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

Quoc Hung NGUYEN[0000-0002-6363-0160]*, Hoai An THAI[0009-0004-4093-522X], Duy Tan NGUYEN[0009-0000-2832-8647], Vy LE[0009-0003-0387-8302], and Thi Thuy Duong NGUYEN[0009-0000-3548-2212]

UEH College of Technology and Design, University of Economics Ho Chi Minh City
Corresponding author: hungngq@ueh.edu.vn

Abstract. Vietnamese ASR is increasingly deployed in production, where offline word error rate (WER) alone is insufficient for model selection under streaming latency and runtime budgets. We present a reproducible streaming evaluation framework that standardizes chunking, right context (look-ahead), overlap handling, and metric computation, and reports WER and real-time factor (RTF) under multiple deployment-oriented latency profiles (1200/2400/4000 ms) across representative model families. We evaluate on VIVOS, VLSP2020, Viet YouTube ASR v2, and Speech-MASSIVE_vie to produce a robustness scorecard under streaming stressors and distribution shifts. We further provide a preliminary Vietnamese-specific error taxonomy (numerals, punctuation, named entities, and Vietnamese–English code-switching) to translate benchmark outcomes into deployment guidance.

Keywords: Vietnamese ASR · Streaming evaluation · Latency · Real-time factor · Robustness

1 Introduction

Vietnamese automatic speech recognition (ASR) is widely used in production [23, 6]. Model selection, however, is still often driven by offline word error rate (WER). This is misaligned with streaming deployment, where systems must satisfy latency budgets and limited compute [20, 16]. The problem is sharper for Vietnamese because evaluation data often differ from real usage: read speech vs. spontaneous and in-the-wild speech [3, 6], combined with strong accent variability [1, 15, 8, 5].

To facilitate transparent evaluation and reproducibility, we release the full benchmark implementation and configuration. [1]

We present a system-oriented benchmark for Vietnamese streaming ASR. Our contributions are:

- We define a reproducible streaming evaluation protocol (with a reference implementation) covering chunking, look-ahead, overlap handling, and metric computation.
- We report WER, ΔWER (streaming minus offline), and RTF under three streaming latency profiles representative of practical deployment.
- We provide a robustness scorecard measuring degradation under streaming stressors and realistic distribution shifts.
- We provide a preliminary Vietnamese-specific error taxonomy and summarize practical implications for deployment.

2 Related Work

ASR evaluation often reports offline word error rate (WER), which reflects transcription accuracy but not the constraints of streaming deployment. In streaming decoding, limited right context and bounded computation can change both accuracy and latency, so the reported WER depends on settings such as chunk size, look-ahead, and overlap. Because these settings are not always described in a consistent way, comparing results across papers can be difficult [22, 20].

Several toolkits report runtime metrics such as latency, real-time factor (RTF), and throughput alongside WER, which helps evaluate deployment trade-offs. Streaming toolkits also expose decoding settings and make it possible to report latency and RTF under controlled configurations [20]. Open leaderboards standardize reporting, but many rankings still emphasize offline accuracy and provide limited support for streaming controls or stress testing [19].

Vietnamese ASR performance on public benchmarks does not always match real-world usage. Many studies still evaluate on clean read speech, while production audio often includes spontaneous speech, in-the-wild recordings, and strong accent variation [1, 15, 8]. In addition, Vietnamese transcripts are sensitive to text normalization choices such as numerals and punctuation [18, 21]. Named entities and Vietnamese–English code-switching further contribute errors that aggregate WER may not capture well [14, 10].

