SINERGI Vol. 29, No. 2, June 2025: 269-278
http://publikasi.mercubuana.ac.id/index.php/sinergi
http://doi.org/10.22441/sinergi.2025.2.001

Performance of speech enhancement models in video conferences: DeepFilterNet3 and RNNoise

Muhammad Iqbal Maulana, Muhammad Fadhlillah Raisul Akbar, Zendi Iklima*
Department of Electrical Engineering, Faculty of Engineering, Universitas Mercu Buana, Indonesia

Abstract
As remote work and online education continue to gain prominence, the importance of clear audio communication becomes crucial. Deep Learning-based Speech Enhancement has emerged as a promising solution for processing data in noisy environments. In this study, we conducted an in-depth analysis of two speech enhancement models, RNNoise and DeepFilterNet3, selected for their respective strengths. DeepFilterNet3 leverages time-frequency masking with a Complex Mask filter, while RNNoise employs Recurrent Neural Networks with lower complexity. The performance evaluation in training revealed that RNNoise demonstrated impressive denoising capabilities, achieving low loss values, while DeepFilterNet3 showed superior generalization. Specifically, "DeepFilterNet3 (Pre-Trained)" exhibited the best overall performance, excelling in intelligibility and speech quality. RNNoise also performed well in subjective quality measures. Furthermore, we assessed the real-time processing efficiency of both models. Both RNNoise variants processed speech signals almost in real-time, whereas DeepFilterNet3, though slightly slower, remained efficient. The findings demonstrate significant improvements in speech quality, with "DeepFilterNet3 (Pre-Trained)" emerging as the top-performing model. The implications of this study have the potential to enhance video conference experiences and contribute to the improvement of remote work and online education.

Keywords:
Accuracy Metrics;
DeepFilterNet3;
RNNoise;
Speech Enhancement;
Speed Metrics;

Article History:
Received: July 30, 2024
Revised: October 5, 2024
Accepted: November 7, 2024
Published: May 1, 2025

Corresponding Author:
Zendi Iklima
Electrical Engineering Department, Faculty of Engineering, Universitas Mercu Buana, Indonesia
Email:
zendi.iklima@mercubuana.ac.id

INTRODUCTION

As the world embraces remote work, video conferences have become essential communication tools [1]. They enable seamless connection and collaboration, overcoming physical distance in today's fast-paced world [2]. However, audio quality often suffers due to background noise, echo, and poor microphones, impacting clear communication [3]. While noise in online learning moderately affects students' task performance, overall noise levels don't significantly impact perceived performance difficulties [4].

Speech Enhancement techniques have gained significance [5] as they tackle the challenge of improving sound quality in video conferences. However, traditional noise reduction methods have limitations, leading to speech signal distortions. Therefore, there is a growing need for new and innovative approaches to ensure a better user experience [5].

Recent advance of deep learning method has been widely implemented in many sectors such as medical [6], industry, robotic [7], virtual reality, and so on. In sector virtual conference, deep learning-based of speech enhancement [8] have demonstrated promising results in addressing challenges related to processing data in noisy environments [9]. Complex value processing, particularly through the technique of time-frequency masking using a Complex Mask (CM) filter, has significantly improved Deep Learning for Speech Enhancement and signal extraction [10]. Unlike real-valued filters, CM allows phase modification, leading to enhanced

signal quality [11]. The application of CM in speech enhancement and signal extraction tasks has shown better outcomes [12].

Through the comparative study of various models for speech quality enhancement, including RNNoise [13], DCCRN [14], NSNet2 [15], FullSubNet+ [16], GaGNet [17], FRCRN [18], DeepFilterNet [19], DeepFilterNet2 [20], and DeepFilterNet3[21], two models, RNNoise and DeepFilterNet3, were specifically chosen for in-depth analysis in this research due to their promising performance in different aspects. RNNoise was selected as a lightweight and efficient solution, while DeepFilterNet3 demonstrated state-of-the-art speech quality enhancement with a good balance between computational efficiency and superior results.

DeepFilterNet is an algorithmic model [21] that enhances speech quality using CM for speech enhancement through Deep Filtering. It consists of two stages: one for enhancing the spectral envelope with human frequency perception modeling [19], and another for enhancing periodic speech components using deep filtering. The approach leverages speech's perceptual nature and enforces network sparsity through separable and extensive grouping in linear and recurrent layers [9]. The same researchers extended the model to create DeepFilterNet2, optimized for real-time video conference Speech Enhancement, achieving better performance and running on embedded devices like Raspberry Pi 4 with a real-time factor of 0.42 [19]. This makes DeepFilterNet2 ideal for speech enhancement in video conferences and real-time applications. The model was further refined into DeepFilterNet3, improving noise suppression performance, fixing bugs, and introducing attenuation limiting at the implementation level using the LADSPA plugin [20].

