6th International Conference on Computing and Communication Networks
(ICCCNet-2026)
ORGANISED BY : MANCHESTER METROPOLITAN UNIVERSITY, MANCHESTER, UNITED KINGDOM
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)
17th - 19th July, 2026

Benchmarking Streaming ASR for Real-time Deployment: A Robustness Scorecard and Error Taxonomy for Vietnamese

Quoc Hung Nguyen, Hoai An Thai, Duy Tan Nguyen, Vy Le, Thi Thuy Duong Nguyen

UEH College of Technology and Design — University of Economics Ho Chi Minh City

Corresponding author: hungngq@ueh.edu.vn

ICCCNet-2026 · Manchester, UK · 17 July 2026

6th International Conference on Computing and Communication Networks
(ICCCNet-2026)
ORGANISED BY : MANCHESTER METROPOLITAN UNIVERSITY, MANCHESTER, UNITED KINGDOM
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)
17th - 19th July, 2026

1. INDEX

Outline

- Motivation — why offline WER misleads streaming deployment
- Related work & the gap for Vietnamese streaming ASR
- Evaluation pipeline: chunking, latency profiles, metrics
- Results: offline vs. streaming WER, latency–accuracy frontier
- Robustness & stability under distribution shift
- Comparative analysis: accuracy vs. runtime trade-off
- Conclusion, error taxonomy & future work

6th International Conference on Computing and Communication Networks
(ICCCNet-2026)
ORGANISED BY : MANCHESTER METROPOLITAN UNIVERSITY, MANCHESTER, UNITED KINGDOM
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)
17th - 19th July, 2026

2. ABSTRACT

Offline WER alone is insufficient for streaming model selection

- Vietnamese ASR is increasingly deployed in production — under latency budgets and limited compute, not offline conditions.
- A reproducible streaming evaluation framework: standardized chunking, right-context (look-ahead), overlap handling & metrics.
- Reports WER and real-time factor (RTF) under three latency profiles: 1200 / 2400 / 4000 ms.
- Evaluated on 4 public datasets — VIVOS, VLSP2020, Viet YouTube ASR v2, Speech-MASSIVE_vie — as a robustness scorecard.
- Adds a preliminary Vietnamese error taxonomy: numerals, punctuation, named entities & Vietnamese–English code-switching.

➜ **One framework that reports accuracy, latency & robustness together — the numbers a deployment decision needs.**

6th International Conference on Computing and Communication Networks
(ICCCNet-2026)
ORGANISED BY : MANCHESTER METROPOLITAN UNIVERSITY, MANCHESTER, UNITED KINGDOM
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)
17th - 19th July, 2026

3. INTRODUCTION

Model selection by offline WER is misaligned with real-time deployment

- **The problem**
  - Production ASR runs incrementally under latency budgets & limited compute — offline full-context decoding does not.
  - Streaming WER depends on chunk size, look-ahead & overlap; described inconsistently → papers hard to compare.
- **Sharper for Vietnamese**
  - Benchmarks favor clean read speech, but real usage is spontaneous, in-the-wild and accent-variable.
  - Numerals, punctuation, named entities & Vietnamese–English code-switching add errors aggregate WER can hide.

➜ **We need a system-oriented benchmark that measures models the way they are actually deployed.**

6th International Conference on Computing and Communication Networks
(ICCCNet-2026)
ORGANISED BY : MANCHESTER METROPOLITAN UNIVERSITY, MANCHESTER, UNITED KINGDOM
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)
17th - 19th July, 2026

4. LITERATURE REVIEW

No prior work jointly reports WER, RTF & streaming degradation for Vietnamese

Work	Lang.	Streaming	Latency / RTF	Robustness	Error analysis	Repro.
WeNet toolkit [20]	Multi	Yes	Yes	No	No	Yes
Open ASR Leaderboard [19]	Multi	No	Yes	No	No	Yes
Vietnamese ASR benchmarks [1,15]	Vi	No	No	Partial	No	Partial
This work	Vi	Yes	Yes	Yes	Yes	Yes

Table 1 — Related ASR benchmarks (green = supported, red = not).