Robustness under distribution shift has been studied using controlled tests with noise, speaking rate variation, reverberation, and domain mismatch [2, 7]. These studies report clear performance drops outside the training domain. However, robustness is rarely evaluated together with streaming constraints. Because streaming decoding operates with a truncated right context, partial hypotheses are revised more frequently, especially under domain mismatch. Existing streaming evaluations and toolkits report latency/RTF and WER but do not jointly

[1] Repository: `https://github.com/hoaianthai345/ASR_Vietnamese_Benchmark.git`.

analyze robustness under distribution shift [22, 20]. We are not aware of prior Vietnamese benchmarks that jointly report WER, RTF, and streaming degradation under realistic shifts. Table 1 summarizes these gaps.

Table 1. Comparison of related ASR benchmarks.

Work	Language	Streaming	Latency / RTF	Robustness	Error Analysis	Reproducible
Representative Streaming ASR models [22]	Multi	Yes	Partial	No	No	Partial
WeNet toolkit [20]	Multi	Yes	Yes	No	No	Yes
Open ASR Leaderboard [19]	Multi	No	Yes	No	No	Yes
Vietnamese ASR benchmarks [1, 15]	Vietnamese	No	No	Partial	No	Partial
This work	Vietnamese	Yes	Yes	Yes	Yes	Yes

3 Evaluation Pipeline and Setup

The evaluation pipeline aims to support reproducible streaming ASR experiments. It includes four stages: dataset preparation, streaming simulation and inference, metric computation, and analysis. Figure 1 shows the overall structure.

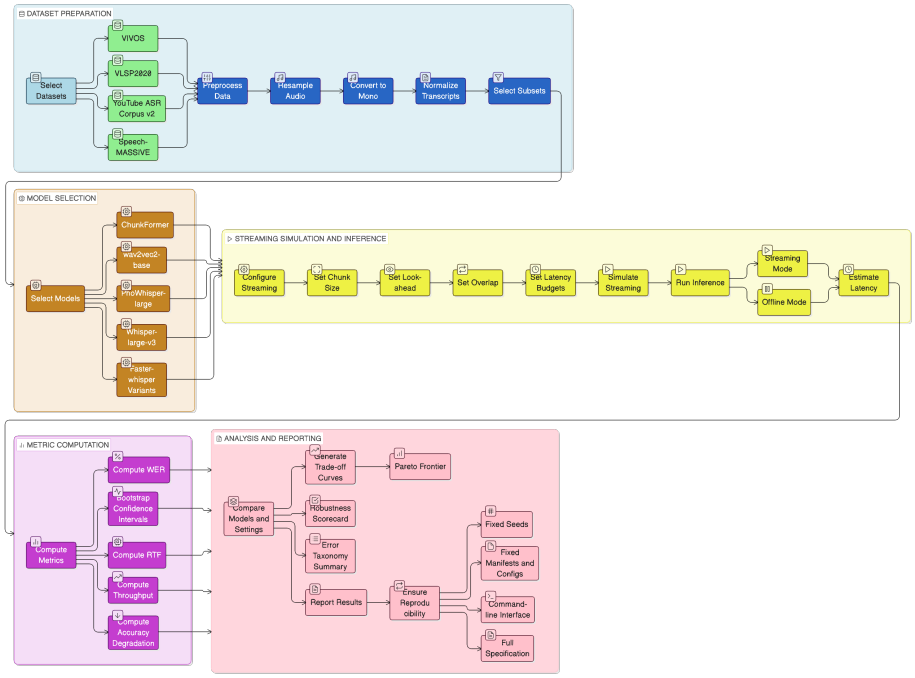

Fig. 1. Overview of the evaluation pipeline.

We evaluate models on four public Vietnamese speech datasets. VIVOS is used as a clean read-speech baseline [1]. VLSP2020 includes spontaneous speech and represents domain shift [15]. The Vietnamese YouTube ASR Corpus v2 contains in-the-wild recordings across different domains [11]. Speech-MASSIVE (Vietnamese) consists of short assistant-style utterances with strict latency requirements [9].

For latency-sweep experiments, we use fixed 300-utterance subsets per dataset. All audio is resampled to a common rate and converted to mono. Reference transcripts are normalized using a unified procedure.

