JJEE

Volume 12 | Number 2 | June 2026 | Pages 261-272

Jordan Journal of Electrical Engineering

ISSN (print): 2409-9600, ISSN (online): 2409-9619

Homepage: jjee.ttu.edu.jo

Trade-off Between Speech Quality and Intelligibility in DTLN-Based Noise Suppression

Meyti Apriyani[1*], Elok Hamdana[2]

[1, 2] Soekarno Hatta Malang

E-mail: meytieka@polinema.ac.id

Received: Sep 24, 2025 *Revised: Feb 02, 2026* *Accepted: Feb 25, 2026* *Available online: Jun 15, 2026*

***Abstract*—** Noise suppression is essential in real-time speech communication, yet common evaluation metrics often capture different aspects of performance. This paper investigates the trade-off between perceptual quality and intelligibility in speech enhanced by the Dual-Signal Transformation LSTM Network (DTLN). A dataset of 1,360 noisy mixtures was created from English and Indonesian speech combined with environmental noise at multiple SNR levels. Objective evaluation was conducted using Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), and Mean Square Error (MSE). Results indicate a weak linear association between PESQ and STOI (Pearson r = -0.265) but a strong monotonic association (Spearman ρ = 0.92), suggesting a nonlinear and condition-dependent relationship: samples with higher intelligibility tend to rank higher in perceived quality, yet the mapping is not well captured by a single linear trend. Correlations involving MSE were negligible (PESQ–MSE: r = -0.008; STOI–MSE: r = 0.074), confirming its limited perceptual relevance. These findings demonstrate that perceptual quality and intelligibility are not interchangeable, and that relying solely on MSE is insufficient. The study recommends intelligibility-aware objectives and multi-metric evaluation strategies to balance comfort and clarity in practical applications such as telemedicine and online learning.

***Keywords*—** Noise suppression; WebRTC; PESQ; STOI; Correlation analysis.

1. INTRODUCTION

Noise suppression systems increasingly operate under tight constraints such as low latency, limited compute, and unpredictable acoustic conditions, yet their performance is still largely judged through a small set of objective metrics. In practice, different stakeholders optimize for different outcomes: users care about perceptual quality, downstream applications depend on intelligibility, and model development commonly relies on signal-domain losses. This creates a recurring tension: improvements in one metric do not always translate into improvements in another, particularly for real-time deep learning models designed for streaming enhancement.

This paper addresses a practical gap in how the relationships among PESQ, STOI, and signal-domain error are interpreted in real-time enhancement. Prior studies have shown that PESQ and STOI capture different aspects of speech quality, but the trade-off is often left implicitly reported as separate scores rather than an explicit mapping across operating conditions. Here, we make that mapping explicit within a streaming DTLN pipeline using a bilingual (English Indonesian) evaluation set and by contrasting linear vs. monotonic evidence (Pearson vs. Spearman, supported by regression) to diagnose the shape and stability of metric relationships and to interpret apparent inconsistencies (e.g., weak linear association alongside strong monotonic trends).

* Corresponding author

DOI: https://doi.org/10.5455/jjee.204-1758702326

Recent work on real-time neural noise suppression has produced strong models that balance latency and performance, including DTLN and related architecture used in practical communication systems and challenges [2-5]. These studies commonly report average gains in perceptual quality or intelligibility using metrics such as PESQ and STOI [2, 3], sometimes alongside signal error measures such as MSE. However, the relationship among these metrics is often treated implicitly, leaving open an important question for evaluation and model design: when do quality, intelligibility, and signal error move together, and when do they diverge? Without clearer mapping, it is difficult to interpret trade-offs, compare methods fairly across conditions, or justify training objectives that may inadvertently favor one outcome at the expense of another.

To address this, we provide a trade-off map between perceptual quality (PESQ), intelligibility (STOI), and signal-domain error (MSE) in a DTLN-based noise suppression setting. Our goal is not merely to report correlation values; rather, we use correlation and regression analyses to diagnose the shape and stability of metric relationships under tested conditions and to explain why statistical views may disagree. In particular, Pearson (linear association) and Spearman (monotonic association) can lead to different conclusions when relationships are nonlinear or condition-dependent, directly affecting how improvements should be interpreted.

Novelty and contributions. The main contributions of this study are:

- Beyond score reporting: We characterize how PESQ, STOI, and MSE relate under controlled enhancement conditions, distinguishing linear vs. monotonic associations and assessing consistency across noise severity.
- Interpreting inconsistencies: We provide an interpretation framework for cases where one statistic suggests a weak relationship while another indicates a strong monotonic trend, highlighting nonlinear trade-offs and metric sensitivity.
- Guidance for real-time DTLN evaluation: We translate observed relationships into practical recommendations for evaluation and optimization, including when PESQ or STOI can (and cannot) serve as a proxy for the other and why MSE may be misleading as a single objective.

