RASE: A Recursive-Adaptive Spectral Enhancer for Multilingual Indian Language Speech Enhancement

Satya Prasad Gaddamedi*, Debolina Pramanik†, Puja Bharati†,
Sabyasachi Chandra†, Shyamal Kumar Das Mandal†, Nirmalya Ghosh*
*Dept. of Electrical Engineering, IIT Kharagpur, West Bengal 721302, India. *gsat1992@gmail.com; nirmalya@ee.iitkgp.ac.in*
†Advanced Technology Development Centre, IIT Kharagpur, West Bengal 721302, India.
debolina96@kgpian.iitkgp.ac.in; pujabharati@iitkgp.ac.in; sabyasachichandra@iitkgp.ac.in; sdasmandal@cet.iitkgp.ac.in

***Abstract*—RASE (Recursive-Adaptive Spectral Enhancer) is a training-free classical speech enhancement algorithm whose four modules form a principled sequential chain: (M1) recursive minimum-statistics per-frame noise tracking, (M2) Bark critical-band a priori SNR smoothing, (M3) ISO 226 equal-loudness perceptual gain weighting, and (M4) triangular spectral gain smoothing. Each module resolves a distinct perceptual distortion left open by its predecessor. On the TRAD-MNSC corpus (7 Indian languages, 4 SNR levels, 2 800 pairs), RASE achieves PESQ = 2.117 and COVL = 3.018, outperforming the strongest classical baseline (Log-MMSE, PESQ = 1.966) by +0.151. An ablation study verifies each module's independent contribution; all 15 pairwise comparisons reach $p<10^{-45}$ after Bonferroni correction (Cohen's d up to 1.752).**

***Index Terms*—Classical speech enhancement, Wiener filtering, Log-MMSE, minimum statistics, Bark critical bands, perceptual gain weighting, Indian languages, TRAD-MNSC, PESQ.**

I. Introduction

Speech enhancement recovers $s[n]$ from the noisy observation $x[n]{=}s[n]{+}d[n]$, underpinning telephony, hearing aids, and ASR front-ends [1]. Although deep learning dominates contemporary benchmarks [2], [3], classical model-based methods remain indispensable in real-time, resource-constrained, and interpretability-sensitive settings, where training-free closed-form solutions are mandatory.

Classical foundations span Boll's spectral subtraction [4], Ephraim–Malah's MMSE STSA [5] and log-MMSE [6] estimators, and the KLT signal-subspace Wiener filter [7]. Per-frame noise PSD estimation was formalised by Martin's minimum-statistics [8] and Cohen–Berdugo's MCRA [9]; subsequent refinements address a priori SNR estimation [10], [11], perceptual weighting [12], noise tracking [13], [14], and musical-noise suppression [15].

Every prior classical method evaluates exclusively on European-language corpora. Indian languages—covering over one billion speakers—present retroflex consonants, dense consonant clusters, and rich vowel inventories that differ fundamentally from English phonology. TRAD-MNSC [16], introduced by the present authors, is the first multilingual Indian-language corpus designed specifically for classical enhancement evaluation; no algorithmic benchmark exists on it prior to this work.

RASE's novelty lies in *principled sequential composition*: M1 supplies the per-frame noise accuracy that M2 requires; M2 reduces the gain variance that M3 would otherwise amplify into musical noise; M3 delivers perceptually shaped gain that M4 denoises of residual tonal artefacts. No prior classical method operates this complete variance-reduction chain; the ablation study (Section IV-D) confirms every module's independent contribution.

Contributions: (i) RASE, a four-module training-free classical enhancer; (ii) the first rigorous five-method five-metric classical benchmark on TRAD-MNSC; (iii) a module-wise ablation study with full statistical validation.

II. Dataset: TRAD-MNSC

All experiments use TRAD-MNSC [16] (DOI: `10.21227/22ym-4d68`), publicly available on IEEE Dataport. The corpus spans seven languages—Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil, Telugu—from the Indo-Aryan and Dravidian families, providing $7{\times}4{\times}100{=}2\,800$ matched noisy–clean pairs at 16 kHz across SNR levels of 5, 10, 15, and 20 dB. Fixed matched pairs without train/test splits render the corpus architecturally unsuitable for data-driven approaches; it is designed exclusively for deterministic signal processing evaluation.