We evaluate several ASR models with public checkpoints. ChunkFormer is used as the main streaming model in our analysis. wav2vec2-base (Vietnamese, 250h) is used as a lightweight CTC-based baseline amenable to streaming simulation. PhoWhisper-large represents a strong Vietnamese-specific offline upper bound, while Whisper-large-v3 serves as a multilingual baseline. Deployment-oriented variants based on faster-whisper are left as future work due to environment/toolchain compatibility constraints in this run. No model is retrained or fine-tuned; all results are obtained through inference-only evaluation to isolate the effects of streaming constraints and evaluation protocol choices.

Streaming inference is simulated by segmenting input audio into fixed-size chunks and restricting right context using a configurable look-ahead window. For streaming-capable models (ChunkFormer and wav2vec2), we evaluate three deployment-oriented profiles: 1200 ms, 2400 ms, and 4000 ms chunk size, each with 200 ms overlap and 0 ms look-ahead. We apply the same chunking and overlap settings to all streaming models. Offline decoding with full context is used as reference. With 0 ms look-ahead, the latency proxy equals the chunk size.

We use word error rate (WER) as the main accuracy metric and report 95% bootstrap confidence intervals [4, 12]. Runtime efficiency is measured by real-time factor (RTF), defined as inference time divided by audio duration, and we report throughput when available. Streaming degradation is the difference between streaming and offline WER. All runtime numbers are measured on the same hardware.

We document the dataset subsets, streaming settings, normalization rules, and inference scripts used in our experiments. Runs use fixed random seeds and can be reproduced with a single command. The repository also supports rerunning the benchmark with new models or datasets.

4 Results and Analysis

Table 2 lists WER for each model on each dataset under offline and streaming decoding. We report 95% bootstrap confidence intervals. Offline results provide the full-context baseline, and the gap to streaming varies across datasets and model families.

PhoWhisper-large has low WER on VIVOS but degrades on VLSP2020 (Table 2). This pattern is consistent with a domain shift from read speech to spon-

taneous, multi-speaker audio, and it may also be affected by decoding choices and acoustic variability.

Table 2. Overall WER (%) with 95% confidence intervals. Lower is better. N/A denotes offline-only models.

Model	Dataset	Offline WER (95% CI)	Streaming WER (95% CI)	ΔWER
ChunkFormer-CTC-large (110M)	VIVOS	3.44 [2.68, 4.39]	6.17 [4.85, 7.57]	2.74
	VLSP2020	3.83 [3.26, 4.45]	50.64 [46.27, 54.42]	46.81
	Viet YouTube ASR v2	4.48 [3.83, 5.18]	5.44 [4.45, 6.83]	0.96
	Speech-MASSIVE_vie	18.78 [15.49, 22.28]	21.38 [17.75, 24.98]	2.60
wav2vec2-base-vi (95M)	VIVOS	8.88 [7.69, 10.22]	12.32 [10.54, 13.95]	3.44
	VLSP2020	10.09 [9.14, 11.35]	52.74 [49.03, 56.17]	42.65
	Viet YouTube ASR v2	7.90 [6.81, 8.86]	8.31 [7.02, 9.51]	0.40
	Speech-MASSIVE_vie	27.71 [24.12, 31.29]	30.38 [26.95, 34.14]	2.68
PhoWhisper-large (1.55B; offline)	VIVOS	3.01 [2.32, 3.80]	N/A	N/A
	VLSP2020	17.49 [13.74, 21.60]	N/A	N/A
	Viet YouTube ASR v2	10.55 [8.80, 12.69]	N/A	N/A
	Speech-MASSIVE_vie	17.85 [14.60, 20.78]	N/A	N/A
Whisper-large-v3 (1.55B; offline)	VIVOS	9.49 [8.27, 11.11]	N/A	N/A
	VLSP2020	24.91 [21.89, 28.25]	N/A	N/A
	Viet YouTube ASR v2	36.27 [30.08, 42.86]	N/A	N/A
	Speech-MASSIVE_vie	22.39 [18.98, 25.82]	N/A	N/A

Table 3 reports real-time factor (RTF) and throughput measured under the same hardware setup. Larger models achieve lower offline WER, but require more computation. Deployment decisions should consider this trade-off.