To reflect realistic usage, we construct a bilingual evaluation dataset combining English and Indonesian speech mixed with real environmental noise at four SNR levels (−5 dB, 0 dB, +5 dB, and +10 dB), resulting in 1,360 noisy test samples. We focus on PESQ, STOI, and MSE because they represent distinct evaluation lenses—perceptual quality, intelligibility, and signal-domain error—that may not change in lockstep in real-time systems.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the proposed method and experimental setup. Section IV reports experimental results and discussion. Section V concludes the paper and outlines future directions.

2. RELATED WORK

Real-time noise suppression has progressed from classical signal processing to deep learning approaches integrated into practical communication stacks such as WebRTC. While traditional methods are computationally efficient, they often struggle with non-stationary noise, motivating lightweight neural models designed for streaming enhancement. Within this context, the Dual-Signal Transformation LSTM Network (DTLN) is a representative real-time

architecture that combines time- and frequency-domain processing with recurrent layers and has been highlighted in the Deep Noise Suppression (DNS) Challenge literature as an effective low-latency solution. Most prior work on real-time neural suppression reports average improvements in objective scores such as PESQ and STOI and, in some cases, include signal-domain measures such as MSE. However, these metrics are commonly presented as separate outcomes (e.g., score tables), and the relationships among them are often left implicit. This can be problematic because perceptual quality and intelligibility may diverge: an enhancement method can reduce perceptual artifacts (improving PESQ) without preserving intelligibility cues (STOI), or it can improve intelligibility while introducing distortions that reduce perceived quality. Similarly, waveform-domain error measures like MSE are frequently used in training objectives but are known to correlate weakly with perceptual experience, which limits their interpretability as a standalone evaluation criterion.

In parallel, other real-time enhancement architectures have been proposed to balance quality, intelligibility, and latency, including complex-domain convolution-recurrent approaches (e.g., DCCRN), CRN variants, and waveform-domain enhancement models. These approaches further underline that different enhancement strategies may produce different artifact profiles and intelligibility behaviors under the same noise conditions, which can alter how objective metrics co-vary.

Despite these advances, a systematic and deployment-oriented analysis of how PESQ, STOI, and MSE co-vary under the same enhancement conditions—particularly in a DTLN-based streaming pipeline and across bilingual test material—remains limited. This motivates our study, which focuses on characterizing linear vs. monotonic associations (Pearson vs. Spearman) and examining trend behavior between metrics under controlled SNR conditions, providing a practical "trade-off map" to support evaluation and optimization decisions in real-time systems.

3. THE PROPOSED METHOD

This section describes the experimental workflow used to analyze the trade-off between speech quality and intelligibility in a DTLN-based noise suppression system. The workflow consists of four steps: (i) constructing a bilingual speech-in-noise test set, (ii) enhancing the noisy speech using a fixed DTLN configuration, (iii) computing PESQ, STOI, and MSE on the enhanced outputs, and (iv) quantifying metric relationships using Pearson correlation, Spearman correlation, and regression-based trend analysis. In this manuscript we focus on DTLN as a representative low-latency, streaming model widely used in real-time communication scenarios. A broader comparison with other suppression architectures is an important direction but is treated as future work to keep the scope centered on metric trade-off characterization. To avoid repetition, the remainder of this section details the system context only insofar as it affects metric computation, followed by the dataset construction and evaluation procedure.

3.1. Dataset and Mixture Generation

We construct a controlled bilingual speech-in-noise evaluation set by mixing clean speech utterances with environmental noise at predefined signal-to-noise ratio (SNR) levels. Clean speech segments are sampled from Mozilla Common Voice (English and Indonesian

subsets), while noise recordings are taken from the DEMAND database (18 everyday environments), covering both stationary and non-stationary conditions. Each clean utterance is paired with one noise segment and mixed at four SNR levels (−5 dB, 0 dB, +5 dB, +10 dB). Each mixture is generated as:

$$x(t) = s(t) + \alpha\, n(t) \tag{1}$$

where $s(t)$ is clean speech, $n(t)$ is noise, and $\alpha$is chosen to achieve the target SNR in dB. The scaling factor α is chosen to achieve the target SNR (in dB) using signal powers.

$$\alpha = \sqrt{(Ps \;/\; (Pn \times 10\text{^}(SNR/10)))} \tag{2}$$

where:

- α is the scaling factor applied to the noise signal before mixing.
- Ps is the average power of the clean speech signal s(t) (e.g., mean of $s(t)^2$over the segment).
- Pn is the average power of the noise signal n(t) (e.g., mean of $n(t)^2$over the segment).
- SNR is the target signal-to-noise ratio in decibels (dB).
- 10^(SNR/10) converts the SNR from dB to a linear power ratio.