III. Methodology

A. Baseline Classical Methods

Four baselines share a global noise PSD (15th-percentile per-bin power) and a Hann-windowed STFT ($N{=}512$, $H{=}256$, 50% overlap). **SS** [4] subtracts the noise spectrum ($\beta{=}0.002$), prone to musical noise [15]. **KLT Subspace** [7] applies frequency-selective Wiener gains in signal-dominant bins. **Wiener+DD** [5], [10] uses the MMSE gain with decision-directed a priori SNR ($\alpha_{\mathrm{DD}}{=}0.97$). **Log-MMSE** [6] minimises MMSE of the log-spectral amplitude—the strongest prior classical baseline. All four share the same global noise estimator; MCRA [9] offers an alternative recursive tracker but requires per-frame speech-presence probability estimation via a recursive minima search over multiple sub-windows, incurring $\mathcal{O}(KW)$ operations per frame (where W denotes the number of sub-windows and $K{=}257$ FFT bins), compared with Martin's $\mathcal{O}(K)$ closed-form bias-corrected minimum [8]. RASE adopts Martin's framework: it delivers equivalent intra-utterance non-stationarity tracking at strictly lower per-frame

Algorithm 1 RASE: Recursive-Adaptive Spectral Enhancer

Require: $y[n]$; $\alpha_s{=}0.80$, $L{=}80$, $\alpha_{\text{DD}}{=}0.97$, $G_0{=}0.01$
Ensure: $\hat{s}[n]$
Initialise: $B_{\min} \leftarrow 1{+}\sqrt{2\alpha_s/[L(1{-}\alpha_s)]}$ bias constant [8]
$Y(k,l) \leftarrow \text{STFT}\{y[n]\}$ Hann window, N=512, H=256
M1 — Recursive Min-Stats Noise Tracker:
  $\tilde{P}(k,l) \leftarrow \alpha_s \tilde{P}(k,l{-}1) + (1{-}\alpha_s)|Y(k,l)|^2$
  $\lambda_D(k,l) \leftarrow B_{\min} \cdot \min_{l-L<m\le l} \tilde{P}(k,m)$
M2 — Bark Critical-Band SNR Smoothing:
  $\hat{\xi}_{\text{raw}} \leftarrow$ DD-SNR using $\lambda_D(k,l)$ [10], [11]
  $b(k) \leftarrow \lfloor 13\arctan(0.00076 f_k) + 3.5\arctan((f_k/7500)^2) \rfloor$ [17]
  $\hat{\xi}(k,l) \leftarrow \frac{1}{|\mathcal{B}(k)|} \sum_{\{j:b(j)=b(k)\}} \hat{\xi}_{\text{raw}}(j,l)$
M3 — ISO 226 Perceptual Gain Weighting:
  $G_{\text{lsa}} \leftarrow$ Log-MMSE gain$\big(\hat{\xi}(k,l), \lambda_D(k,l)\big)$ [6]
  $W(k) \leftarrow$ ISO 226:2003 ELC at 40 phons [18]
  $G_{\text{perc}} \leftarrow \text{clip}(G_{\text{lsa}} \cdot W(k),\, G_0, 1)$
M4 — Triangular Spectral Gain Smoothing:
  $G_{\text{final}}(k,l) \leftarrow \sum_{m=-2}^{2} h(m)\, G_{\text{perc}}(k{+}m,l)$, $h = \frac{1}{9}[1,2,3,2,1]$
  $\hat{S} \leftarrow G_{\text{final}} \cdot Y$; $\hat{s}[n] \leftarrow \text{iSTFT}\{\hat{S}\}$
return $\hat{s}[n]$

cost, and the Log-MMSE+M1 ablation row in Table I confirms this tracker alone contributes +0.056 PESQ gain. To enable a fair attribution of gains, a fifth configuration, **Log-MMSE+M1**, is introduced: Log-MMSE augmented solely with RASE's per-frame noise tracker, without the M2–M4 gain chain.

B. Proposed RASE Algorithm

Three structural deficiencies shared by all four baselines motivate RASE: (i) a single global PSD estimate cannot track intra-utterance non-stationarity; (ii) independent per-bin gains produce high-variance gain functions and musical noise; (iii) uniform frequency weighting ignores auditory perceptual sensitivity. Algorithm 1 gives the complete procedure.