Four gaps we close: G1 streaming settings not standardized · G2 latency & cost reported in isolation · G3 robustness not tied to streaming · G4 no Vietnamese error breakdown

➜ **First Vietnamese benchmark to unify streaming, latency/RTF, robustness, error analysis & reproducibility.**

6th International Conference on Computing and Communication Networks
(ICCCNet-2026)
ORGANISED BY : MANCHESTER METROPOLITAN UNIVERSITY, MANCHESTER, UNITED KINGDOM
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)
17th - 19th July, 2026

5. RESEARCH GAPS

Four gaps motivate a system-oriented streaming benchmark

G1 Streaming settings not standardized

Chunk size, look-ahead and overlap vary across papers, so reported streaming WER is not comparable.

G2 Latency & cost reported in isolation

WER, ΔWER (streaming − offline) and RTF are rarely reported together under the same latency profiles.

G3 Robustness not tied to streaming

Distribution-shift robustness is studied offline; streaming can amplify degradation but it goes unmeasured.

G4 No Vietnamese error breakdown

Numerals, punctuation, named entities and code-switching are not analyzed for Vietnamese streaming ASR.

➜ **Our contribution: a reproducible protocol + robustness scorecard + preliminary error taxonomy that close G1–G4.**

6th International Conference on Computing and Communication Networks
(ICCCNet-2026)
ORGANISED BY : MANCHESTER METROPOLITAN UNIVERSITY, MANCHESTER, UNITED KINGDOM
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)
17th - 19th July, 2026

6. PROPOSED METHODOLOGY

A reproducible pipeline: fixed-size chunking, three latency budgets, one command

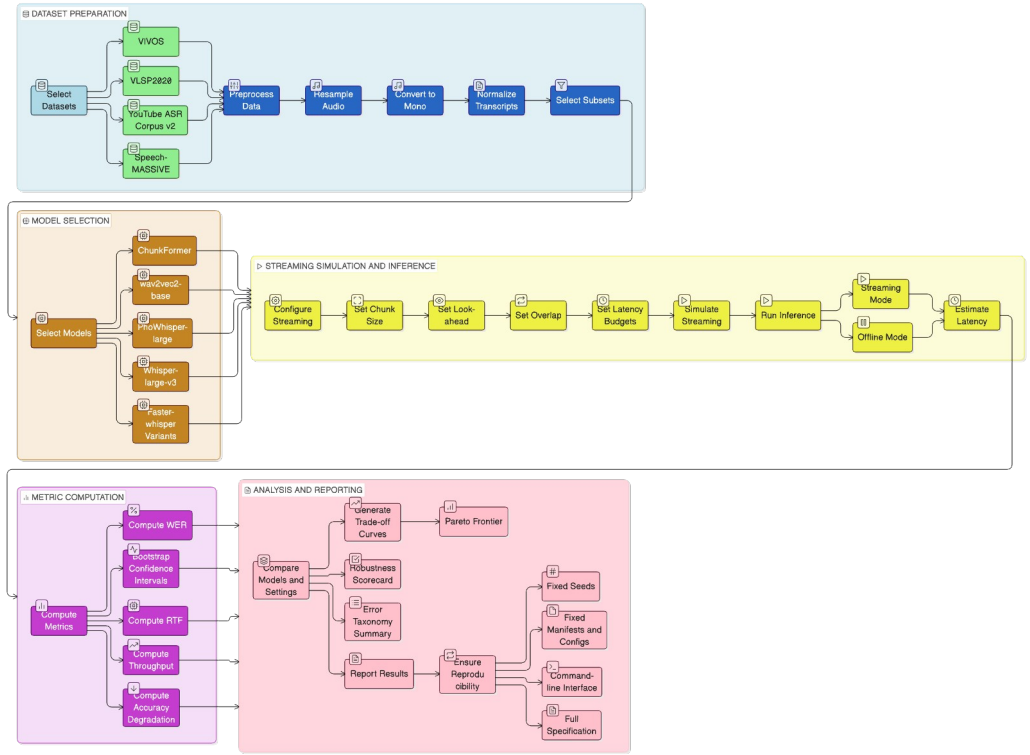

Fig. 1 — Evaluation pipeline (4 stages).