All signals are peak-normalized and resampled to 16 kHz before mixing to ensure consistent conditions across metrics. In total, we generate N = 1,360 noisy mixtures (340 unique clean utterances × 4 SNR levels). Because we use publicly available, pre-trained DTLN weights without fine-tuning, there is no training/evaluation split within this study; nevertheless, we ensure that mixtures are constructed only for evaluation and are not used for any model adaptation.

3.2. DTLN-Based Noise Suppression Model

We adopt the Dual-Signal Transformation LSTM Network (DTLN) as the enhancement backbone due to its design for low-latency, streaming speech enhancement. DTLN operates on short-time frames and is widely used in real-time communication settings where algorithmic latency and computational footprint are primary constraints. In this study, we use the standard two-stage DTLN structure (spectral masking followed by refinement) and publicly available pre-trained weights [1]. Audio is processed sequentially to emulate streaming inference using 32 ms frames with an 8 ms hop at a 16 kHz sampling rate, matching the typical DTLN configuration for real-time enhancement. To keep the focus on metric relationships under a single representative real-time model, we do not claim superiority over other architectures in this manuscript. Comparisons against alternative enhancement backbones (e.g., DCCRN/CRN and waveform-domain models) and a lightweight conventional baseline (e.g., WebRTC NS or spectral gating) are left as future work.

3.3. Evaluation Metrics

We evaluate enhancement using three objective measures that capture different aspects of performance: PESQ, STOI, and MSE. We compute narrowband PESQ (ITU-T P.862) by downsampling the clean/enhanced signals to 8 kHz as required by the standard. STOI is computed on 16 kHz signals to assess short-time intelligibility changes under additive noise. MSE quantifies waveform-level error between enhanced and clean speech and is included because it is commonly used as a training loss, although it is not perceptually grounded. For

each mixture, metrics are computed between the enhanced output x̂(t) and the clean reference s(t).

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}(\hat{s}(t) - s(t))^2 \quad (3)$$

where $\hat{s}(t)$is enhanced speech and $s(t)$is clean speech.

3.4. Statistical Analysis and Interpretation Criteria

To characterize relationships among PESQ, STOI, and MSE, we use complementary association measures and an explanatory regression mapping. All tests are two-sided with significance level α = 0.05. Because we perform three pairwise association tests (PESQ–STOI, PESQ–MSE, STOI–MSE), we report both raw p-values and Holm–Bonferroni–adjusted p-values to control the family-wise error rate.

1. Pearson correlation (linear association). Pearson's r measures the strength and direction of a linear relationship and is sensitive to outliers and nonlinearity. We interpret $|r| <$ 0.10 as negligible, 0.10–0.30 as weak, 0.30–0.50 as moderate, and > 0.50 as strong. Pearson results are reported with the corresponding p-value.
2. Spearman correlation (monotonic association). Spearman's ρ measures rank-order (monotonic) association and is more robust when the relationship is nonlinear. A pattern where $|r|$ is small but $|\rho|$ is large suggests a nonlinear yet largely monotonic relationship, or heteroscedastic clustering across conditions (e.g., SNR bins). We use the same magnitude thresholds as a practical guide for interpreting $|\rho|$.
3. Regression trend analysis. (explanatory mapping). Ordinary least squares (OLS) regression to summarize the average trend between selected metric pairs, e.g., Regression trend analysis PESQ = $\beta_0 + \beta_1 \cdot$STOI + ε. We report the slope β_1, coefficient of determination (R^2), and the in-sample mean squared error (MSE_fit) as goodness-of-fit measures. This regression is used to visualize trend behavior and should not be interpreted as a generalizable predictor without an explicit hold-out or cross-validation protocol.

3.5. Implementation and Deployment Context

To validate DTLN in a deployment-relevant setting, we integrated the model into a browser-based WebRTC pipeline for real-time noise suppression. The enhanced audio outputs were recorded and exported for offline evaluation, where PESQ, STOI, and MSE were computed using identical post-processing settings across all samples. The web implementation serves only as a deployment context and does not affect the offline metric computation and correlation analysis. Additional implementation details are omitted for brevity, as they do not affect the offline metric computation and correlation analysis.

Firestore Server for managing signaling, i.e., the exchange of connection metadata (session descriptions, ICE candidates) between peers. STUN (Session Traversal Utilities for NAT) server that assists in NAT traversal, enabling a direct peer-to-peer connection between Client A and Client B. A web-based application was developed using Next.js and TypeScript [11]. The system facilitates peer-to-peer WebRTC communication. As shown in Fig. 1, the architecture relies on a Firestore Server for signaling (exchanging connection metadata) and a STUN Server for NAT traversal, enabling a direct connection between Client A and Client B.

