drift, and they should remain stable across changes in noise, rooms, and languages. While large, non-causal systems can excel offline, there is a growing need for methods that deliver high perceived quality when run in a streaming setting.

Early real-time speech enhancement often combined traditional signal processing with compact neural networks. Valin [1] proposed RNNoise, which predicts critical-band gains from features such as Bark-frequency cepstral coefficients and pitch correlations, then applies a frequency-domain pitch filter. This hybrid design achieves very low latency and embedded-class efficiency, and it outperforms minimum mean-square spectral estimators in perceptual quality. Research then advanced sequence modeling for streaming noise suppression. Westhausen and Meyer [2] introduced DTLN with two cascaded cores. One core operates in the frequency-domain and the other uses a learned time-domain representation. Both rely on LSTM layers with instant layer normalization, enabling stable frame-synchronous processing with fewer than one million parameters and benefiting from both magnitude and phase cues. Braun and Tashev proposed NSNet2 [3], a recurrent model that estimates spectral suppression gains from log power spectral features. Its level-invariant loss normalization and broad augmentation over SNR, spectral shape, and signal level yield a robust real-time baseline. Another line of work moved to the waveform domain. Defossez et al. [4] proposed DEMUCS, an encoder-decoder with U-Net-style skip connections and a recurrent bottleneck that operates directly on raw audio. Training uses a mix of waveform loss and multi-resolution spectral losses with strong augmentation. This approach improves naturalness and intelligibility under real-time constraints. Kong et al. proposed CleanUNet [5] as a refinement of this design. It strengthens the bottleneck with stacked masked self-attention and combines waveform and spectral objectives, including a high-band loss that improves the handling of silence and high-frequency detail. Although relatively large, CleanUNet maintains low latency and delivers consistent quality. Work in the time-frequency domain focused on stronger cross-band modeling while preserving local structure. Hao et al. [6] proposed FullSubNet, which encodes long-range context with a full-band LSTM. The frame-wise embedding is concatenated with local frequency neighborhoods and passed to a shared sub-band LSTM that predicts a complex ideal ratio mask [7] (cIRM) with tanh compression. This design links global spectral trends with local stationarity cues. Dual-path recurrent models were optimized to balance efficiency and streaming performance. Le et al. [8] proposed DPCRN, an encoder-decoder with causal two-dimensional convolutions and skip connections. A dual-path module models spectral structure within each frame and temporal dynamics across frames, under causal constraints. Rong et al. proposed GTCRN [9] as a direct follow-up. It uses grouped temporal convolutions and grouped dual-path recurrent units, merges high-frequency bands through an *equivalent*

[*]D. Rika and N. Sapir contributed equally to this work.
Code, pretrained models, and a real-time demo are available at https://github.com/ceva-ip/DPDFNet

rectangular bandwidth (ERB) filter bank, and applies temporal recurrent attention. The result keeps the parameter count in the tens of thousands while remaining competitive with heavier systems. Most recently, Pei et al. proposed aTENNuate [10], a deep state space autoencoder for streaming enhancement on raw audio. The model stacks causal state space blocks with skip connections and light resampling. It targets online operation with steady latency and compact size, and it trains with objectives in both time and frequency domains. Overall, the field has made a substantial progress from classical-plus-RNN hybrids to highly capable pure *end-to-end* deep neural-network models, for higher intelligibility and naturalness under challenging real-time constraints.

Although a wide array of methods developed in the recent years, including the aforementioned ones, we based our work over *Schröter et al.* [11]'s DeepFilterNet2 architecture: a compact two-stage architecture that offers a strong balance of efficiency and quality for streaming SE. This architecture provides a favorable trade-off between computational complexity and enhancement quality, enabling real-time, low-latency deployment. Moreover, the architecture permits training at lower sampling rates (e.g., 16 kHz) while deploying at higher rates with negligible performance loss. This leverages the wider availability of low-rate data and reduces the computational load of online augmentations. Motivated by both technical considerations and community interest in advanced speech enhancement, we believe that enhancing the current DeepFilterNet2 architecture will contribute much to the field.

Accordingly, we integrate causal dual-path modules into the DeepFilterNet2 encoder, resulting in *DPDFNet*. This enhanced design delivers a substantial performance gain while preserving the overall architecture. We also add an over-attenuation loss, used alongside the original *multi-resolution loss* [11], to mitigate rare cases where the model over-suppresses the target speech, and proposed a fine-tuning stage to exposed the models to long-range dependencies. Finally, we curate a new evaluation set focused on low-SNR conditions with realistic noise types (car, pub, office, etc.) with multilingual coverage spanning 12 languages from the Speech-MASSIVE test set [12].

The remainder of the paper is organized as follows. Section II defines the single-channel denoising problem, reviews the DeepFilterNet2 architecture, and introduces our dual-path module along with its integration into DeepFilterNet2. Section III details the training pipeline, including datasets, augmentation strategies and the loss functions. Section IV presents the experimental results along with our proposed fine-tuning phase and deployment on Ceva-NeuPro™-Nano. Section V concludes the work.