Profile	Chunk	Overlap	Look-ahead
Low latency	1200 ms	200 ms	0 ms
Balanced	2400 ms	200 ms	0 ms
High context	4000 ms	200 ms	0 ms

- Data: VIVOS (clean) · VLSP2020 (spontaneous) · Viet YouTube v2 (in-the-wild) · Speech-MASSIVE_vie (assistant) — 300-utt subsets, unified normalization.
- Streaming models: ChunkFormer-CTC-large (110M) · wav2vec2-base-vi (95M) — swept over all 3 profiles.
- Offline baselines: PhoWhisper-large & Whisper-large-v3 (1.55B).
- Metrics: strict WER + 95% bootstrap CI · RTF & throughput (1× H100) · streaming stability. Inference-only — no retraining.

6th International Conference on Computing and Communication Networks
(ICCCNet-2026)
ORGANISED BY : MANCHESTER METROPOLITAN UNIVERSITY, MANCHESTER, UNITED KINGDOM
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)
17th - 19th July, 2026

7. RESULTS & DISCUSSION

Streaming inflates WER far beyond the offline gap — worst on spontaneous speech

Model	Dataset	Offline	Streaming	ΔWER
ChunkFormer-CTC (110M)	VIVOS	3.44	6.17	+2.74
	VLSP2020	3.83	**50.64**	**+46.81**
	Viet YouTube v2	4.48	5.44	+0.96
	Speech-MASSIVE	18.78	21.38	+2.60
wav2vec2-base-vi (95M)	VIVOS	8.88	12.32	+3.44
	VLSP2020	10.09	**52.74**	**+42.65**
	Viet YouTube v2	7.90	8.31	+0.40
	Speech-MASSIVE	27.71	30.38	+2.68

Table 2 — WER (%). Offline PhoWhisper / Whisper baselines in paper.

- **Offline gap is small everywhere — streaming reveals it.**
  - VLSP2020 (spontaneous) collapses to ~50% streaming: +40–47 pp.
  - Clean (VIVOS) & in-the-wild (YouTube) stay well-behaved.
  - Robustness scorecard @ 4000 ms: VLSP2020 shift +44 pp / +720% vs clean VIVOS.
- **Offline ranking ≠ streaming ranking.**

➜ **Two models with near-identical offline WER diverge by tens of points once streamed.**

6th International Conference on Computing and Communication Networks
(ICCCNet-2026)
ORGANISED BY : MANCHESTER METROPOLITAN UNIVERSITY, MANCHESTER, UNITED KINGDOM
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)
17th - 19th July, 2026