Table 3. Runtime efficiency measured by real-time factor (RTF) and throughput on the same hardware. Lower RTF and higher throughput indicate better performance.

Model	RTF (median [IQR])	Throughput (xRT)	Hardware
ChunkFormer-CTC-large (110M)	0.0111 [0.0016]	90.1	1×NVIDIA H100 80GB
wav2vec2-base-vi (95M)	0.0059 [0.0009]	169.5	1×NVIDIA H100 80GB
PhoWhisper-large (1.55B)	0.1034 [0.0170]	9.7	1×NVIDIA H100 80GB
Whisper-large-v3 (1.55B)	0.0903 [0.0186]	11.1	1×NVIDIA H100 80GB

Figure 2 shows streaming WER across three latency settings. We use chunk sizes of 1200, 2400, and 4000 ms with 200 ms overlap and no look-ahead.

4.1 Latency–Accuracy Frontier Analysis

Figure 2 shows that increasing chunk size reduces streaming WER for both models. From 1200 ms to 4000 ms, ChunkFormer improves from 34.8% to 20.9%, and wav2vec2 improves from 43.5% to 25.9%. Latency setting has a strong impact on performance.

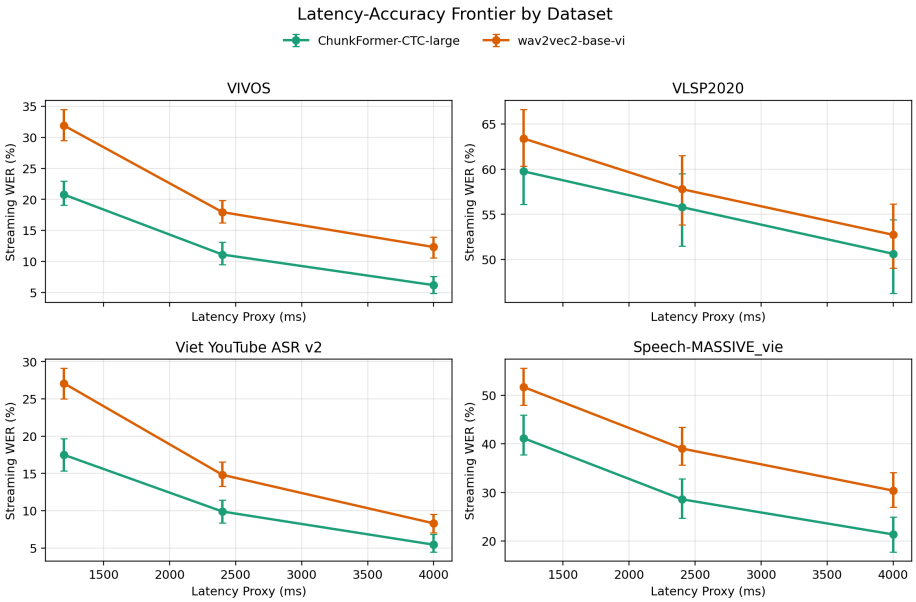

Fig. 2. Streaming WER with 95% CI across 1200, 2400, and 4000 ms settings.

Figure 3 shows ΔWER across the same latency settings. For both models, ΔWER decreases as chunk size increases. Larger chunks provide more context. VLSP2020 remains the most difficult dataset. Degradation is still large at 4000 ms.

4.2 VLSP2020 Failure Analysis

The degradation on VLSP2020 is consistent across settings. At 4000 ms, both models still exceed 50% streaming WER, while offline WER remains low. VLSP2020 also shows more chunks per utterance and a higher change rate than other datasets. At 1200 ms, the number of chunks increases and the change rate rises further. These results point to unstable decoding under multi-chunk processing. This instability poses a risk for long or variable utterances.