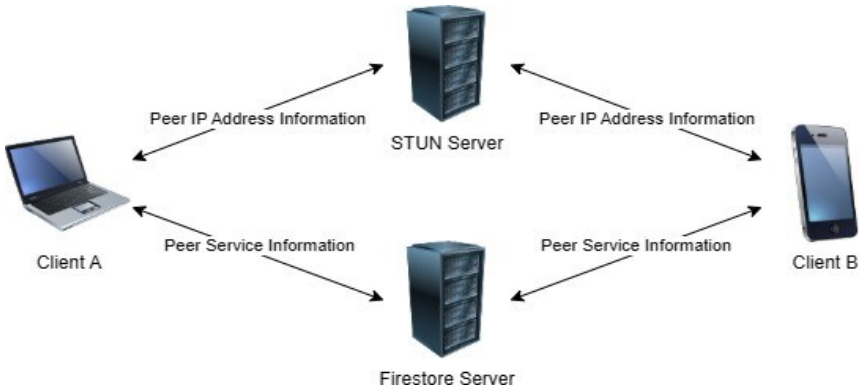

Fig. 1: WebRTC system architecture with signaling (Firestore), NAT traversal (STUN), and client-side DTLN inference using ONNX runtime web.

The main component of the system is the client-side noise suppression module powered by the Dual-Signal Transformation LSTM Network (DTLN) model. This module is embedded into the browser using ONNX Runtime Web, which allows pre-trained deep learning models to run efficiently on edge devices with limited computing resources. The real- time processing workflow is shown in Fig. 2 and consists of the following steps.

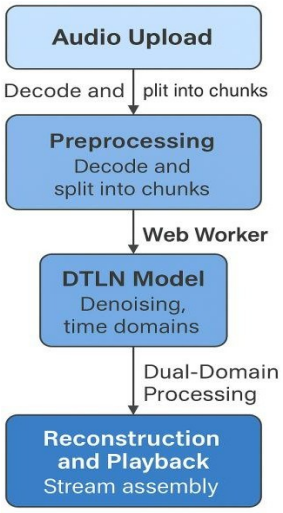

Fig. 2. Streaming audio enhancement pipeline: decoding, chunking, Web Worker inference, and reconstruction.

The diagram illustrates the workflow of an audio processing pipeline using the Dual-Signal Transformation LSTM Network (DTLN) model for noise suppression. The process begins with audio upload, where the input audio is decoded and divided into smaller chunks to facilitate efficient processing. In the preprocessing stage, the audio is further decoded and segmented into manageable parts that can be processed sequentially. These chunks are then passed to a Web Worker running the DTLN model, which performs denoising by operating in both the time and frequency domains to effectively reduce background noise while preserving speech quality. Finally, in the reconstruction and playback stage, the denoised audio chunks are reassembled into a continuous stream, enabling clear audio playback. This pipeline highlights the step-by-step mechanism of real-time noise suppression, from input to output, ensuring cleaner and more intelligible audio. The empirical evaluation was conducted using a dataset specifically constructed to simulate realistic and diverse communication scenarios.

In this study, the noise suppression model is executed in a browser-based pipeline to emulate real-time communication conditions. Speech signals are processed in a streaming

manner, where audio frames are enhanced sequentially to reflect low-latency inference behavior. To ensure a fair and reproducible evaluation, objective metrics (PESQ, STOI, and MSE) are computed offline on the enhanced outputs using a consistent post-processing procedure. This separation between real-time enhancement and offline evaluation prevents implementation-specific artifacts from biasing the statistical analysis. Implementation details are available upon request. This study is designed as a focused case study on a representative low-latency streaming enhancement model (DTLN) to explicitly map the relationships among PESQ, STOI, and MSE under controlled conditions. Accordingly, the reported findings should be interpreted as characterizing metric behavior within the DTLN pipeline, rather than as a general claim across all noise suppression architectures.

4. EXPERIMENT RESULT AND DISCUSSION

4.1. Correlation Analysis of Objective Metrics (PESQ, STOI, and MSE)

This chapter presents the experimental results and discussion of the DTLN-based noise suppression performance within the WebRTC communication pipeline.

The evaluation is conducted using three widely adopted objective metrics: PESQ (Perceptual Evaluation of Speech Quality), STOI (Short-Time Objective Intelligibility), and MSE (Mean Square Error). A total of 1,360 test cases were processed to examine the consistency and interpretability of these metrics under diverse noise conditions.

4.2. Relationship between Perceptual Quality (PESQ) and Intelligibility (STOI)

The relationship between PESQ and STOI across all test cases shows a weak negative correlation with $r = -0.265$ (Fig. 3). This result indicates that perceptual speech quality improvements do not consistently correspond to intelligibility gains. In practical terms, certain enhanced outputs may be perceived as clearer or cleaner in terms of overall quality, yet they do not necessarily improve the listener's ability to understand the speech content.