II. Methods

A. Denoising Framework

We consider a noisy speech signal modeled as

$$x = s * r + n, \tag{1}$$

where s denotes the clean speech signal, r is the room impulse response (RIR) characterizing the acoustic environment, and n represents additive background noise. This formulation captures both the reverberation introduced by the environment and the presence of noise.

Following the original DeepFilterNet2 framework, we apply the Short-Time Fourier Transform (STFT) to the time-domain signal x to obtain its time-frequency representation X. This representation is then used to derive two primary input features: *complex features* and *ERB features*.

The complex features consist of the lower 96 frequency bins of X, corresponding to frequencies up to 4800 Hz. This range is chosen because it encompasses the majority of the periodic components of speech. We denote these features as X_{df}. In the original DeepFilterNet2, deep-filters (DF) were applied only to these frequency bins.

To obtain the ERB features, the power spectrogram $|X|^2$ is passed through a bank of 32 ERB filters. This compresses the spectral information in a way that approximates the human auditory system's perception of frequency and energy. The resulting perceptual representation is denoted as X_{erb}.

The denoising framework consists of two sequential stages: *masking* and *reconstruction*.

In the masking stage, the model predicts ERB gains $G_{\mathrm{erb}}(k, b)$, where k denotes the time frame and b denotes the ERB bin index. These gains are applied to the noisy spectrogram as follows:

$$\begin{aligned} G(k, f) &= \mathrm{inter}(G_{\mathrm{erb}}(k, b)), \\ Y_G(k, f) &= X(k, f) \cdot G(k, f), \end{aligned} \tag{2}$$

where $G(k, f)$ is the full-band mask obtained via inverse interpolation of the ERB filter banks (practically, using the transposed ERB filter bank matrix), and $Y_G(k, f)$ is the masked spectrogram.

In the reconstruction stage, the model predicts complex DF coefficients $C(k, i, f_{\mathrm{df}})$, which are applied to the periodic part of the masked spectrogram up to 4800 Hz:

$$Y(k, f) = \sum_{i=0}^{N} C(k, i, f) \cdot Y_G(k - i + \ell, f), \tag{3}$$

where N is the DF order, ℓ is the *look-ahead* ($N = 5$ and $\ell = 2$ in our case), and the dot product is performed in the complex domain.

B. DeepFilterNet2 Architecture

To predict the ERB gains $G_{\mathrm{erb}}(k, b)$ and the DF coefficients $C(k, i, f_{\mathrm{df}})$, the DeepFilterNet2 architecture first uses an *Encoder* that extracts and fuses information from both ERB features and complex features into a unified latent representation, denoted by:

$$\mathcal{E} = \mathcal{F}_{\mathrm{enc}}\left(X_{\mathrm{erb}}, X_{\mathrm{df}}\right). \tag{4}$$

This embedding $\mathcal{E}$ then passes through the *ERB Decoder* $\mathcal{F}_{\mathrm{erb\_dec}}$ and the *DF Decoder* $\mathcal{F}_{\mathrm{df\_dec}}$ to predict the ERB gains and DF coefficients, respectively.

More specifically, the *Encoder* consists of two separate branches, each composed of several convolutional blocks. Each convolutional block comprises a separable convolutional layer with 64 channels and a kernel size of $(1, 3)$

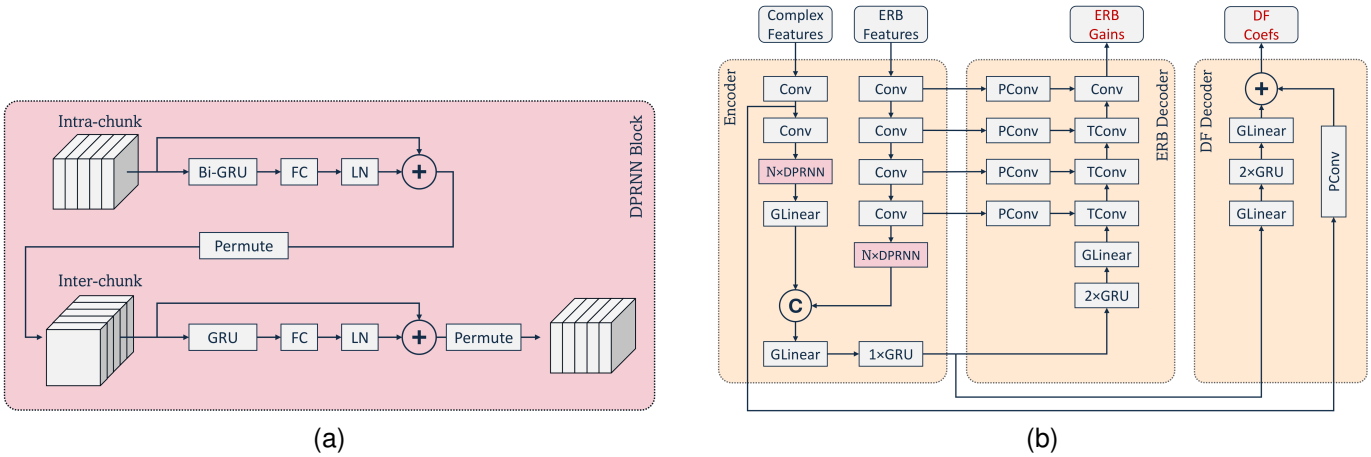

Fig. 1. Overview of the proposed DPDFNet architecture. (a) single DPRNN block; (b) DeepFilterNet2 scheme integrated with DPRNN blocks.