4.3 Statistical Significance of ΔWER

We perform paired bootstrap tests over utterances ($n = 5000$ resamples per run) to verify whether streaming degradation is statistically different from zero. Across the 8 streaming runs (4 datasets $\times$ 2 models), 7 runs show significant degradation with one-sided $p < 0.05$. The only non-significant case is *viet_youtube_asr_v2_300 + wav2vec2-base-vi*, where ΔWER is small (0.40 percentage points), the 95% CI crosses zero ([-0.59, 1.62] pp), and $p = 0.2628$. All but one ΔWER cases are statistically significant.

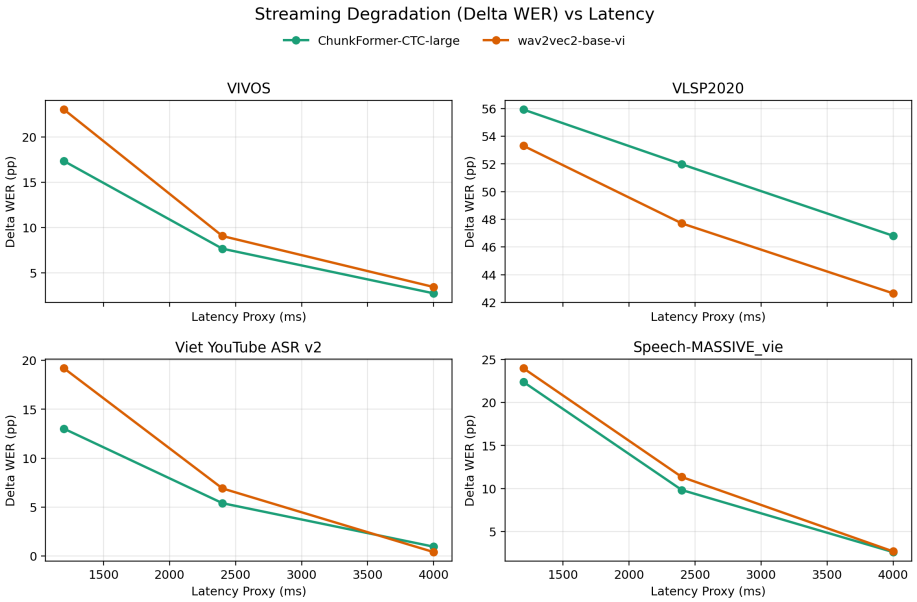

Fig. 3. Latency sweep for streaming degradation (ΔWER in percentage points). Lower is better.

4.4 Streaming Stability

Following prior work on incremental and streaming ASR evaluation that emphasizes hypothesis stability under partial decoding, we analyze incremental-hypothesis stability using change rate and edit-overhead metrics [17, 13]. Change rate measures the proportion of tokens revised between successive partial hypotheses, while edit-overhead quantifies the cumulative edit operations required to reach the final hypothesis relative to its length. Figure 5 shows that stability improves markedly as latency increases: mean change rate drops from approximately 81% (1200 ms) to 31% (4000 ms), and mean edit overhead drops from approximately 1.55 to 0.44. VLSP2020 is consistently the least stable condition, with change rates around 0.88–0.92 at 1200/2400 ms and still around 0.80 at 4000 ms.

Table 4 reports accuracy degradation under distribution shifts, including spontaneous speech, in-the-wild recordings, and assistant-style utterances. For each condition, we report both absolute and relative shifts against the clean read-speech baseline (VIVOS) at 4000 ms. Reporting both metrics avoids overemphasis from ratio-only views when the clean baseline is small.