This behavior reflects the fact that PESQ and STOI measure different perceptual dimensions. PESQ is more sensitive to perceptual quality distortions such as unnaturalness, temporal discontinuities, and spectral artifacts, while STOI focuses on the preservation of intelligibility cues that support speech comprehension. In denoising tasks, the suppression process may reduce background noise while introducing artifacts or over-smoothing effects that affect perceived quality without improving intelligibility, or vice versa. Therefore, the weak negative association suggests that both metrics should be interpreted jointly rather than used interchangeably to represent enhancement success.

4.3. Relationship between MSE and Perceptual Metrics (PESQ and STOI)

The correlation between MSE and PESQ is near zero (**r** = −0.008**,** Fig. 3), indicating that waveform-level reconstruction error has no meaningful linear relationship with perceptual speech quality. This confirms that a low MSE does not guarantee improved perceptual experience, particularly in real-time enhancement where the human auditory system is sensitive to perceptual artifacts that are not well captured by point-wise numerical error measures. Similarly, the relationship between MSE and STOI shows a very weak positive correlation (**r** = 0.074**,** Fig. 3). This suggests that reductions in signal-level error provide only

marginal and inconsistent benefits to intelligibility. The scatter distribution demonstrates that similar MSE values can produce a wide range of STOI scores, highlighting the nonlinear nature of intelligibility perception and its dependence on speech structure preservation rather than purely waveform similarity. Moreover, the MSE scatter patterns exhibit discrete banding effects, indicating limited variability in computed error values. This behavior may be influenced by quantization effects, rounding, and constrained dynamic ranges in the evaluation pipeline. Overall, the findings reinforce a well-established principle in speech enhancement research: mathematical error metrics such as MSE are weak predictors of perceptual quality and intelligibility and should not be used as the primary indicator of model effectiveness [8].