(time, frequency), followed by batch normalization and a ReLU activation function. To combine the extracted ERB and complex features, both feature representations are flattened, concatenated, and subsequently processed through *grouped linear* (GLinear) and GRU layers.

The ERB decoder contains two GRU layers, followed by several transposed convolutional blocks, to predict the ERB gains. Each transposed convolutional block comprises a separable transposed convolutional layer, batch normalization, and a ReLU activation, except for the final block, which uses a Sigmoid activation to predict the mask. To establish a U-Net-style architecture, convolutional blocks from the ERB encoder branch are connected to corresponding transposed convolutional blocks in the decoder through skip connections.

The DF decoder consists of two GRU layers, preceded by a GLinear layer and followed by another GLinear layer. Using the output from this structure combined with a skip connection from the encoder, the decoder predicts the complex DF coefficients.

C. Dual-Path RNN Block

Dual-Path Recurrent Neural Networks, introduced by Luo *et al.* [13], address long-context modeling through two alternating recurrent stages that operate along complementary axes of the data: an *intra stage* that aggregates information within each time frame and an *inter stage* that propagates information across time.

Let the input be a tensor $X \in \mathbb{R}^{B\times T\times F\times D}$, where B denotes batch size, T the number of time frames, F the number of frequency bins, and D the feature dimension.

Intra stage. To capture spectral dependencies within a time frame, reshape

$$X \mapsto X_{\text{intra}} \in \mathbb{R}^{B\cdot T\times F\times D}$$

and apply a bidirectional RNN along the F axis, one sequence per time frame. The output is then reshaped back to $\mathbb{R}^{B\times T\times F\times D}$.

Inter stage. To model temporal evolution at each frequency bin, permute and reshape

$$X \mapsto X_{\text{inter}} \in \mathbb{R}^{B\cdot F\times T\times D}$$

and apply a unidirectional RNN along the T axis, one sequence per frequency bin. A single parameter set is shared across all frequencies, while the recurrent states remain distinct. The result is finally reshaped to $\mathbb{R}^{B\times T\times F\times D}$.

This dual-path strategy was adopted by DPCRN [8], which replaces the GRU bottleneck in a U-Net-style encoder-decoder with DPRNN blocks and reports strong performance. In addition, DPCRN appends a position-wise fully connected (FC) layer and Layer Normalization (LN) [14] after each DPRNN stage, both intra and inter, in order to refine features and stabilize training. Figure 1a depicts the scheme of the proposed DPRNN block.

D. DPDFNet

As outlined in Section II-B, DeepFilterNet2 comprises three core components: a shared encoder and two decoders. The encoder takes two complementary feature branches, the ERB features and the Complex features. The ERB branch captures perceptually motivated spectral information, whereas the Complex branch emphasizes periodic structure up to 4800 Hz, which helps restore the naturalness of human speech.

Before fusion, each branch is processed in parallel by its own stack of convolutional blocks that extract local patterns from the corresponding input. The resulting representations are then merged in a combined layer. To enhance temporal and cross-frequency modeling beyond local contexts, we insert N DPRNN blocks immediately after the convolutional stacks for each branch. The number of channels in these convolutions is set to the feature dimension D. These DPRNN blocks enable richer interactions across time and frequency bins, improving the capacity to model complex speech characteristics. The complete DPDFNet architecture is shown in Figure 1b.

III. Training Framework

A. Datasets and Augmentations

Many recent SE models, including our baseline DeepFilterNet2, have been trained on the widely adopted Deep Noise Suppression 4 (DNS4) Challenge [15] dataset. This full-band

(48 kHz) dataset provides 760 hours of clean speech recordings in six languages from 3,230 unique speakers, along with 181 hours of noise comprising 62,000 clips across 150 noise classes. It also includes 248 real and 60,000 synthetic RIRs for generating reverberant signals. Although DNS4 remains a widely adopted dataset for training SE models, the clean speech portion is not entirely free of residual noise, which may constrain the achievable performance of trained models. In contrast, a variety of high-quality, open-source wideband (16 kHz) datasets exist for both clean speech and noise, offering valuable resources for more effective model training. Furthermore, since the DeepFilterNet2 architecture can be adapted to full-band operation via a lightweight postprocessing step with negligible impact on performance, utilizing these larger and higher-quality wideband datasets presents a particularly compelling direction for training. Therefore, we downsampled the DNS4 dataset to wideband and enriched it with 4,000 hours of English read speech from the Multilingual LibriSpeech (MLS) corpus [16], a large-scale dataset derived from LibriVox released by Facebook. This addition improves both quality and speaker diversity compared to DNS4 alone. To further ensure the cleanest possible speech signals, all audio was processed through the DPCRN [8] model. Furthermore, we incorporated noise clips from the MUSAN [17] and FSD50K [18] datasets to further diversify the noise conditions. To improve model generalization, a set of online data augmentations is applied during training. Random second-order filtering is used to expose the model to natural spectral colorations, improving robustness across devices and recording environments. Gain perturbations are introduced to prevent overfitting to specific loudness levels and to ensure invariance to overall volume. In addition, reverberation with early reflection targets is applied, enabling the model to suppress late reverberation while preserving the direct path and early reflections that are beneficial for speech intelligibility.