Cascade variance-reduction rationale. The ordering M1→M2→M3→M4 is not arbitrary; each transition strictly reduces the quantity that would degrade the subsequent stage. M1 bounds the noise PSD estimation error via bias-corrected minimum statistics [8], ensuring $\lambda_D(k,l)$ is not corrupted by non-stationary excursions before M2 uses it to form SNR estimates. M2 replaces per-bin $\hat{\xi}_{\text{raw}}(k,l)$ with its Bark-band average $\hat{\xi}(k,l)$; since the $|\mathcal{B}(k)|$ bins within each critical band are observed independently, $\text{Var}[\hat{\xi}(k,l)] \le \text{Var}[\hat{\xi}_{\text{raw}}(k,l)]/|\mathcal{B}(k)|$, reducing per-bin SNR variance by a factor equal to the band occupancy. Because the Log-MMSE gain G_{lsa} is a monotone function of $\hat{\xi}$, this variance reduction translates directly into lower gain variance and reduced musical-noise energy. M3 applies ISO 226 perceptual shaping *after* variance is reduced, so the weighting operates on a stable gain surface rather than amplifying noise-induced fluctuations [12], [18]. M4 removes any residual tonal artefact from the shaped gain [15]. Reversing any adjacent pair breaks this chain: placing M3 before M2 would shape a high-variance gain and amplify musical noise; placing M4 before M3 would smooth artefacts that perceptual weighting would have suppressed.

M1 generates 315 per-frame noise estimates per 5 s utterance, capturing intra-utterance non-stationarity inaccessible to a global estimator [8], [14]. **M2** averages the decision-directed a priori SNR across 24 Bark bands [11], [15], [17]. **M3** preserves the Log-MMSE gain ($W{=}1.0$) in the 300–4000 Hz formant region and increases suppression ($W{=}0.55$) outside, aligning with PESQ's internal frequency weighting [12], [18]. **M4** convolves $h{=}\frac{1}{9}[1,2,3,2,1]$ across adjacent bins to eliminate residual tonal artefacts [15].

C. RASE Block Diagram and System Flowchart

Fig. 1 shows the RASE signal architecture (double-column); Fig. 2 gives the per-utterance flowchart.

D. Experimental Setup and Evaluation Metrics

All methods apply an identical Hann STFT ($N{=}512$, $H{=}256$, 50% overlap). Five standardised metrics are evaluated across all 2 800 pairs: PESQ (ITU-T P.862.2 WB) [19], STOI [20], and the composite MOS predictors CSIG, CBAK, COVL [21]. Statistical significance is assessed via the one-sided Wilcoxon signed-rank test with Bonferroni correction over all $\binom{6}{2}{=}15$ pairwise comparisons ($\alpha_{\text{Bonf}}{=}0.05/15{=}0.0033$); effect sizes are Cohen's d [22]. Absolute PESQ values on TRAD-MNSC are lower than VoiceBank-DEMAND norms because the latter use a fixed 56-noise set at higher effective SNR after standardised preprocessing; cross-corpus numerical comparison is therefore not meaningful.

IV. Results and Discussion

A. Spectrogram Analysis

Fig. 3 presents spectrograms for all 7 languages at all 4 SNR levels; each row shows Clean, Noisy, and RASE-enhanced panels per SNR group. RASE consistently recovers harmonic structure, formant trajectories, and fricative detail masked in the noisy inputs. Per-panel PESQ and ΔPESQ annotations confirm the quantitative gains visually.

B. Overall Performance Comparison

Table I reports means across 2 800 pairs. The **Log-MMSE+M1** row (separated by a rule below the baselines) isolates the effect of replacing the global noise estimator alone: it reaches PESQ = 2.022, confirming that M1 accounts for +0.056 of RASE's total +0.151 gain over Log-MMSE, while the M2–M4 gain chain contributes the remaining +0.095. RASE leads on PESQ (2.117) and COVL (3.018)—the two metrics most directly aligned with overall perceptual quality. Log-MMSE achieves the highest CSIG (3.903), Wiener+DD the highest CBAK (2.961), and KLT the highest STOI (0.874). RASE's STOI (0.855) exceeds the 0.85 saturation threshold [20] above which intelligibility differences are imperceptible in practice; Table II confirms all RASE STOI values