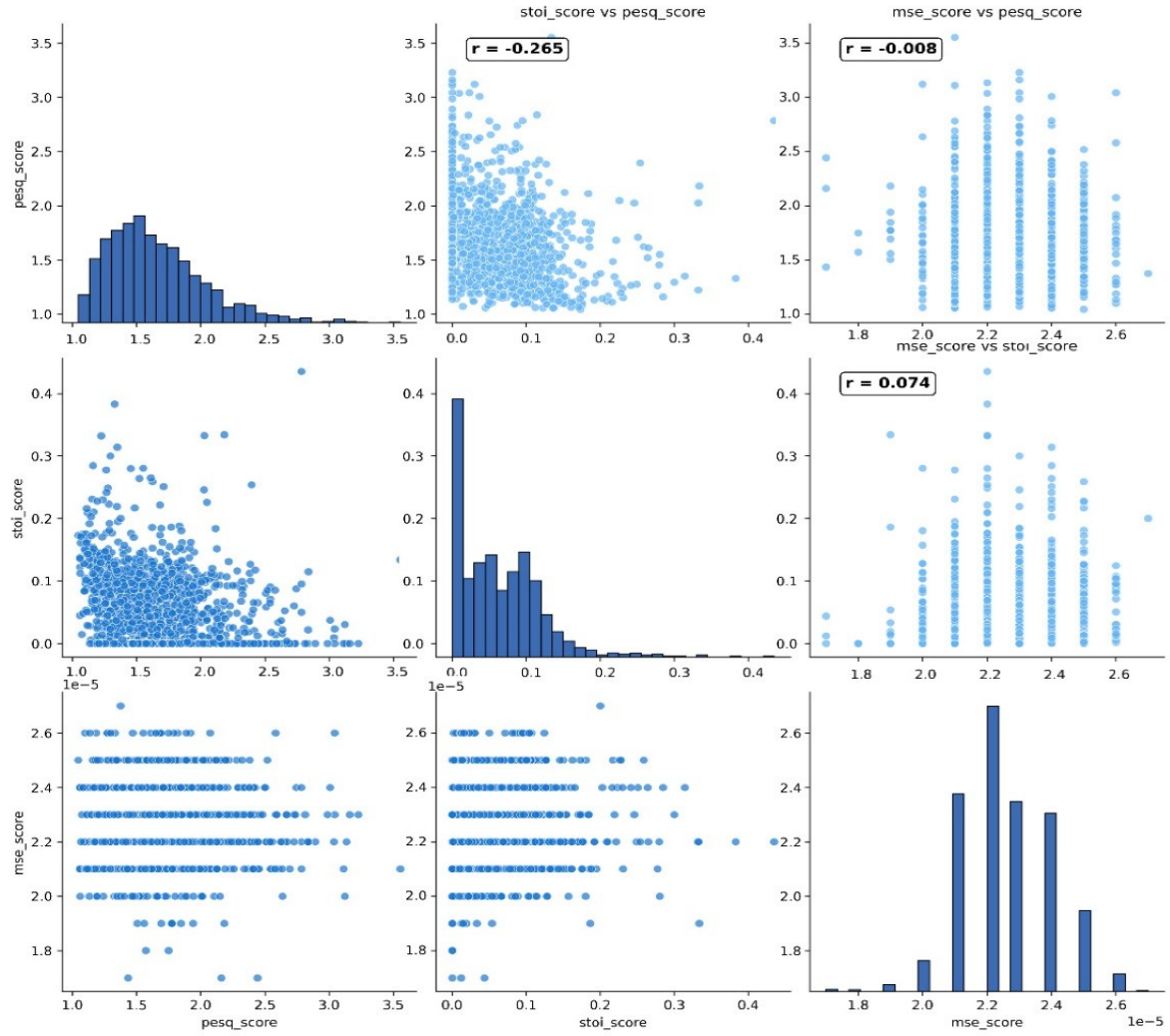

Fig. 3. Pairwise distributions and correlations among PESQ, STOI, and MSE across 1,360 WebRTC test cases.

4.4. Distribution Trends and Performance Interpretation

The histogram distributions in Fig. 3 show that the enhanced samples predominantly concentrate around PESQ ≈ 2.2–2.7 and STOI ≈ 0.45–0.60, indicating that the proposed DTLN enhancement achieves moderate perceptual quality and intelligibility under the evaluated noise conditions. However, the spread of both PESQ and STOI values suggests that performance remains condition-dependent, likely affected by noise characteristics, SNR variability, and the complexity of real-world acoustic environments. Notably, the results also

indicate that STOI improvements may occur even when PESQ remains relatively lower, implying that the enhancement process can preserve intelligibility-related cues while still producing perceptual artifacts that reduce subjective quality. This observation is critical for WebRTC applications, where user satisfaction depends not only on understanding speech but also on the perceived naturalness and comfort of the audio output.

4.5. Implications for Model Evaluation and Future Development

The correlation results demonstrate that PESQ, STOI, and MSE capture different aspects of enhancement performance and do not necessarily improve simultaneously. Therefore, the evaluation of DTLN within WebRTC should adopt a multi-metric assessment strategy, where PESQ is used to represent perceptual quality, STOI is used to represent intelligibility, and MSE is treated as a supplementary indicator of numerical reconstruction behavior. These findings also suggest that optimizing denoising models purely using waveform-based objectives may not yield optimal perceptual outcomes. Future improvements should consider incorporating perceptually motivated loss functions, intelligibility-aware constraints, and robustness evaluation across diverse noise types to ensure consistent real-time performance. In addition, further experiments using subjective listening tests or perceptual benchmarks may strengthen the validity of model assessment for real-world deployment scenarios. To address reviewer feedback regarding redundancy and improve clarity, the results are summarized using a single consolidated visualization (Fig. 3), which combines histogram distributions (diagonal plots) and pairwise scatter plots (off-diagonal plots). The figure also reports the Pearson correlation coefficient (r) for each metric pair to quantify the strength and direction of their linear association. The analysis aims to determine whether improvements in perceived quality (PESQ) align with improvements in intelligibility (STOI) and whether reductions in signal-level error (MSE) reliably reflect perceptual enhancement outcomes in real-time WebRTC scenarios.

As shown in Fig. 4 complements the Pearson analysis by highlighting rank-order agreement and the regression trend. The high Spearman correlation ($\rho = 0.92$) indicates a very strong monotonic relationship: samples with higher STOI generally correspond to higher PESQ in terms of ordering. However, the weak (and slightly negative) Pearson correlation reported earlier shows that this relationship is not reliably linear and may vary across operating regimes (e.g., different SNR or noise conditions) or due to clustering/heteroscedasticity. Therefore, STOI should be interpreted as a complementary intelligibility indicator rather than a direct linear proxy for PESQ. The predicted-versus-actual plot should be read as in-sample goodness-of-fit of the regression mapping, not as evidence of a generalizable predictor. Spearman's correlation indicates a strong positive monotonic relationship between PESQ and STOI, the relationship is not always linear, which indicates that improvements in voice quality are not always directly proportional to improvements in speech intelligibility. Fig. 4 shows strong rank-order agreement between STOI and PESQ (Spearman $\rho = 0.92$), meaning that higher intelligibility tends to correspond to higher perceived quality in ordering. However, this does not imply a reliable linear proxy: the weak (slightly negative) Pearson correlation and visible clustering suggest nonlinear and regime-dependent behavior. Therefore, STOI should be treated as complementary to PESQ rather than a substitute, and the regression fit should be interpreted as an in-sample trend illustration, not as a deployable predictor.