B. New Evaluation Set

To evaluate and compare different SE models, two test sets are commonly employed: (1) the VoiceBank+DEMAND test set [19] and (2) the DNS4 blind test set [15]. While these datasets have been widely adopted as benchmarks within the single-channel SE domain, they fail to fully capture the diversity of real-world acoustic conditions. For instance, the VoiceBank+DEMAND test set primarily consists of short clips, typically only a few seconds in duration, recorded at relatively high SNRs. Conversely, the DNS4 blind test set provides longer recordings; however, each clip includes only a single speaker and lacks extended periods of background noise without speech - an element often present in realistic environments. Furthermore, both datasets exhibit limited linguistic diversity, as they consist almost exclusively of English speech.

In light of these limitations, we developed a new evaluation set designed to better approximate real-world scenarios. Specifically, we incorporated nine noisy environments that reflect common day-to-day situations: *airport*, *car*, *office*, *pub*, *rain*, *restaurant*, *street*, *subway*, and *train*. Given the inherently high noise levels in these settings, we selected low SNR values of 0, 5, and 10 dB. To assess cross-linguistic generalization, we used clean speech samples from 12 languages included in the Speech-MASSIVE test set: *Arabic*, *Dutch*, *French*, *German*, *Hungarian*, *Korean*, *Polish*, *Portuguese*, *Russian*, *Spanish*, *Turkish*, and *Vietnamese*. Each clip is approximately 2.5 minutes in duration and features a unique combination of environmental noise, SNR level, and language. Within a single clip, multiple speakers may occur, and speech-free intervals of up to 15 seconds are included to enhance realism. In total, the constructed dataset comprises 324 clips, amounting to roughly 13.5 hours of challenging evaluation material.

C. Loss Functions

We adopt the Multi-Resolution loss from [11]. The enhanced signal y is analyzed with several STFTs, specifically with window sizes $i \in \{5, 10, 20, 40\}$ ms, to form the following loss:

$$\mathcal{L}_{\mathrm{MR}} = \sum_i \left\| \widetilde{Y}_i - \widetilde{S}_i \right\|_2^2 + \left\| \widehat{Y}_i - \widehat{S}_i \right\|_2^2 \tag{5}$$

where the magnitude and phase are represented as

$$\widetilde{Y}_i = |Y_i|^c, \quad \widetilde{S}_i = |S_i|^c \tag{6}$$

$$\widehat{Y}_i = \widetilde{Y}_i e^{j\phi_{Y_i}}, \quad \widehat{S}_i = \widetilde{S}_i e^{j\phi_{S_i}} \tag{7}$$

For each resolution i, let

$$Y_i = \mathrm{STFT}_i\{y\}, \qquad S_i = \mathrm{STFT}_i\{s\},$$

denote the complex spectrograms of the enhanced and clean signals, respectively, with phases ϕ_{Y_i} and ϕ_{S_i}. The magnitude compression exponent set to $c = 0.3$ follow the original DeepFilterNet2 training [11].

During our experiments, we observed that the models suffered to over-attenuation (OA). To address this, we added a loss which penalizes enhanced bins with relatively low energy comparing to the clean target. For each resolution i, define a freq-bin binary mask as follows:

$$\mathbf{M}_i(k, f) = \mathbb{1}\{|\mathbf{S}_i|(k, f) > |\mathbf{Y}_i|(k, f)\} \tag{8}$$

which then applied on the same Multi-Resolution loss:

$$\mathcal{L}_{\mathrm{OA}} = \sum_i \left\| (\widetilde{Y}_i - \widetilde{S}_i) \odot M_i \right\|_2^2 + \left\| (\widehat{Y}_i - \widehat{S}_i) \odot M_i \right\|_2^2 \tag{9}$$

The total objective combines the both of the lost functions:

$$\mathcal{L} \;=\; \lambda_{\mathrm{MR}} \mathcal{L}_{\mathrm{MR}} + \lambda_{\mathrm{OA}} \mathcal{L}_{\mathrm{OA}}. \tag{10}$$

where $\lambda_{\mathrm{MR}} = \lambda_{\mathrm{OA}} = 500$.

A detailed quantitative evaluation, contrasting performance with and without OA loss, is summarized in Table I.