In other way RNNoise, which shares the same function but relies on Recurrent Neural Networks (RNN) with lower complexity [13]. While DeepFilterNet3 and RNNoise has shown promising results in offline scenarios [20], its real-time performance, especially in video conferences for remote learning, remains understudied. This study aims to comprehensively evaluate both performance in real-time speech enhancement during video conferences through simulation. The assessment will cover training, validation, and final testing using selected samples. The model will be compared against Both models will be trained on the Voicebank + DEMAND dataset, containing multi person speech and noise samples [22], for comparative analysis. Additionally, the MIT IR Survey dataset of Room Impulse Responses (RIR) [23] will be used for training DeepFilterNet3.

The evaluation will compare two models in Speech Enhancement using samples from the JL Corpus dataset [24] for Clean Speech and the ESC 50 dataset [25] for background sounds. Key metrics include PESQ [26], STOI [27], "CSIG" (Signal-to-Noise Ratio improvement), "CBAK" (distortion introduced), and "COVL" (coherence) [28]. Additional metrics are SegSNR [29] and SiSDR [30]. Model parameters (Params[M]) [31], computational complexity (MACS[G]) [32], and Real-Time Factor (RTF) [33] will also be considered. These comparisons will identify each model's strengths and weaknesses for specific applications.

In conclusion, this research aims to analyze the performance of Speech Enhancement in audio for video conferences using the DeepFilterNet3 model and RNNoise as the tested implementation in online learning cases. The study can potentially enhance video conference experiences and improve remote work and online education.

MATERIALS AND METHODS

This chapter explores the selected speech enhancement models and the datasets used in our study.

DeepFilterNet3

DeepFilterNet3 [21] is an advanced real-time speech enhancement model, an evolution of the DeepFilterNet framework. It utilizes deep filtering to estimate complex filters in the frequency domain, effectively leveraging short-term correlations in speech signals. The model achieves high efficiency by combining domain knowledge of speech production and psychoacoustic perception while providing comparable performance to state-of-the-art speech enhancement benchmarks. DeepFilterNet3 has been trained on a multilingual dataset, delivering improved results on objective evaluation metrics. The model's real-time capabilities enable users to experience enhanced audio directly, with the flexibility to configure noise suppression settings dynamically. Additionally, DeepFilterNet3 can be applied for direct noise reduction during activities like video calls, making it a valuable tool for various speech processing and communication systems. DeepFilterNet processes audio signals at a high 48 kHz sampling rate, making it well-suited for high-fidelity speech enhancement applications. It operates on 20 ms frames, using a 10 ms hop size to overlap windows, which helps maintain temporal

consistency. The additional 2-frame look-ahead contributes to a total latency of 40 ms, which, while slightly higher than real-time, is still low enough for many interactive applications such as voice communication or hearing aids.

In the first stage, DeepFilterNet uses the real-valued ERB (Equivalent Rectangular Bandwidth) domain, a perceptually-motivated scale that mirrors how the human ear processes sound. Here, 32 ERB-scaled gains are predicted, which are then pointwise multiplied with the noisy spectrum. This process helps reconstruct the speech envelope, which is essential for maintaining the naturalness and intelligibility of speech while reducing noise. The second stage focuses on refining lower-frequency components by applying a 5-tap complex filter, limited to the lowest 96 frequency bins, corresponding to frequencies up to 4.8 kHz. This focus on lower frequencies is important, as speech intelligibility largely depends on these frequencies. For higher frequencies, the model leverages the ERB-gain output from the first stage, which is computationally efficient and helps preserve the finer details of the speech signal.By combining the DF output for lower frequencies with the ERB gain output for higher frequencies, DeepFilterNet effectively balances computational efficiency and enhancement quality, making it suitable for real-time or near-real-time applications in noise suppression, such as in teleconferencing, hearing aids, or mobile devices. Figure 1 represent DeepFilterNet framework.

The mixture signal can be identified as $x(k)$, where combine the clean speech signal $s(k)$ and the interfering background noise $z(k)$, can be formulated as,

$$x(k) = s(k) + z(k) \quad (1)$$

Additionally, the noise reduction can be operated in time domain, can be formulated as,

$$X(t,f) = S(t,f) + Z(t,f) \quad (2)$$

where $X(t,f)$ is the STFT representation of the time domain signal in $x(k)$ and t, and f represents as time and frequency bins. Deep filtering for a complex filter in TF-domain can be formulated as

$$Y(k,f) = \sum_{i=0}^{N} C(k,i,f).X(k-i+l,f) \quad (3)$$

where C is the complex coefficients or filter order N that are applied to the input spectrogram X, and Y is the spectrogram enhacement.

RNNoise

The RNNoise [13] speech enhancement model is a real-time noise suppression algorithm that combines classic signal processing with deep learning techniques.