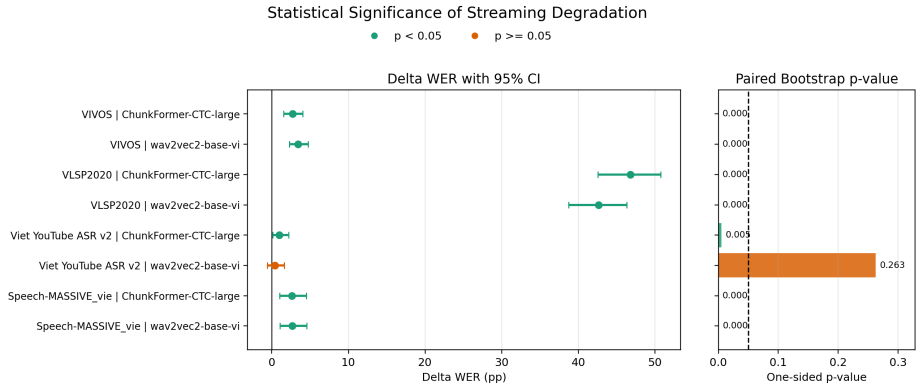

Fig. 4. Paired-bootstrap significance summary of streaming degradation. Left: ΔWER (pp) with 95% CI. Right: one-sided p-values with threshold $p = 0.05$.

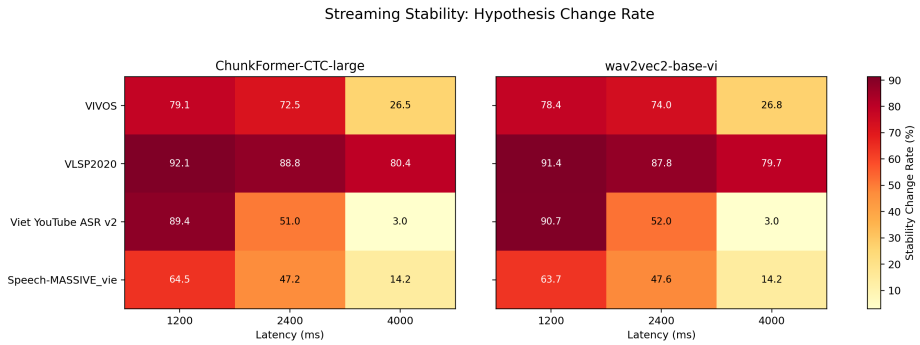

Fig. 5. Streaming stability heatmap (hypothesis change rate, %) across datasets and latency profiles.

4.5 Robustness with Absolute and Relative Shift

To make robustness comparison explicit, we define a model-wise robustness index at fixed latency:

$$\mathrm{RI}(d, m) = \frac{\mathrm{WER}_{\mathrm{stream}}(d, m)}{\mathrm{WER}_{\mathrm{stream}}(\mathrm{VIVOS}, m)}.$$

This is equivalent to $1 + \frac{\text{Relative Shift}}{100}$, and complements absolute shifts (percentage points).

Based on the observed latency-sweep results, we derive deployment-oriented insights. ChunkFormer consistently outperforms wav2vec2 across all datasets and all three latency profiles, but both models degrade substantially on VLSP2020. Streaming evaluation under fixed latency settings should be reported together with offline WER.

Table 4. Robustness scorecard at 4000 ms streaming profile. Absolute shift is in percentage points (pp) from VIVOS. Relative shift is % change from VIVOS; negative values indicate improvement.