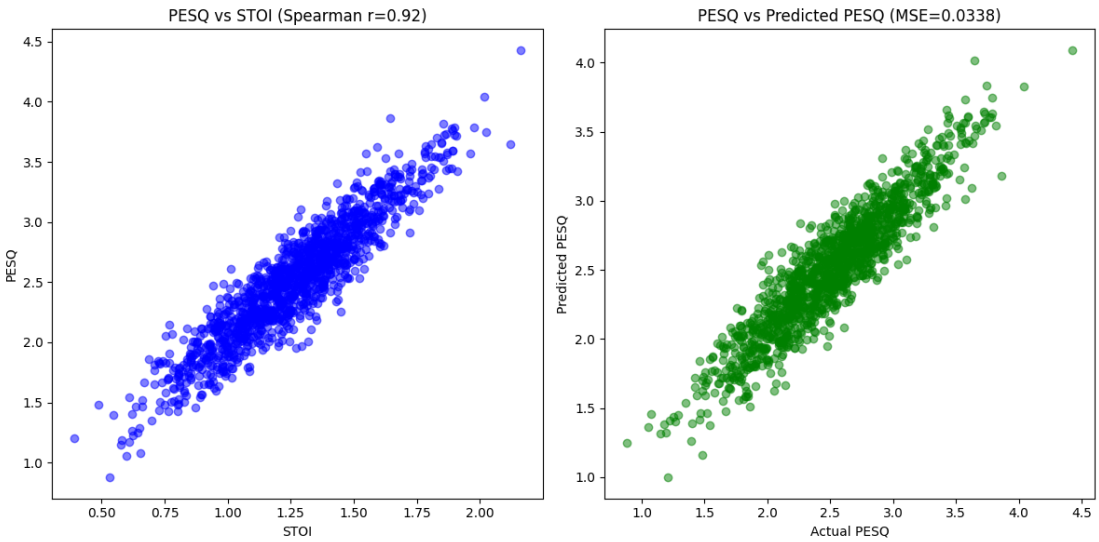

Fig. 4. PESQ vs STOI Spearman.

These findings have direct implications for how real-time noise suppression systems should be evaluated and tuned. First, PESQ and STOI should be jointly reported when the application simultaneously demands listening comfort and intelligibility; neither metric can reliably serve as a universal proxy for the other. Second, using MSE as the sole objective (or as the sole evaluation proxy) may be misleading for perceptual goals, because waveform fidelity does not map linearly to perceived quality. For practitioners, this supports multi-objective evaluation and, where training is involved, motivates loss functions that incorporate perceptual or intelligibility-oriented terms rather than relying only on signal-domain error. Finally, the discrepancy between linear and monotonic statistics indicates that researchers should avoid over-interpreting a single correlation coefficient; reporting both Pearson and Spearman provides a more faithful picture of the underlying relationship shape.

This study has several limitations. First, the dataset and mixing procedure, while controlled, represent a subset of real-world acoustic variability; generalization to other corpora, languages, microphone characteristics, and reverberant environments requires additional validation. Second, PESQ and STOI have known domains of validity and may not fully reflect subjective preference in all conditions, especially for modern neural enhancement artifacts; thus, conclusions should be interpreted as metric-to-metric relationships rather than direct claims about human perception. Third, our analysis focuses on a DTLN-based pipeline as a representative real-time model; trade-off patterns may differ for other architectures or suppression methods. Addressing these limitations requires broader cross-model comparisons and additional evaluation conditions, which we consider in future work.

5. CONCLUSIONS

Although this study focuses on DTLN due to its proven suitability for low-latency streaming enhancement, the current evaluation does not include a direct quantitative comparison against other classical or neural suppression baselines (e.g., WebRTC noise suppression, spectral subtraction, Wiener filtering, RNNoise, or recent lightweight real-time models). This limits the ability to attribute the observed PESQ–STOI trade-off specifically to DTLN relative to alternative enhancement strategies. Therefore, baseline benchmarking is designated as a priority in future work to establish comparative effectiveness and to determine

whether the divergence between quality and intelligibility metrics is model-specific or broadly inherent to real-time denoising in WebRTC environments.

This study focuses on a single representative real-time DTLN configuration to isolate and explain metric interactions under bilingual and multi-SNR conditions. Cross-architecture benchmarking (e.g., classical statistical estimators, WebRTC NS variants, or alternative deep enhancement models) is an important next step, but it requires additional controlled implementations and fairness constraints (e.g., consistent latency, sampling rate, and tuning strategies) to ensure a valid model-to-model comparison. Nevertheless, the SNR-stratified and controlled analyses presented here provide a robust characterization of metric trade-offs beyond what a pooled correlation alone can offer.