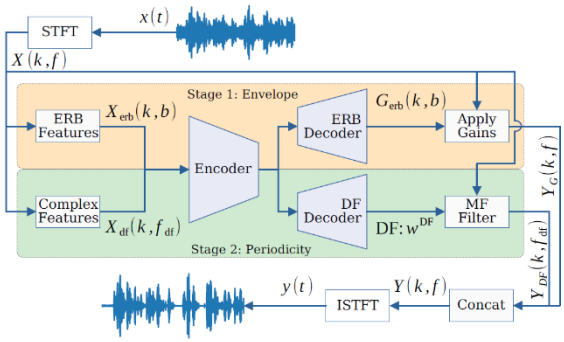

Figure 1. DeepFilterNet3 Block Diagram

It utilizes Gated Recurrent Units (GRUs), a type of recurrent neural network, to model time sequences, enabling effective noise estimation and suppression. The input to the neural network consists of cepstral coefficients based on the Bark scale, along with derivatives, pitch period, pitch gain, and a non-stationarity value. The neural network computes per-band gains to attenuate noise while preserving the speech signal. The model's deep architecture comprises three GRU layers, which map to the traditional steps of noise suppression (presented in Figure 2). Training data is generated by combining separate recordings of clean speech and various types of noise. The design and training of the neural network are performed in Python using Keras, while the runtime code is implemented in C to achieve real-time processing. The resulting model is small and efficient, making it suitable for deployment on resource-constrained devices like the Raspberry Pi.

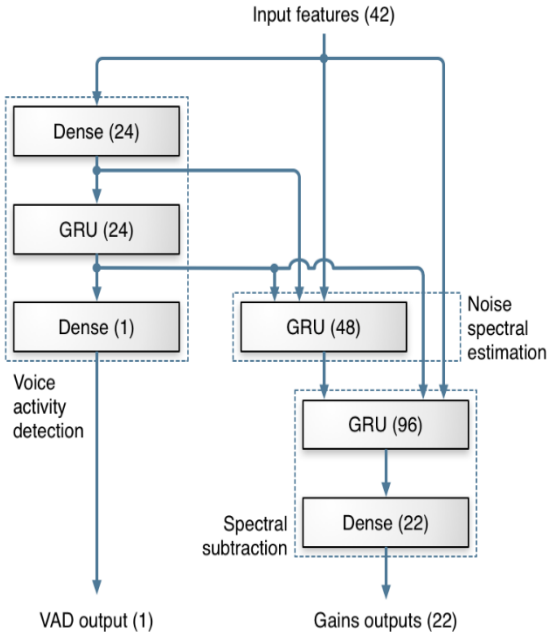

Figure 2. Topology of the RNNoise

Datasets

The study makes use of the following datasets:

a. Clean Speech Training Dataset [22] (Voicebank): This dataset comprises 23,074 samples collected from 56 different speakers, with a combined size of 6.12 GB.
b. Noise Training Dataset [22] (DEMAND): The noise training dataset consists of 272 samples obtained from 17 distinct environmental scenarios, totaling 7.29 GB.
c. Room Impulse Response (RIR) Training Dataset [23] (MIT IR Survey): This dataset contains 271 samples recorded at 14 different locations and is 16.4 MB in size.
d. Clean Speech Test Dataset [24] (JL Corpus): For the purpose of testing, 6 random samples (951 KB) are selected, featuring 3 female and 3 male voices.
e. Noise Test Dataset [25] (ESC 50): Also, for testing purposes, the noise test dataset, named ESC 50, includes 6 random samples (2.52 MB) of environmental background noises.

The carefully selected datasets are vital for conducting research experiments and evaluations. They provide a wide variety of speech and noise samples, crucial for training and evaluating the proposed models.

METHODS

The research process is outlined through a system flowchart diagram (Figure 3), simplifying the steps in studying Speech Enhancement models, specifically DeepFilterNet3 and RNNoise. It begins with model selection based on dataset characteristics and research objectives. The source code is obtained from the official repository and uploaded to Google Drive, accessible through Google Colab for development and training. The dataset is preprocessed, organized, and converted to hdf5 format before training the selected model using default hyperparameters. In order to ensure consistency and fair comparison between the models, common hyperparameters are set, such as a maximum epoch of 120 and a batch size of 32. The dataset is preprocessed, organized, and converted to hdf5 format before training the selected model using these shared hyperparameters.

During the evaluation phase, the models' effectiveness is compared using relevant performance metrics, and the default model are among those compared to identify areas that require improvement.