Model	Dataset	WER (%)	Absolute Shift (pp)	Relative Shift (%)
ChunkFormer-CTC-large	VIVOS	6.17	0.00	0.0
	VLSP2020	50.64	+44.46	+720.2
	Viet YouTube ASR v2	5.44	-0.73	-11.9
	Speech-MASSIVE_vie	21.38	+15.21	+246.3
wav2vec2-base-vi	VIVOS	12.32	0.00	0.0
	VLSP2020	52.74	+40.42	+328.2
	Viet YouTube ASR v2	8.31	-4.01	-32.6
	Speech-MASSIVE_vie	30.38	+18.07	+146.7

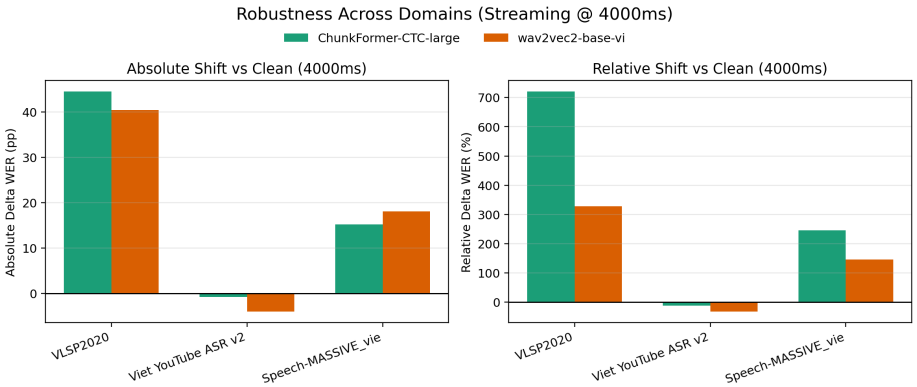

Fig. 6. Robustness comparison at 4000 ms profile: absolute and relative shifts from VIVOS.

5 Conclusion and Future Work

We propose a system-oriented benchmark for Vietnamese streaming ASR that moves beyond offline accuracy-centric evaluation. We introduce a unified evaluation pipeline for streaming ASR. We evaluate models under practical latency constraints and report both accuracy and runtime cost. Our results demonstrate that model selection based solely on offline WER can be misleading in deployment scenarios, as streaming constraints and runtime cost substantially alter the relative performance of different model families.

Through latency–profile comparisons of WER, ΔWER, and RTF, we identify configurations with favorable accuracy–efficiency trade–offs. In addition, the proposed robustness scorecard reveals that streaming constraints often amplify performance degradation under distribution shifts, including spontaneous speech and in-the-wild conditions that are common in Vietnamese applications. Our preliminary Vietnamese-specific error taxonomy (numerals, abbreviations, named

entities, and code-switching) exposes likely failure patterns and motivates future quantitative error breakdown.

This work focuses on evaluation rather than model training, and several limitations remain. We do not retrain or fine-tune models under streaming constraints, and robustness analysis is limited to a fixed set of datasets and stress conditions. Although we evaluate three practical latency profiles, we do not claim an exhaustive continuous latency–accuracy frontier. RTF results are measured on a single H100 80GB setup and should be interpreted as throughput upper bounds rather than edge-deployment guarantees. Future work may extend this benchmark by incorporating additional streaming architectures, low-resource adaptation techniques, explicit latency-budget sweeps, and more fine-grained latency measurements that account for end-to-end system effects. We also plan to expand the robustness analysis to cover a broader range of acoustic and linguistic variations, and to integrate the benchmark into an open, continuously updated evaluation platform for Vietnamese ASR.

By providing an open, reproducible, and deployment-aligned evaluation framework, we hope this work serves as a reference point for future research and practical system development in Vietnamese streaming ASR.

Acknowledgement

This research is funded by the University of Economics Ho Chi Minh City (UEH).