TABLE I
INTRUSIVE, NON-INTRUSIVE AND PRISM SCORES ON OUR NEW MULTILINGUAL LOW-SNR TEST SET FOR OPEN-SOURCE CAUSAL MODELS AND OUR DPDFNET-$\{k\}$ VARIANTS (k = NUMBER OF DPRNN BLOCKS). THE BASELINE USES THE DEEPFILTERNET2 ARCHITECTURE WITHIN OUR TRAINING FRAMEWORK.

Model	**Params** [M]	**MACs** [G]	**PESQ**	**STOI**	**SI-SNR**	**DNSMOS**				**NISQA**					**PRISM**
						SIG	**BAK**	**OVL**	**P.808**	**MOS**	**NOI**	**DIS**	**COL**	**LOUD**	
Noisy	–	–	1.27	83.2	0.38	2.02	1.66	1.56	2.53	2.00	1.69	3.40	2.68	2.68	0.04
DTLN [2]	0.99	0.12	1.94	88.8	10.83	2.51	3.50	2.15	2.85	2.13	2.51	2.79	2.39	2.85	0.46
GTCRN [9]	**0.023**	**0.039**	2.00	87.2	9.11	2.56	3.76	2.24	3.00	2.53	3.27	3.17	2.80	3.01	0.49
RNNoise [1]	0.087	0.04	2.05	88.5	8.97	2.81	3.93	2.50	3.01	1.88	3.68	2.20	2.19	3.14	0.51
NSNet2 [3]	2.60	0.26	2.06	86.8	8.63	2.68	3.82	2.36	2.92	2.63	3.39	3.25	2.76	3.30	0.51
FullSubNet [6]	5.60	30.00	2.19	89.6	10.63	2.90	3.41	2.43	3.06	2.76	2.78	3.83	3.41	3.29	0.63
aTENNuate [10]	0.80	0.33	2.43	88.9	9.65	3.05	4.03	2.74	2.94	3.01	3.35	3.90	2.97	3.92	0.69
DPCRN [8]	0.53	1.10	2.47	90.7	11.93	2.73	3.78	2.39	3.03	2.92	3.11	3.61	3.23	3.40	0.69
CleanUNet [5]	46.07	15.44	2.38	91.2	11.67	2.97	3.95	2.66	3.07	2.57	3.41	2.72	2.86	3.57	0.69
DEMUCS [4]	33.53	7.70	2.27	91.5	12.36	3.00	3.95	2.69	3.09	2.52	3.56	2.70	2.94	3.79	0.71
DeepFilterNet2 [11]	2.31	0.36	2.35	91.3	11.95	2.92	3.87	2.58	3.19	3.25	3.74	3.62	3.47	3.60	0.76
DeepFilterNet3 [20]	2.14	0.35	2.45	90.6	12.10	3.03	4.01	2.71	3.22	3.84	4.12	3.98	3.75	3.85	0.82
Baseline*	2.31	0.36	2.48	89.2	10.25	3.10	4.08	2.80	3.16	3.95	4.37	4.17	3.87	4.02	0.80
+OA Loss	2.31	0.36	2.56	90.2	11.21	3.13	4.06	2.82	3.20	3.80	4.28	4.07	3.79	4.00	0.84
+Fine-Tuning	2.31	0.36	2.68	91.6	13.27	3.14	4.07	2.83	3.21	3.96	4.31	4.17	3.83	4.05	0.91
DPDFNet-2	2.49	1.35	2.79	92.6	13.72	3.16	4.08	2.85	3.25	4.06	4.36	4.23	3.92	4.10	0.95
DPDFNet-4	2.84	2.36	2.82	93.0	14.11	3.17	4.08	2.86	**3.28**	4.15	4.40	4.30	3.96	4.14	0.98
DPDFNet-8	3.54	4.37	**2.85**	**93.4**	**14.47**	**3.19**	**4.09**	**2.89**	**3.28**	**4.21**	**4.43**	**4.34**	**4.00**	**4.18**	**1.00**

*This is the baseline without OA loss and the additional fine-tuning phase.

TABLE II
DNSMOS RESULTS COMPARISON BETWEEN DEEPFILTERNET2/3 TO OUR PROPOSED DPDFNET VARIANTS ON THE DNS4 BLIND TEST SET.