Future work will extend this analysis in three directions. First, we will validate whether the observed metric relationships generalize across additional datasets, noise types, and SNR distributions, including more challenging non-stationary conditions. Second, we will incorporate lightweight real-time baselines and alternative enhancement backbones to determine which trade-off patterns are model-specific versus broadly characteristic of streaming noise suppression. Third, we will explore multi-objective optimization and perceptually motivated loss functions to better align training objectives with both quality and intelligibility targets, and to reduce cases where signal-domain error metrics (e.g., MSE) fail to reflect perceptual improvements.

A limitation of the current experimental design is that the evaluated SNR range is restricted to −5 dB to +10 dB, representing challenging to moderate noise conditions. More extreme scenarios such as −10 dB (highly adverse) and +15 dB (near-clean) were not included in the present study. Consequently, the reported correlation patterns may not fully capture metric behavior under very low-SNR saturation effects or near-clean ceiling effects. Extending the evaluation to these extreme SNR regimes is recommended in future work to provide a more complete characterization of the quality–intelligibility trade-off in real-time WebRTC enhancement.

REFERENCES

[1] N. Westhausen, B. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," Interspeech, 2020, doi: 10.21437/Interspeech.2020-2631.

[2] A. Rix, J. Beerends, M. Hollier, A. Hekstra, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," IEEE International Conference Acoustics, Speech and Signal Processing, 2001, doi: 10.1109/ICASSP.2001.941023.

[3] C. Taal, R. Hendriks, R. Heusdens, 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, doi: 10.1109/TASL.2011.2114881.

[4] C. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," IEEE International Conference Acoustics, Speech and Signal Processing, 2021, doi: 10.1109/ICASSP39728.2021.9415105.

[5] F. Gelderblom, T. Tronstad, T. Svendsen, T. Myrvoll, "On the predictive power of objective intelligibility metrics for the subjective performance of deep complex convolutional recurrent speech enhancement networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 32, pp. 215–226, 2023, doi: 10.1109/TASLP.2023.3329378.

[6] R. Senthilkumar, T. Raghavasimhan, R. Tejeshwini, C. Kavipriya, P. Jayanthi, "Real-time suppression of non-stationary noise for web-based calling applications," *Lecture Notes in Networks and Systems*, vol. 771, pp. 131–140, 2023, doi: 10.1007/978-981-99-5652-4_14.

[7] Y. Li, Z. Zhang, H. Chen, Z. Ma, "Mamba: bringing multi-dimensional ABR to WebRTC," ACM International Conference on Multimedia, 2023, doi: 10.1145/3581783.3611915.

[8] N. Smirnov, S. Tomforde, "Real-time rate control of WebRTC video streams in 5G networks: Improving quality of experience with deep reinforcement learning," *Journal of Systems Architecture*, vol. 148, p. 103066, 2024, doi: 10.1016/j.sysarc.2024.103066.

[9] S. Braun, H. Gamper, C. Reddy, I. Tashev, "Towards efficient models for real-time deep noise suppression," IEEE International Conference Acoustics, Speech and Signal Processing, 2021, doi: 10.1109/ICASSP39728.2021.9413580.

[10] Y. Hu et al., "DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement," arXiv preprint, 2020, doi: arxiv.org/abs/2008.00264.

[11] A. Défossez, G. Synnaeve, Y. Adi, "Real time speech enhancement in the waveform domain," arXiv preprint, 2020, doi: arxiv.org/abs/2011.01843.

[12] N. Westhausen, B. Meyer, "Acoustic echo cancellation with the dual-signal transformation LSTM network," IEEE/ACM Transactions on Audio, Speech and Language Processing, 2021, doi: 10.1109/ICASSP39728.2021.9413510.

[13] F. Jia et al., "Empowering in-browser deep learning inference on edge devices with just-in-time kernel optimizations," arXiv preprint, 2023, doi: arxiv.org/abs/2309.08978.

[14] M. Buffa et al., "Web audio modules 2.0: An open web audio plugin standard," Companion Proceedings of the Web Conference, 2022, doi: 10.1145/3487553.3524225.

[15] H. Kwon, Y. Kim, H. Yoon, D. Choi, "Selective audio adversarial example in evasion attack on speech recognition system," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 526–538, 2020, doi: 10.1109/TIFS.2019.2925452.

[16] M. Vaithianathan, "Digital signal processing for noise suppression in voice signals," *Journal of Computer, Signal, and System Research*, vol. 1, no. 4, pp. 198–208, 2024, doi: 10.61359/11.2206-2417.

[17] S. Zhang, Y. Kong, S. Lv, Y. Hu, L. Xie, "F-T-LSTM based complex network for joint acoustic echo cancellation and speech enhancement," Interspeech, 2021, doi: 10.21437/Interspeech.2021-1359.