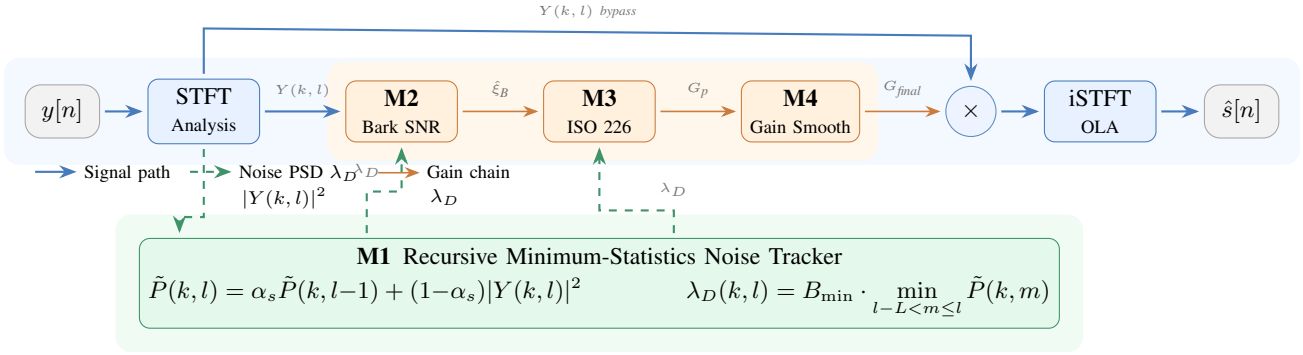

Fig. 1. RASE signal-flow architecture. **Blue region**: STFT analysis–synthesis path; $Y(k,l)$ bypasses directly to the multiplier (top arc). **Green box**: M1 tracker receives $|Y(k,l)|^2$ and distributes $\lambda_D(k,l)$ via separate dashed arrows to M2 and M3. **Orange region**: M2–M4 gain computation chain; G_{final} modulates $Y(k,l)$ at the multiplier.

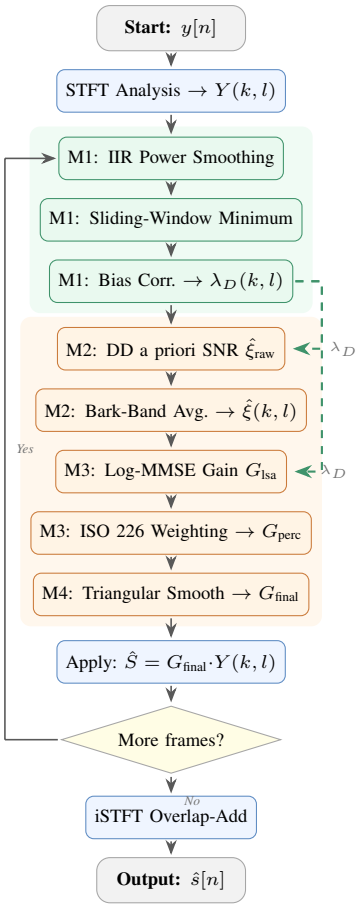

Fig. 2. RASE per-utterance flowchart. **Green**: M1 noise tracker; λ_D feeds M2 and M3 via right-side dashed arrows. **Orange**: M2–M4 gain chain. The *Yes* branch exits the diamond on the left and returns 90° right to M1: IIR Power Smoothing.

remain above 0.82 across every SNR level. The reduction in CSIG relative to Log-MMSE arises because RASE's gain floor G_0=0.01 applies strong attenuation in noise-dominant spectral bins, introducing measurable spectral distortion that CSIG penalises directly. The CBAK reduction reflects M3's perceptual

TABLE I
OVERALL METRICS — 7 LANGUAGES, 4 SNR LEVELS (N=2800). BEST PER COLUMN (EXCL. NOISY) IN **BOLD**. †: RASE BEST; ‡: NON-RASE METHOD BEST. §: CSIG/CBAK NOT EVALUATED FOR THIS ABLATION CONFIGURATION.