Model	SIG	BAK	OVL	P.808
Noisy	3.231	2.406	2.257	3.013
DeepFilterNet2	3.318	3.997	3.019	3.748
DeepFilterNet3	3.279	3.947	2.958	3.729
Baseline	3.374	4.052	3.091	3.815
DPDFNet-2	3.386	4.054	3.104	3.844
DPDFNet-4	3.389	4.047	3.105	3.848
DPDFNet-8	**3.403**	**4.060**	**3.122**	**3.849**

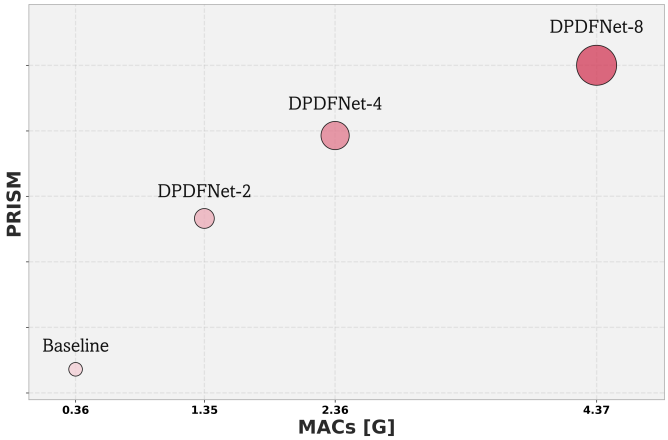

Fig. 2. Impact of dual-path block depth in DPDFNet on PRISM performance, showing a direct performance gain with depth, alongside changes in complexity (X-axis) and footprint (bubble size).

IV. EXPERIMENTAL RESULTS

A. Implementation Details

We evaluate model performance using a combination of established *intrusive* and *non-intrusive* evaluation metrics. The intrusive metrics include PESQ [21], STOI [22], and SI-SNR [23], which rely on comparisons between the enhanced signal and a clean, time-aligned reference. The non-intrusive metrics consist of DNSMOS P.835 (SIG, BAK, and OVRL [24]), P.808 MOS [25], and NISQA v2.0 [26], which evaluates Overall Quality, Noisiness, Coloration, Discontinuity, and Loudness. Audio files were resampled to 16 kHz prior to processing. Training examples were created by segmenting clean speech into 3 sec chunks and mixing them with noise at SNRs randomly sampled from -5 dB to 40 dB. For STFT parameters, we followed the original DeepFilterNet2 setup, employing a window length of 20 ms, a hop size of 10 ms, and a Vorbis window function. We trained the models for 1.6M iterations with a batch size of 32, employing the AdamW [27] optimizer. The learning rate was decayed according to a cosine-annealing schedule, with a pick value of 1e-3.

Among the open-source, we select only those that meet two key criteria: (1) they are causal models suitable for streaming use, and (2) their weights are publicly available, allowing us to run the official versions on our evaluation set. Consequently, all models mentioned in Section I are included in our evaluations. For the models introduced in this work, we begin with a *Baseline* model that starts from the same architecture as DeepFilterNet2 [11], retrained within our framework for

methodological consistency. We then present three variants of our proposed DPDFNet architecture: *DPDFNet-2*, *DPDFNet-4*, and *DPDFNet-8*. The suffix $k \in 2, 4, 8$ indicates the number of DPRNN blocks in the dual-path stack, scaling the model's capacity from small ($k = 2$), to medium ($k = 4$), and large ($k = 8$).

B. Fine-Tuning

After training the models for 1.6M iterations - covering 42,667 hours of speech - we obtained high-quality speech enhancement systems. However, when deployed in an "always-on" streaming setting, we observed instability, including truncation of the initial speech segment following extended silence (3-4 s) and occasional leakage of background noise after such intervals. We hypothesize that these issues arise because stateful components were not sufficiently exposed to state transitions during training, as the models were primarily trained on short, isolated segments.

To address this, we performed an additional fine-tuning stage with batch size of 1, continuous input segments of 30-40 sec, and 5,000 iterations at a fixed learning rate of 1e-5. This procedure encourages the model to better handle long-term temporal dependencies and transitions between speech and non-speech regions encountered in real-time operation.

Following this fine-tuning, all our proposed models exhibited markedly improved stability during continuous inference. A quantitative comparison, with and without this fine-tuning, on our evaluation set is provided in Table I.

C. Performance Relative Integrated Scaled Metric

While the individual metrics mentioned in Section IV already able to demonstrate the strength of a SE model, they each emphasize different aspects of enhancement quality. Intrusive measures reward reconstruction fidelity but may undervalue perceptual improvements, whereas non-intrusive predictors better reflect subjective listening impressions but can sometimes favor aggressive noise suppression. As such, viewing each metric in isolation can obscure the holistic trade-off between speech preservation, noise removal, and overall perceptual quality.

To address this limitation, we introduce the *Performance Relative Integrated Scaled Metric* (PRISM). PRISM employs a hierarchical scoring framework that integrates both intrusive and non-intrusive objective metrics into a single normalized score.

At the first stage, each metric group Intrusive, DNSMOS P.808 & P.835, and NISQA is individually normalized using *min-max normalization* across all models, mapping the lowest-performing result to 0 and the highest to 1. The normalized metrics are then averaged within each group, forming three composite scores: one for intrusive, one for DNSMOS, and one for NISQA. The DNSMOS and NISQA composites are subsequently combined to represent the non-intrusive category.

Finally, the overall PRISM score is obtained by taking the mean of the intrusive and non-intrusive composite scores into a unified measure of model performance. This hierarchical design provides a more interpretable and balanced aggregation of quality metrics, capturing both signal-based and perceptual performance aspects on a consistent, unified scale. See Table I for the PRISM scores of all evaluated models.