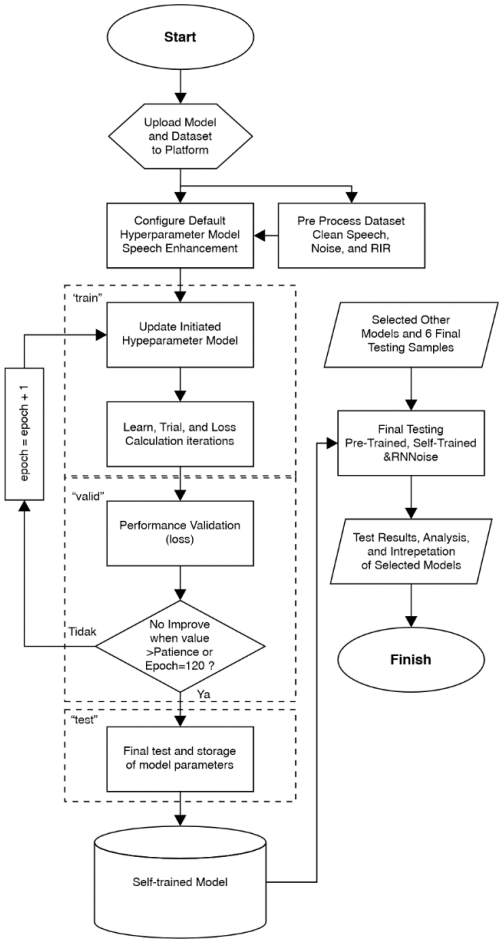

Figure 3. Research Flowchart Diagram

The research findings are then interpreted, leading to conclusive observations and recommendations for future investigations in Speech Enhancement. Throughout the research, cloud resources, including an Intel E5-2686 v4 CPU, GPU T4, and GPU V100, handle computational demands efficiently. This ensures comprehensive exploration and analysis of selected models, optimizing productivity and yielding valuable insights for the study.

To enhance overall efficiency, a block diagram (Figure 4) represents research components and their interconnected flow. The visual representation aids in identifying potential obstacles and streamlining the research process. These diagrams combination facilitates a thorough investigation of Speech Enhancement models.

RESULTS AND DISCUSSION

This chapter thoroughly discusses the chosen speech enhancement model's training process and performance metrics.

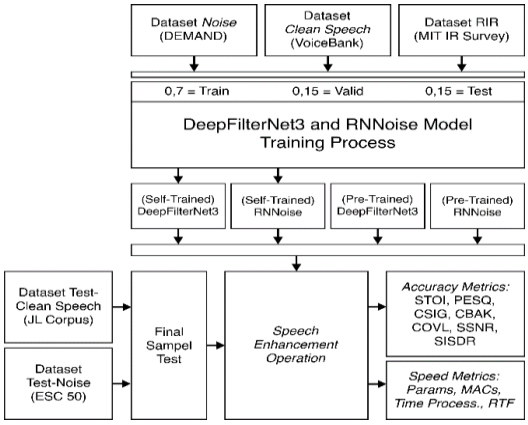

Figure 4. Research Block Diagram

The goal is to examine the training results and evaluate the models' accuracy and speed in speech enhancement operations.

This study presents a comprehensive evaluation of two speech enhancement models, DeepFilterNet3 and RNNoise, in both pre-trained and self-trained conditions. The models are assessed based on several perceptual quality and intelligibility metrics, including STOI, PESQ, CSIG, CBAK, COVL, SegSNR, and SiSDR, as well as computational performance metrics such as parameters, MACs, average sample time, processing time, and real-time factor (RTFavg). These results are compared to baseline performance before enhancement and a theoretical upper limit for each metric.

Models Training Result

In this technical comparison of training results, we assess two speech enhancement models: RNNoise and DeepFilterNet3. It's crucial to acknowledge that directly comparing their "loss" results might lead to misconceptions as their respective authors interpret these values differently. To ensure a fair evaluation using common data, both models were trained on the combined dataset of Voicebank and DEMAND. The obtained results are depicted in Figure 5.

RNNoise underwent training for 120 epochs, accomplishing this in an impressive 21 minutes, with an average epoch time of 10.8 seconds. The model achieved remarkably low loss values, ranging from 0.0014 to 0.000446118, indicating effective denoising according to its specific loss function. However, it also displayed signs of overfitting in the early epochs.

On the other hand, DeepFilterNet3 also underwent 120 epochs of training, but this process took significantly longer, lasting 15 hours and 8 minutes. Throughout the training, both the training and validation losses consistently decreased, although a slight validation loss increase was observed towards the end, indicating a potential issue with overfitting. It's crucial to remember that the interpretation of loss values is unique to DeepFilterNet3's loss function.

Despite the differing interpretations of loss, both models demonstrated improved speech quality on the common dataset. RNNoise proved to be particularly adept at denoising, while DeepFilterNet3 exhibited better generalization with a lower validation loss. For a comprehensive assessment of practical speech enhancement, it is advisable to consider additional metrics and conduct real-world testing on various datasets. This aspect will be explored in the next two sections.

Accuracy Metrics Analysis

Both Accuracy and Speed testing is done through 5 synthetic samples created using random selected sample from JL Corpus and ESC 50. Accuracy metrics result can be analyzed from the data presented in Table 1. The first model, "DeepFilterNet3 (Pre-Trained)," demonstrates promising results across various metrics. It achieves a high Short-Time Objective Intelligibility (STOI) score of 0.810, indicating that the enhanced speech is highly intelligible compared to the original. Moreover, the Perceptual Evaluation of Speech Quality (PESQ) score of 1.553 suggests that the quality of the enhanced speech is good. The model also performs well in terms of the Subjective Mean Opinion Scores (CSIG and CBAK), which measure the subjective quality of speech and background noise, respectively. Additionally, it yields respectable results in terms of the Overall Mean Opinion Score for Listening Quality (COVL) and the Segmental Signal-to-Noise Ratio (SegSNR).