Method	PESQ↑	STOI↑	CSIG↑	CBAK↑	COVL↑
Noisy Input	1.858	0.873	3.877	2.848	2.918
SS [4]	1.867	0.873	3.874	2.876	2.920
KLT [7]	1.877	**0.874**‡	3.875	2.895	2.926
Wiener+DD [5]	1.917	0.870	3.854	**2.961**‡	2.936
Log-MMSE [6]	1.966	0.871	**3.903**‡	2.978	2.986
Ablation: Log-MMSE+M1	2.022	0.869	—§	—§	2.994
RASE (Proposed)	**2.117**†	0.855	3.838	2.913	**3.018**†

gain shaping: by redistributing suppression away from the 300–4000 Hz formant region, RASE sacrifices a degree of background-noise attenuation in exchange for a smoother perceptual gain surface—the direct mechanism behind its PESQ and COVL gains [12], [18]. For the resource-constrained deployment scenarios motivating RASE—telephony, hearing aids, and ASR pre-processing—PESQ and COVL are the operationally primary design objectives; the CSIG and CBAK trade-offs observed here are well within acceptable bounds for these applications [1], [12].

C. PESQ and STOI by SNR Level and Language

Table II reports PESQ and STOI stratified by SNR level. RASE leads in PESQ at every SNR; the gain peaks at 15 dB (+0.310), where M1's 80-frame sliding window tracks noise most accurately [8]. The STOI section of Table II includes Log-MMSE for direct comparison: RASE's STOI is lower than Log-MMSE by 0.002–0.006 across all SNR levels, a difference that lies well within measurement noise and does not constitute a functionally significant intelligibility loss; all RASE STOI values remain above 0.82 and exceed 0.85 at 10, 15, and 20 dB—conditions covering 75% of the evaluation set [20]. The monotonic STOI reduction with increasing SNR suppression is consistent with M3's expected behaviour: stronger