Now that we have a unified metric through PRISM, we can more clearly examine the trade-offs between model quality and efficiency. Figure 2 plots PRISM against model complexity (MACs) and bubble size (#Params). The results reveal a consistent pattern: as the number of DPRNN blocks increases from 0 (*i.e.*, Baseline) to 8, the PRISM score rises monotonically, indicating that quality improvements scale reliably with depth. Meanwhile, computational cost and model size grow only moderately. This offers deployment flexibility: smaller variants are preferable when efficiency is critical, while deeper ones can be chosen when maximizing enhancement quality is the priority.

D. Results

We begin by presenting the objective evaluation of all models on the new multilingual low-SNR evaluation set. Table I reports both intrusive and non-intrusive metrics, including reconstruction fidelity measures (PESQ, STOI, SI-SNR) and perceptual quality predictors (DNSMOS P.835 SIG/BAK/OVRL, P.808 MOS, and NISQA). This comprehensive set of indicators enables us to assess not only how accurately speech is restored relative to the clean reference, but also how the enhanced signals are expected to be perceived by human listeners.

Several clear trends emerge. The Baseline model - without the OA loss and fine-tuning phase - shows comparable performance to DeepFilterNet2 and its successor, DeepFilterNet3. However, incorporating these two components into the training pipeline allows the Baseline model to surpass both across all metrics (except for P.808, where DeepFilterNet3 leads by only 0.01 points), highlighting their critical role in achieving high-performing speech enhancement.

Turning to our proposed DPDFNet architectures, all variants consistently outperform the baseline models. Moreover, performance scales positively with the number of blocks (*i.e.*, $k \in \{2, 4, 8\}$), with the largest model, DPDFNet-8, achieving the best overall results. When compared to strong causal baselines such as DPCRN, DPDFNet-8 delivers consistently higher fidelity and perceptual quality. Even against substantially larger waveform-domain models like DEMUCS and CleanUNet, DPDFNet variants achieve superior enhancement performance while requiring significantly fewer parameters and computational resources. Overall, the DPDFNet family compares favorably with both lightweight and heavyweight alternatives.

For completeness, although we consider the two benchmarks mentioned in Section III-B to be less representative of real-world conditions, we chose the DNS4 blind test set [15] because it provides longer, 10-second audio samples. We therefore evaluated our proposed models on DNS4 and compared their DNSMOS metrics with those of DeepFilterNet2/3. The results are shown in Table II.

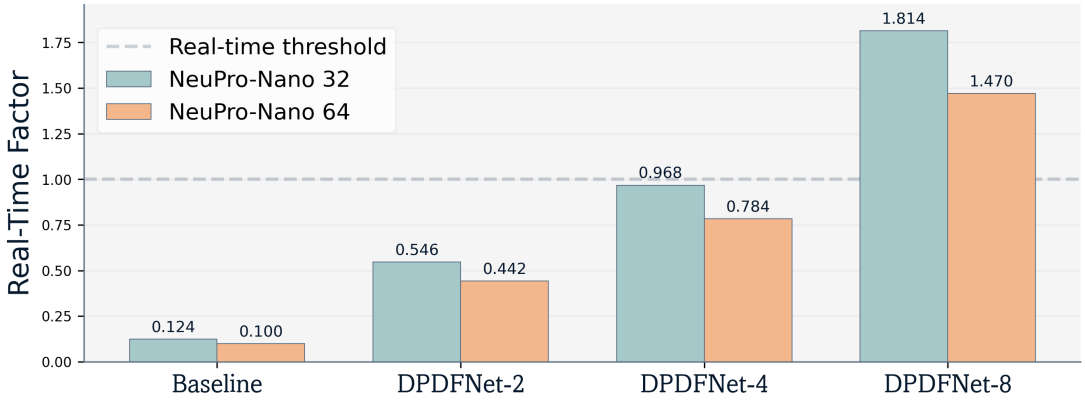

Fig. 3. Real-time factor of our proposed models on NPN32 and NPN64, assuming deployment with **int8** weights and **int16** activations.

E. Deployment DPDFNet on NeuPro-Nano

Ceva-NeuPro™-Nano[1] (NPN) is a programmable edge NPU IP optimized for embedded deep-learning under stringent power and area constraints. It integrates control, DSP, and neural execution within a single core, removing the need for an external host processor. The architecture supports two configurations (*NPN32* and *NPN64*), integer precisions from 4 to 32 bits, and includes advanced features such as native transformer operators, sparsity acceleration, fast re-quantization, and Ceva's NetSqueeze™ technology for on-the-fly weight compression to minimize memory traffic.

Meeting real-time (RT) constraints is a key challenge when deploying SE models on hardware with limited memory and computational resources. Many prior works reported RT factors measured on high-performance processors such as the Intel i7. While useful as a baseline, these results do not reflect the conditions of actual edge devices, such as earbuds or smartwatches, where strict power limitations must be considered. To bridge this gap, we evaluate our models directly on edge-oriented cores, NPN32 and NPN64. The results show that while the largest model, DPDFNet-8, falls short of real-time performance, DPDFNet-4 successfully achieves it on NPN32 with an RT factor of 0.97. Furthermore, NPN64 delivers a performance improvement, consistently lowering RT factors across all model variants.