On the other hand, the "DeepFilterNet3 (Self Trained)" model exhibits comparatively lower performance across all metrics. The STOI score drops to 0.653, indicating a decrease in intelligibility compared to the pre-trained version. The PESQ score of 1.145 also suggests a decline in speech quality. Furthermore, both CSIG and CBAK scores decrease, reflecting a reduction in subjective quality for speech and background noise. The Overall Mean Opinion Score for Listening Quality (COVL) and the Segmental Signal-to-Noise Ratio (SegSNR) also show a decrease, indicating a drop in the overall listening experience and increased noise interference.

Moving on to the "RNNoise (Pre-Trained)" model, we observe mixed results. While the STOI score of 0.743 indicates reasonably good intelligibility, the PESQ score of 1.233 suggests that the speech quality is acceptable but not exceptional.

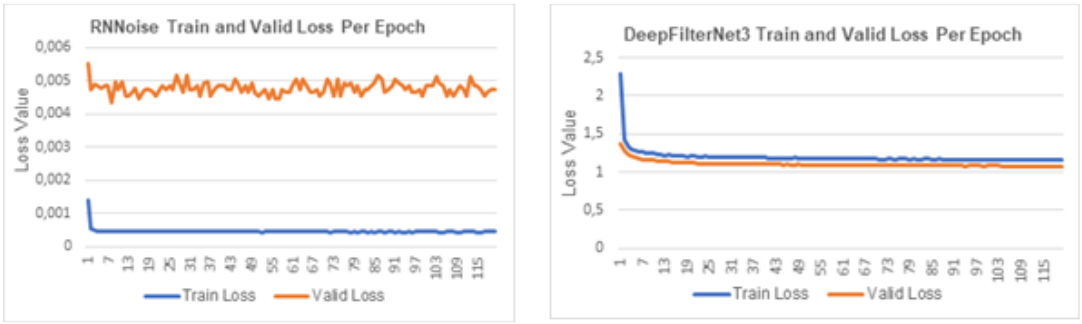

Figure. 5 RNNoise and DeepFilterNet3 Model Training Result

Table 1. Summary of Accuracy Metrics Before and After Speech Operations

Model	STOI	PESQ	CSIG	CBAK	COVL	SegSNR	SiSDR
DeepFilterNet3 (Pre-Trained)	**0.810**	**1.553**	**2.786**	**2.423**	**2.081**	**5.711**	**7.717**
DeepFilterNet3 (Self Trained)	0.653	1.145	2.167	1.870	1.565	1.511	0.545
RNNoise (Pre-Trained)	0.743	1.233	2.228	1.586	1.620	-2.800	-10.637
RNNoise (Self-Trained)	0.719	1.095	1.765	1.273	1.333	-7.401	-1.776
Before Speech Enhancement	0.727	1.101	2.001	1.613	1.444	-2.461	-0.885
Theoretical Upper Limit	1.000	4.644	5.000	5.000	5.000	35.000	80.681

However, the model's performance in terms of subjective quality (CSIG and CBAK) is still decent. Interestingly, the Segmental Signal-to-Noise Ratio (SegSNR) and the Scale-Invariant Signal-to-Distortion Ratio (SiSDR) for this model are negative, indicating that the speech enhancement process has introduced some distortion and noise to the speech signal, leading to a worse listening experience.

The "RNNoise (Self-Trained)" model also faces challenges in performance. Its STOI score of 0.719 indicates a decrease in intelligibility compared to the pre-trained version. The PESQ score of 1.095 suggests a reduction in speech quality, and the subjective quality scores (CSIG and CBAK) also show a decline. Similar to the pre-trained RNNoise model, the Segmental Signal-to-Noise Ratio (SegSNR) and the Scale-Invariant Signal-to-Distortion Ratio (SiSDR) are negative, indicating introduced distortion and noise.

Comparing the model results to the "Before Speech Enhancement" state reveals that the speech enhancement process generally improves intelligibility (STOI increases) but may not consistently improve speech quality (mixed results in PESQ). The subjective quality scores (CSIG and CBAK) vary across models, suggesting that different models may excel in enhancing certain aspects of the speech signal.