8. COMPARATIVE ANALYSIS

Latency helps every model — but ChunkFormer leads accuracy, wav2vec2 leads speed

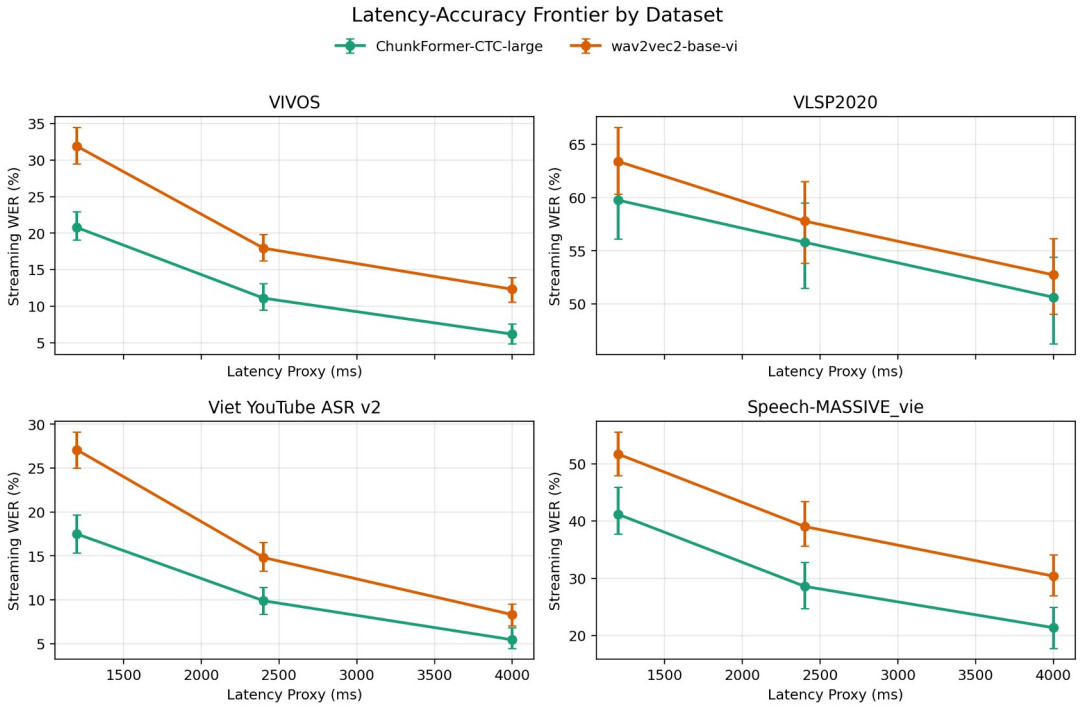

Fig. 2 — Streaming WER (95% CI) across 1200 / 2400 / 4000 ms.

Model	RTF	Throughput	Mode
ChunkFormer-CTC (110M)	0.0111	**90.1×**	**Yes**
wav2vec2-base-vi (95M)	**0.0059**	**169.5×**	**Yes**
PhoWhisper-large (1.55B)	0.1034	9.7×	Offline
Whisper-large-v3 (1.55B)	0.0903	11.1×	Offline

Table 3 — RTF & throughput on 1× H100 80GB (lower RTF / higher throughput = faster).

- Bigger chunks cut streaming WER: mean 35%→21% (ChunkFormer), 44%→26% (wav2vec2).
- ChunkFormer wins accuracy everywhere; wav2vec2 ~1.9× faster; big offline models ~10–17× costlier.
- VLSP2020 stays >50% even at 4000 ms → instability, not just latency.

➜ **No model wins both — deployment is a point on the WER × RTF × latency frontier.**

6th International Conference on Computing and Communication Networks
(ICCCNet-2026)
ORGANISED BY : MANCHESTER METROPOLITAN UNIVERSITY, MANCHESTER, UNITED KINGDOM
Springer LNNS Approved Conference (Indexed in Scopus, EI, WoS and Many More)
17th - 19th July, 2026

9. CONCLUSION & FUTURE WORK

A deployment-aligned benchmark changes which Vietnamese ASR model you ship

Conclusions

- Offline-WER-only selection can mislead — streaming constraints & runtime cost reorder model families.
- Matched-latency WER / ΔWER / RTF comparison exposes favorable accuracy–efficiency configurations.
- Robustness scorecard: streaming amplifies degradation on spontaneous & in-the-wild Vietnamese speech.

Future work

- Quantify the error taxonomy: numerals, abbreviations, named entities, Vietnamese–English code-switching.
- More streaming architectures & low-resource adaptation; explicit latency-budget sweeps; finer end-to-end latency.
- Grow into an open, continuously updated Vietnamese streaming ASR evaluation platform.

➜ **Open & reproducible: github.com/hoaianthai345/ASR_Vietnamese_Benchmark · hungngq@ueh.edu.vn**

10. REFERENCES

References (selected)

[1] AILAB, VNUHCM-US. VIVOS: Vietnamese speech corpus for ASR. Zenodo (2016).

[2] Barker, J. et al. The third CHiME speech separation and recognition challenge. Comput. Speech Lang. 46 (2017).

[3] Benzeghiba, M. et al. ASR and speech variability: a review. Speech Communication 49 (2007).

[4] Bisani, M., Ney, H. Bootstrap estimates for confidence intervals in ASR evaluation. ICASSP (2004).

[5] Dinh, N.V. et al. Multi-dialect Vietnamese: task, dataset, baseline models and challenges. EMNLP (2024).

[9] Speech-MASSIVE: multilingual assistant-style speech dataset (Vietnamese subset).

[11] Viet YouTube ASR v2: in-the-wild Vietnamese speech corpus.

[15] VLSP2020 Vietnamese ASR shared task dataset.

[19] Open ASR Leaderboard. [20] WeNet: streaming end-to-end speech recognition toolkit.

[22] Streaming ASR models — latency/RTF under controlled configurations.

Full list (23 entries) in the paper. Repository: github.com/hoaianthai345/ASR_Vietnamese_Benchmark