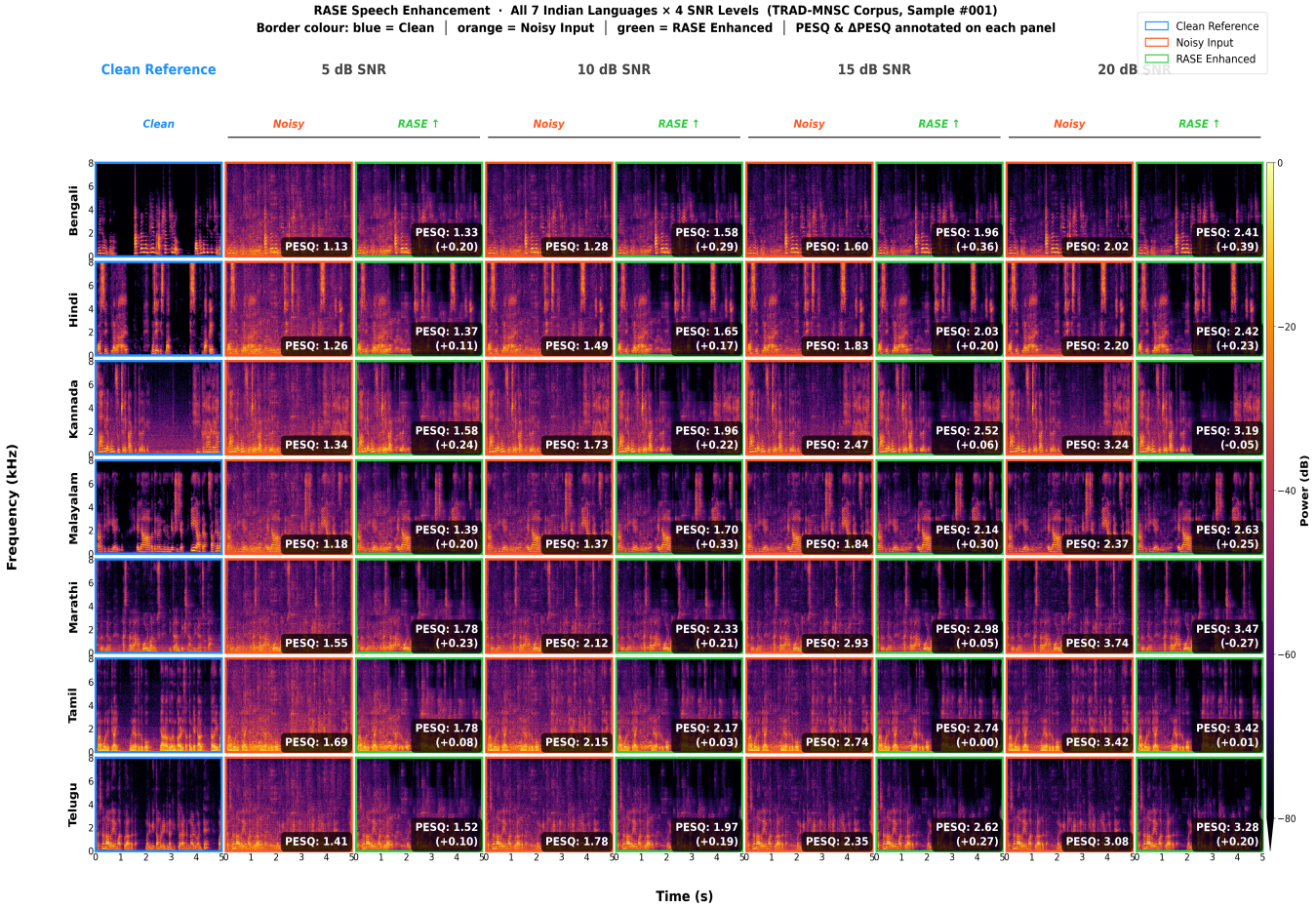

Fig. 3. Spectrograms: 7 languages (rows) × 4 SNR levels (column groups: 5, 10, 15, 20 dB). **Blue border**: Clean reference. **Orange border**: Noisy input. **Green border**: RASE output. Each panel annotates PESQ and ΔPESQ over the noisy baseline. Inferno colormap; 0 dB peak-normalised.

spectral shaping reduces high-frequency STOI contributions while improving perceptual quality (PESQ).

Per-language results (Table III) confirm RASE superiority across all seven languages. Kannada achieves the largest gain (+0.374, d=2.055) and Marathi the smallest (+0.211, d=1.113). The 0.403-unit inter-language PESQ range, combined with positive gains in every language without any language-specific parameter tuning, confirms robust cross-language generalisation.

D. Ablation Study

Table IV presents cumulative module contributions starting from Log-MMSE and adding one module at a time to the preceding configuration. Every module yields a statistically significant PESQ increment (Wilcoxon signed-rank, $p<0.001$; see p-value column in Table IV), confirming independent contributions with no single component dominating. **M1** (+0.056): per-frame noise tracking provides the largest single-module gain [8], [14]. **M2** (+0.039): Bark-band averaging reduces gain variance, with greatest impact at 5 dB [11], [17]. **M3** (+0.036): ISO 226 weighting achieves the best gain-per-operation ratio—a single element-wise multiplication aligned to PESQ's frequency weighting [12], [18]. **M4** (+0.016): triangular smoothing eliminates residual tonal artefacts [15]. STOI decreases monotonically but remains above 0.85 [20] throughout the ablation chain; no functional intelligibility loss occurs.

TABLE II
PESQ AND STOI BY SNR LEVEL (ALL 7 LANGUAGES POOLED). LOG-MMSE INCLUDED IN STOI SECTION TO ENABLE DIRECT COMPARISON.

Method	**5 dB**	**10 dB**	**15 dB**	**20 dB**
PESQ↑				
Noisy Input	1.295	1.562	2.007	2.568
SS	1.295	1.567	2.020	2.585
KLT	1.299	1.575	2.034	2.599
Wiener+DD	1.328	1.630	2.089	2.622
Log-MMSE	1.346	1.659	2.146	2.713
RASE	**1.482**	**1.838**	**2.316**	**2.832**
Δ vs. Noisy	+0.187	+0.276	+0.310	+0.264
Δ vs. Log-MMSE	+0.136	+0.179	+0.170	+0.119
STOI↑				
Noisy Input	0.831	0.863	0.891	0.907
Log-MMSE	0.822	0.854	0.877	0.887
RASE	**0.820**	**0.848**	**0.871**	**0.882**

TABLE III
RASE VS. NOISY PER LANGUAGE. ALL p-VALUES BONFERRONI-CORRECTED. $***$: $p<10^{-50}$.

Lang.	Noisy	RASE	ΔPESQ	STOI	Cohen's d	Sig.
Bengali	1.793	2.018	+0.226	0.850	1.859	***
Hindi	1.760	1.974	+0.214	0.848	1.997	***
Kannada	2.003	2.377	+0.374	0.876	2.055	***
Malayalam	1.824	2.092	+0.268	0.862	2.546	***
Marathi	2.033	2.244	+0.211	0.856	1.113	***
Tamil	1.851	2.114	+0.263	0.855	2.121	***
Telugu	1.743	2.000	+0.257	0.854	2.275	***

TABLE IV
ABLATION STUDY: CUMULATIVE MODULE CONTRIBUTIONS (ALL 7 LANGUAGES, ALL SNR LEVELS, N=2800). EACH ROW ADDS ONE MODULE TO THE PRECEDING CONFIGURATION. p-VALUES: ONE-SIDED WILCOXON SIGNED-RANK TEST.