Finally, the "Upper Limit" represents an ideal performance achievable with perfect speech enhancement. The scores are significantly higher across all metrics, with maximum values for CSIG, CBAK, COVL, SegSNR, and SiSDR. This upper limit serves as a reference for the best possible performance that current models should aim to approach. Figure 6 shows how all speech enhancement model operation affects the spectrogram of a sample. In conclusion, the "DeepFilterNet3 (Pre-Trained)" model demonstrates the best overall performance among the evaluated models, achieving higher scores in most metrics. However, there is still room for improvement, as all models fall short of the ideal upper limit

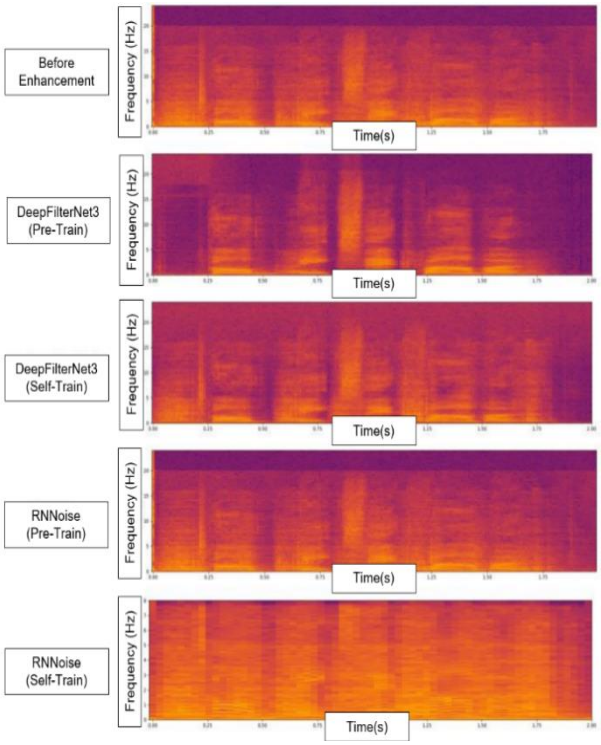

Figure 6. Qualitative Speech Enhancement Operation

Speed Metrics Analysis

Each model in this section was tested on a total of 5 different speech samples to ensure a comprehensive evaluation across various scenarios. For each speech sample, the testing process was repeated 5 times, and the times taken for each run were recorded. The recorded times were then averaged to obtain the "Average Sample Time" and "Average Process Time" metrics. This approach of averaging over multiple runs and samples helps mitigate the impact of any outliers or random variations, resulting in more reliable and representative performance metrics. Speed metrics results can be analyzed from the data presented in Table 2.

Starting with the architecture, both DeepFilterNet3 and RNNoise models seem to have a relatively small number of parameters and operations (MACS), which indicates their efficiency in terms of memory and computational requirements. The fact that they have similar values for parameters and MACS suggests that they might share some similarities in design or complexity. However, without more detailed information on the model architectures, it is challenging to make direct comparisons.

Moving on to the computational efficiency, the provided metrics show the average sample time and average process time. The average sample time, which represents the time taken to process a single sample, is consistent at 2.175 seconds for all models. This indicates that the models process each input sample in approximately the same amount of time, regardless of their complexity or the type of training.

The most critical metric for real-time applications is the Real-Time Factor (RTF_{avg}). This metric reveals how well the models perform in real-time scenarios, with values close to 1 indicating real-time processing. Both Pre-Trained and Self-Trained RNNoise models achieve remarkable RTF_{avg} values of around 0.001, which implies that they can process speech signals nearly in real-time. This is a highly desirable feature, particularly for applications that require immediate feedback, such as live communication systems or voice assistants.

On the other hand, the DeepFilterNet3 models, both Pre-Trained and Self-Trained, have slightly higher RTF_{avg} values, around 0.081 and 0.088, respectively. While these values are not as impressive as RNNoise, they still suggest that DeepFilterNet3 models can process speech in a reasonably efficient manner, being roughly 8 to 9 times slower than real-time.

Computational Efficiency

RNNoise, with just 0.060 million parameters and 0.040 MACs, has a minimal processing time of 0.003 seconds (pre-trained) and 0.002 seconds (self-trained). These figures translate to an RTFavg of 0.001, meaning that RNNoise can process audio nearly in real-time with minimal computational resources. This makes RNNoise highly suitable for large-scale deployment, where efficiency and speed are paramount. Its low processing time and resource requirements suggest it could be easily scaled across thousands of devices in real-time communication systems or online platforms, without significant strain on computational infrastructure. RNNoise is inherently more scalable due to its lightweight architecture and extremely low processing requirements. It can be deployed across a wide range of devices, including low-power mobile phones, tablets, and laptops, without overwhelming the hardware. This makes it ideal for large-scale applications such as remote classrooms, online meetings, or virtual conferences, where thousands of participants may need real-time speech enhancement. Its minimal computational footprint also allows for deployment in edge computing environments, where processing is done locally on the device, reducing the need for server-based processing and minimizing latency.

DeepFilterNet3 requires significantly more computational resources. With 2.135 million parameters and 0.340 MACs, its average processing time is 0.175 seconds for the pre-trained model, resulting in an RTFavg of 0.081.

