

Analyzing the Correlation and Impact of Speech Evaluation Metrics on Real-World Speaker Verification and Speech Recognition

Tan-Loc Le^{1,2}, Van-Huy Nguyen^{3,4}, Tri-Nhan Do³, Trung-Kien Phan³, Dang-Khoa Mac³

¹Ho Chi Minh City University of Technology, ²VNU-HCMC, ³VinBigData, ⁴Thai Nguyen University of Technology

INTRODUCTION

- In practical speech systems like ASR and Speaker Verification, audio quality is critical but clean references rarely exist.
- Existing metrics (PESQ, STOI, etc.) often need a reference ("intrusive") or are mostly English-only.
- Unknown: Which non-intrusive or estimated metrics best predict real-world ASR/SV performance across languages and noise?

Our Contributions:

- Benchmark estimated and non-intrusive quality metrics (PESQ, STOI, SI-SDR, Distill-MOS, Audiobox, etc.) for their correlation with ASR/SV performance.
- Evaluate on both English and Vietnamese datasets under diverse noise.
- Provide practical guidance for quality monitoring in deployment.

METHOD

1. Data & Downstream Tasks

- **ASR Datasets:** We used a *Vietnamese dataset* (studio-quality, with MUSAN-noise augmentation), *VoiceBank+DEMAND* (English speech with diverse synthetic noises), and both *URGENT-Challenge 2024* (English, challenging real-world noise) and *2025* (multilingual, broader distortions) datasets to comprehensively test speech recognition under varied conditions.
- **SV Dataset:** Speaker verification was evaluated using *VoxCeleb1*, a large set of real-world, celebrity speech clips, both clean and with systematically added noise from 1 dB to 20 dB SNR.

2. Models:

- **ASR:** *Whisper* (general multilingual ASR) and *PhoWhisper* (Vietnamese-specific ASR fine-tuned on diverse accents) were used to compare performance across languages and conditions.
- **SV:** For speaker verification, we used *WavLM* (transformer-based, robust to noise) and *ECAPA-TDNN* (strong speaker embedding extractor for SV), focusing on real-world speech variability.

3. Metrics Evaluated

- **Objective/Estimated:** We computed *TorchAudio-Squim*'s reference-free estimates of standard objective metrics (*PESQ* for perceptual quality, *STOI* for intelligibility, *SI-SDR* for signal separation, *Est-SNR* and *WADA-SNR* for reference-free SNR estimation) using deep learning models for each noisy audio sample.
- **Subjective/No-Reference:** We measured subjective speech quality with *Distill-MOS* (distilled MOS model for perceived quality), and with four axes from *Audiobox-Aesthetics* (CE: Content Enjoyment, CU: Content Usefulness, PC: Production Complexity, PQ: Production Quality), to capture human-like judgments without reference signals.
- **Evaluation:** For every dataset and noise level, we calculated the *Pearson correlation* coefficient between each metric (objective or subjective) and downstream task scores—specifically, *WER/CER* (ASR) and *EER* (Speaker Verification) — to identify which metrics best predict real-world system performance.

Experiments

Table 1. Relationship between SNR, EER, and speech quality metrics.

SNR (dB)	EER	PESQ	STOI	SI-SDR	WADA-SNR
1	0.1402	1.1589	0.7600	0.4074	5.1121
5	0.0968	1.2148	0.8214	4.2526	7.3492
10	0.0697	1.3667	0.8822	8.7517	10.2863
15	0.0578	1.6215	0.9236	12.6708	13.0217
20	0.0530	1.9435	0.9479	15.7345	15.2866

Table 2. Correlation between average quality metrics and SV matching scores (Pearson ρ). Higher scores are desired for positive pairs, lower for negative pairs..

Metric	Positive Pairs	Negative Pairs
PESQ_avg	0.275169	-0.079335
STOI_avg	0.375938	-0.096068
SI-SDR_avg	0.307812	-0.079471
EST-SNR_avg	0.191972	-0.064522
WADA-SNR_avg	0.183765	-0.061650

RESULT

- *Estimated STOI* provides the strongest correlation with ASR and SV accuracy across languages and noise conditions.
- *Non-intrusive metrics* like *Distill-MOS* and *Audiobox* also offer valuable insights for real-world system monitoring when reference audio is not available.
- *Future work* will expand testing to more languages and noise types, and explore combining multiple metrics for even better prediction of downstream performance.

Table 3. Correlation between quality metrics and ASR performance. We present the absolute values for clarity on strength. Higher absolute values indicate stronger correlation.

Metric	Vietnamese Dataset		VoiceBank+DEMAND		URGENT-Challenge-2024	
	WER	CER	WER	CER	WER	CER
PESQ	0.399288	0.369979	0.307051	0.325640	0.256634	0.216948
STOI	0.737747	0.721363	0.398421	0.487822	0.468690	0.431914
WADA-SNR	0.437042	0.418519	0.151336	0.160584	0.184393	0.171405

ORGANIZERS



M A P R 2025

Nha Trang, 14-15 August 2025

SPONSORS