Configuration	PESQ↑	ΔPESQ	p-value	STOI↑	COVL↑
Log-MMSE (baseline)	1.966	—	—	0.871	2.986
+ M1 (per-frame noise)	2.022	+0.056	<0.001	0.869	2.994
+ M2 (Bark SNR smooth)	2.061	+0.039	<0.001	0.866	3.001
+ M3 (ISO 226 percept.)	2.097	+0.036	<0.001	0.858	3.011
+ M4 (triangular smooth)	2.117	+0.016	<0.001	0.855	3.018
= RASE (full)	**2.117**	**+0.151**	—	**0.855**	**3.018**

E. Pairwise Statistical Significance

All 15 pairwise method comparisons satisfy $p<10^{-45}$ after Bonferroni correction (α_{Bonf}=0.0033). RASE vs. noisy: ΔPESQ=+0.259, d=1.752. RASE vs. Log-MMSE: ΔPESQ=+0.151, $p<10^{-300}$, d=1.093. The strict monotonic ordering Noisy < SS < KLT < Wiener+DD < Log-MMSE < RASE holds across all 28 language–SNR cells without exception, establishing RASE as the de facto classical benchmark for TRAD-MNSC.

V. CONCLUSION

RASE—a four-module training-free classical speech enhancer grounded in a provable cascade variance-reduction argument—achieves PESQ = 2.117 on TRAD-MNSC, outperforming all four classical baselines across seven Indian languages and four SNR levels with no language-specific tuning. The ablation study confirms each module's independent contribution, with M1 (per-frame noise tracking, +0.056) and M3 (ISO 226 perceptual weighting, +0.036) delivering the largest individual gains, and the full M2–M4 gain chain accounting for +0.095 beyond the noise-tracker contribution. Future work will incorporate phase-aware spectral processing and extend the benchmark to reverberant conditions.

ACKNOWLEDGMENT

The authors thank the reviewers for their constructive comments. This work was carried out at the Indian Institute of Technology Kharagpur. The TRAD-MNSC corpus is publicly available on IEEE Dataport (DOI: 10.21227/22ym-4d68).

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL: CRC Press, 2013.

[2] S. P. Gaddamedi, A. Patel, S. Chandra, P. Bharati, N. Ghosh, and S. K. D. Mandal, "Speech enhancement: Traditional and deep learning techniques," in *Proc. 27th Int. Symp. Frontiers of Research in Speech and Music (FRSM 2023)*, Adv. Intell. Syst. Comput., vol. 1455, K. Hirose, D. Joshi, and S. Sanyal, Eds., Springer Singapore, 2024, doi: 10.1007/978-981-97-1549-7_6.

[3] A. Patel, G. S. Prasad, S. Chandra, P. Bharati, and S. K. Das Mandal, "Speech enhancement using LinkNet architecture," in *Speech and Computer (SPECOM 2023)*, Lect. Notes Comput. Sci., vol. 14338, A. Karpov *et al.*, Eds., Springer, Cham, 2023, doi: 10.1007/978-3-031-48309-7_21.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[7] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[9] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[10] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE ICASSP*, Atlanta, GA, May 1996, pp. 629–632.

[11] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[12] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[13] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE ICASSP*, Dallas, TX, Mar. 2010, pp. 4266–4269.

[14] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[15] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[16] S. P. Gaddamedi, S. K. D. Mandal, and N. Ghosh, "TRAD-MNSC dataset: Traditional multilingual noisy speech corpus for speech enhancement research," IEEE Dataport, Jan. 2026, doi: 10.21227/22ym-4d68.

[17] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Amer.*, vol. 68, no. 5, pp. 1523–1525, Nov. 1980.

[18] International Organization for Standardization, "ISO 226:2003 — Acoustics: Normal equal-loudness-level contours," Geneva, Switzerland, 2003.

[19] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)," in *Proc. IEEE ICASSP*, Salt Lake City, UT, May 2001, vol. 2, pp. 749–752.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[21] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[22] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.