Table 2. Summary of Speech Enhancement Speed Metrics

Model	Params (M)	MACS (G)	Average Sample Time (sec)	Average Process Time (sec)	RTF_{avg}
DeepFilterNet3 (Pre-Trained)	2.135	0.340	**2.175**	0.175	0.081
DeepFilterNet3 (Self-Trained)	2.135	0.340		0.189	0.088
RNNoise (Pre-Trained)	0.060	0.040		0.003	0.001
RNNoise (Self-Trained)	**0.060**	**0.040**		**0.002**	**0.001**

While this is still within acceptable real-time processing limits, it is much less efficient than RNNoise. The self-trained DeepFilterNet3 requires slightly more time, with an RTFavg of 0.088. These values indicate that, while DeepFilterNet3 offers superior performance in terms of speech enhancement quality, its computational demands could pose a challenge for large-scale deployment in environments where bandwidth, processing power, or battery life are limited. DeepFilterNet3, while offering better speech quality, would face challenges in terms of scalability, particularly for users with limited hardware capabilities or in resource-constrained environments. Its higher computational requirements mean that it would be more suitable for environments where performance is prioritized over resource efficiency, such as high-end workstations or dedicated audio processing servers. In large-scale deployments, it could be used in cloud-based systems where server-side processing handles the audio enhancement before streaming it to the client. However, this would introduce potential challenges such as increased latency and higher costs associated with cloud infrastructure.

To further analyze the trade-offs between these models, more information about their architectures, training data, and specific use cases is needed. Additionally, it would be beneficial to compare their speech enhancement performance in terms of objective metrics like signal-to-noise ratio (SNR) improvement and subjective evaluations with human listeners.

In conclusion, the provided deep technical analysis indicates that both RNNoise models (Pre-Trained and Self-Trained) demonstrate exceptional computational efficiency, enabling them to process speech signals almost in real-time. The DeepFilterNet3 models, while slightly slower, still exhibit reasonably efficient speech enhancement capabilities. Depending on the specific application requirements and performance considerations, choosing the most suitable model would require a more in-depth investigation and evaluation.

CONCLUSION

This comprehensive analysis of two speech enhancement models, RNNoise and DeepFilterNet3, has provided valuable insights into their training process, accuracy metrics, and speed performance. Both models demonstrated improvements in speech quality on a common dataset, with RNNoise excelling at denoising and DeepFilterNet3 exhibiting better generalization. The accuracy metrics highlighted that the "DeepFilterNet3 (Pre-Trained)" model achieved the best overall performance, with high scores in intelligibility and speech quality, while the "RNNoise (Pre-Trained)" model also showed reasonable performance in subjective quality measures. However, all models fell short of the ideal upper limit in terms of performance. On the speed front, the RNNoise models showcased exceptional computational efficiency, enabling them to process speech signals nearly in real-time, while DeepFilterNet3 models, though slightly slower, demonstrated reasonable efficiency.

ACKNOWLEDGMENT

The authors extend their heartfelt gratitude to our esteemed colleagues from the Electrical Engineering Department and Research Center of Mercu Buana University, Jakarta, Indonesia, whose unwavering support and collaboration were instrumental in successfully completing this research paper.

REFERENCES

[1] A. Hassan, S. Aftab, R. Khan, and H. Asim, “The analysis on the usage of the video conferencing rooms using classification,” *KIET Journal of Computing and Information Sciences*, vol. 2, pp. 72, 2019.

[2] G. A. Strouse et al., “Zooming through development: Using video chat to support family connections,” *Human Behavior and Emerging Technologies*, vol. 3, no. 4, pp. 552–571, Oct. 2021, doi: 10.1002/hbe2.268.

[3] R. S. Oeppen, G. Shaw, and P. A. Brennan, “Human factors recognition at virtual meetings and video conferencing: How to get the best performance from yourself and others,” *British Journal of Oral and Maxillofacial Surgery*, vol. 58, no. 6, pp. 643–646, Jul. 2020, doi: 10.1016/j.bjoms.2020.04.046.

[4] J. A. Piamonte et al., “Effects of noise sources on the perceived task performance of students during online class,” in Proc. 5th *European International Conference on Industrial Engineering and Operations Management*, Rome, Italy, Jul. 2022, pp. 2700–2709.

[5] N. Das et al., “Fundamentals, present and future perspectives of speech enhancement,” *International Journal of Speech Technology*, vol. 24, no. 4, pp. 883–901, Dec. 2021, doi: 10.1007/s10772-020-09674-2.

[6] T. M. Kadarina et al., "A simplified dental caries segmentation using Half U-Net for a teledentistry system," *SINERGI*, vol. 28, no. 2, pp. 251-258, 2024, doi: 10.22441/sinergi.2024.2.005.