A detailed comparison of both cores is presented in Figure 3.

V. Conclusions

This work introduced *DPDFNet*, an extension of DeepFilterNet2 that augments the encoder with causal dual-path recurrent blocks to strengthen long-range temporal and cross-band modeling under streaming constraints. To facilitate realistic assessment of causal speech enhancement, we curated a multilingual, low-SNR evaluation set covering diverse everyday acoustic scenes, and proposed PRISM, a scale-normalized composite metric that integrates intrusive and non-intrusive measures.

Across this new evaluation set, DPDFNet variants consistently outperform strong causal open-source baselines, including models with substantially higher parameter counts and computational demands. Performance improves monotonically with the depth of the dual-path stack, as reflected by the PRISM aggregate, indicating that the added modeling capacity translates into perceptual gains.

Finally, we validated embedded feasibility on Ceva-NeuPro™-Nano edge NPUs: DPDFNet-4 achieves real-time performance on NPN32 and demonstrates even higher throughput on NPN64. These results indicate that high-quality, causal single-channel enhancement can be delivered within stringent power and latency budgets typical of on-device applications.

[1]https://www.ceva-ip.com/product/ceva-neupro-nano/

References

[1] J.-M. Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," 2018. [Online]. Available: https://arxiv.org/abs/1709.08243

[2] N. Westhausen and N. Zeghidour, "Dual-signal transformation lstm network for real-time noise suppression," in *Proc. Interspeech*, 2020, pp. 2472–2476. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2020/westhausen20_interspeech.html

[3] S. Braun, H. Gamper, C. K. A. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," 2021. [Online]. Available: https://arxiv.org/abs/2101.09249

[4] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," 2020. [Online]. Available: https://arxiv.org/abs/2006.12847

[5] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," 2022. [Online]. Available: https://arxiv.org/abs/2202.07790

[6] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, jun 2021. [Online]. Available: http://dx.doi.org/10.1109/ICASSP39728.2021.9414177

[7] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[8] X. Le, H. Chen, K. Chen, and J. Lu, "Dpcrn: Dual-path convolution recurrent network for single channel speech enhancement," in *Proc. Interspeech*, 2021, pp. 2811–2815. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2021/le21b_interspeech.html

[9] X. Rong, T. Sun, X. Zhang, Y. Hu, C. Zhu, and J. Lu, "Gtcrn: A speech enhancement model requiring ultralow computational resources," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 971–975.

[10] Y. R. Pei, R. Shrivastava, and F. Sidharth, "atennuate: Optimized real-time speech enhancement with deep ssms on raw audio," 2025. [Online]. Available: https://arxiv.org/abs/2409.03377

[11] H. Schröter, A. N. Escalante-B., T. Rosenkranz, and A. Maier, "Deepfilternet2: Towards real-time speech enhancement on embedded devices for full-band audio," 2022. [Online]. Available: https://arxiv.org/abs/2205.05474

[12] B. Lee, I. Calapodescu, M. Gaido, M. Negri, and L. Besacier, "Speech-MASSIVE: A Multilingual Speech Dataset for SLU and Beyond," in *Proc. Interspeech 2024*, 2024.

[13] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," 2020. [Online]. Available: https://arxiv.org/abs/1910.06379

[14] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016. [Online]. Available: https://arxiv.org/abs/1607.06450

[15] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "Icassp 2022 deep noise suppression challenge," 2022. [Online]. Available: https://arxiv.org/abs/2202.13288

[16] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in *Interspeech 2020*. ISCA, Oct. 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2826

[17] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," 2015. [Online]. Available: https://arxiv.org/abs/1510.08484

[18] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: An open dataset of human-labeled sound events," 2022. [Online]. Available: https://arxiv.org/abs/2010.00475

[19] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 2016, pp. 146–152.

[20] H. Schröter, T. Rosenkranz, A. N. Escalante-B., and A. Maier, "Deepfilternet: Perceptually motivated real-time speech enhancement," 2023. [Online]. Available: https://arxiv.org/abs/2305.08227

[21] 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. IEEE ICASSP*, 2001, pp. 749–752.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011. [Online]. Available: https://sps.ewi.tudelft.nl/pubs/Taal2011_1.pdf

[23] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *Proc. IEEE ICASSP*, 2019, pp. 626–630. [Online]. Available: https://www.jonathanleroux.org/pdf/LeRoux2019ICASSP05sdr.pdf

[24] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," 2022. [Online]. Available: https://arxiv.org/abs/2110.01763

[25] ——, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," 2021. [Online]. Available: https://arxiv.org/abs/2010.15258

[26] 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*. ISCA, Aug. 2021. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2021-299

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://arxiv.org/abs/1711.05101