[7] K. Aziz et al., "Multilabel image analysis on Polyethylene Terephthalate bottle images using PETNet convolution architecture," *SINERGI*, vol. 27, no. 2, pp. 163-170, 2023, doi: 10.22441/sinergi.2023.2.003.
[8] D. Michelsanti et al., “An overview of deep-learning-based audio-visual speech enhancement and separation,” *IEEE/ACM IEEE Transactions on Audio, Speech and Language Processing (TASLPRO)*, vol. 29, pp. 1368–1396, Mar. 2021, doi: 10.1109/TASLP.2021.3066305.
[9] H. Schröter et al., “DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, pp. 7407–7411, doi: 10.1109/ICASSP43922.2022.9747666.
[10] D. Yin et al., “PHASEN: A phase-and-harmonics-aware speech enhancement network,” in Proc. AAAI Conference on Artificial Intelligence, 2020, pp. 9458–9465.
[11] J.-M. Valin et al., “A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech,” in *Proc. Interspeech*, Oct. 2020, pp. 2482–2486, doi: 10.21437/Interspeech.2020-2730.
[12] K. Tan and D. L. Wang, “Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 6865–6869, doi: 10.1109/ICASSP.2019.8682834.
[13] J.-M. Valin, “A hybrid DSP/deep learning approach to real-time full-band speech enhancement,” in *Proc. International Workshop on Multimedia Signal Processing (MMSP)*, Aug. 2018, pp. 1–5, doi: 10.1109/MMSP.2018.8547094.
[14] Y. Hu et al., “DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement,” in *Proc. Interspeech*, Oct. 2020, pp. 2472–2476, doi: 10.21437/Interspeech.2020-1558.
[15] S. Braun et al., “Towards efficient models for real-time deep noise suppression,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Jun. 2021, pp. 656–660, doi: 10.1109/ICASSP39728.2021.9413908.
[16] J. Chen et al., “FullSubNet+: Channel attention FullSubNet with complex spectrograms for speech enhancement,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2022, pp. 7857–7861, doi: 10.1109/ICASSP43922.2022.9746476.
[17] A. Li et al., “Glance and gaze: A collaborative learning framework for single-channel speech enhancement,” *Applied Acoustics*, vol. 187, p. 108499, Jun. 2022, doi: 10.1016/j.apacoust.2021.108499.
[18] S. Zhao et al., “FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2022, pp. 9281–9285, doi: 10.1109/ICASSP43922.2022.9746546.
[19] H. Schröter et al., “DeepFilterNet2: Towards real-time speech enhancement on embedded devices for full-band audio,” in 022 International Workshop on Acoustic Signal Enhancement (IWAENC), Sep. 2022, pp. 1–5, doi: 10.1109/IWAENC53105.2022.9914714.
[20] H. Schröter et al., “DeepFilterNet: Perceptually motivated real-time speech enhancement,” in *Proc. Interspeech*, Aug. 2023.
[21] H. Schröter et al., “CLCNet: Deep learning-based noise reduction for hearing aids using complex linear coding,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 6949–6953, doi: 10.1109/ICASSP40776.2020.9053144.
[22] C. Valentini-Botinhao, “Noisy speech database for training speech enhancement algorithms and TTS models,” Univ. Edinburgh, *Technical Reports*, 2016.
[23] J. Traer and J. H. McDermott, “Statistics of natural reverberation enable perceptual separation of sound and space,” *Proc. Natl. Acad. Sci. U.S.A.,* vol. 113, no. 48, pp. 7856–7865, 2016, doi: 10.1073/pnas.1612524113.
[24] J. James, L. Tian, and C. I. Watson, “An open-source emotional speech corpus for human robot interaction applications,” in *Proc. Interspeech*, 2018, pp. 2768–2772, doi: 10.21437/Interspeech.2018-1349.
[25] K. J. Piczak, “ESC: Dataset for environmental sound classification,” in *Proc. ACM Multimedia Conference*, 2015, pp. 1015–1018, doi: 10.1145/2733373.2806390.
[26] J. Kim et al., “End-to-end multi-task denoising for joint SDR and PESQ optimization,” in *Proc. Interspeech*, 2021, pp. 1564–1568, doi: 10.21437/Interspeech.2021-1763.
[27] R. E. Zezario et al., “STOI-Net: A deep learning based non-intrusive speech intelligibility assessment model,” in *Proc. Asia-Pacific Signal and Information*

Processing Association (APSIPA), 2020, pp. 482–486.
[28] P. Krishnamoorthy, “An overview of subjective and objective quality measures for noisy speech enhancement algorithms,” *IETE Technical Review*, vol. 28, no. 4, pp. 292–301, 2011, doi: 10.4103/0256-4602.83550.
[29] R. Kashyap, R. H. Laskar, and N. Shome, “Non-negative frequency-weighted energy-based speech quality estimation for different modes and quality of speech,” *Circuits, Systems, and Signal Processing*, vol. 41, pp. 6788–6826, 2022.
[30] J. Le Roux et al., “SDR – Half-baked or well done?,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 626–630, doi: 10.1109/ICASSP.2019.8683635.
[31] S. Badillo et al., “An introduction to machine learning,” *Clinical Pharmacology & Therapeutics*, vol. 107, no. 4, pp. 871–885, 2020, doi: 10.1002/cpt.1796.
[32] M. A. Nahmias et al., “Photonic multiply-accumulate operations for neural networks,” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–18, 2019, doi: 10.1109/JSTQE.2019.2941485.
[33] T. Saeki et al., “Real-time, full-band, online DNN-based voice conversion system using a single CPU,” in *Proc. Interspeech*, 2020, pp. 1021–1022.