References

1. AILAB, VNUHCM-US: Vivos: Vietnamese speech corpus for automatic speech recognition (2016). https://doi.org/10.5281/zenodo.7068130, https://zenodo.org/records/7068130
2. Barker, J., Marxer, R., Vincent, E., Watanabe, S.: The third CHiME speech separation and recognition challenge: Analysis and outcomes. Computer Speech & Language **46**, 605–626 (2017). https://doi.org/10.1016/j.csl.2016.10.005, https://doi.org/10.1016/j.csl.2016.10.005
3. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C.: Automatic speech recognition and speech variability: A review. Speech Communication **49**(10–11), 763–786 (2007). https://doi.org/10.1016/j.specom.2007.02.006, https://doi.org/10.1016/j.specom.2007.02.006
4. Bisani, M., Ney, H.: Bootstrap estimates for confidence intervals in ASR performance evaluation. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). vol. 1, pp. 409–412 (2004). https://doi.org/10.1109/ICASSP.2004.1326009, https://doi.org/10.1109/ICASSP.2004.1326009
5. Dinh, N.V., Dang, T.C., Nguyen, L.T., Nguyen, K.V.: Multi-dialect vietnamese: Task, dataset, baseline models and challenges. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7476–7498 (2024), https://aclanthology.org/2024.emnlp-main.426/

6. Hai, D.V., Viet, N.V., Anh, N.N., et al.: Asr challenge: Vietnamese automatic speech recognition. VNU Journal of Science: Computer Science and Communication Engineering **38**(1), 1–9 (2022). https://doi.org/10.25073/2588-1086/vnucsce.356, https://js.vnu.edu.vn/CSC/article/view/356
7. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.: Audio augmentation for speech recognition. In: Proc. Interspeech 2015. pp. 3586–3590 (2015). https://doi.org/10.21437/Interspeech.2015-711, https://www.isca-archive.org/interspeech_2015/ko15_interspeech.html
8. Le, T., et al.: Phowhisper: Automatic speech recognition for vietnamese. arXiv preprint (2024), https://arxiv.org/abs/2406.02555
9. Lee, B., et al.: Speech-massive: A multilingual speech dataset for slu and beyond. arXiv preprint (2024), https://arxiv.org/abs/2408.03900
10. Liang, Z., Song, Z., Ma, Z., Du, C., Yu, K., Chen, X.: Improving code-switching and name entity recognition in ASR with speech editing based data augmentation. In: Proc. Interspeech 2023. pp. 919–923 (2023). https://doi.org/10.21437/Interspeech.2023-923, https://www.isca-archive.org/interspeech_2023/liang23b_interspeech.html
11. linhtran92: viet_youtube_asr_corpus_v2 (dataset). Hugging Face Datasets (2024), https://huggingface.co/datasets/linhtran92/viet_youtube_asr_corpus_v2, accessed 2026-02-26
12. Liu, Z., Peng, F.: Statistical testing on ASR performance via block-wise bootstrap. In: Proc. Interspeech 2020. pp. 596–600 (2020). https://doi.org/10.21437/Interspeech.2020-1338, https://www.isca-archive.org/interspeech_2020/liu20c_interspeech.html
13. Ma, Y., et al.: Emformer: Efficient memory transformer for streaming asr. In: Proc. Interspeech (2020)
14. Nguyen, B., Nguyen, V.B.H., Nguyen, H., Phuong, P.N., Nguyen, T.L., Do, Q.T., Mai, L.C.: Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging. arXiv (2019). https://doi.org/10.48550/arXiv.1908.02404, https://arxiv.org/abs/1908.02404
15. Nguyen, H.T.M., et al.: Asr challenge: Vietnamese automatic speech recognition (vlsp 2020-asr). Journal of Computer Science and Cybernetics (2022)
16. Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., Kahn, J., et al.: Scaling up online speech recognition using convnets. In: Proc. Interspeech (2020), https://arxiv.org/abs/2001.03031
17. Selfridge, E., Arizmendi, I., Heeman, P.: Stability and accuracy in incremental speech recognition. In: Proc. SIGDIAL (2011)
18. Sproat, R., Black, A.W., Chen, S., Kumar, S., Ostendorf, M., Richards, C.: Normalization of non-standard words. Computer Speech & Language **15**(3), 287–333 (2001). https://doi.org/10.1006/csla.2001.0169, https://doi.org/10.1006/csla.2001.0169
19. Srivastav, V., et al.: Open asr leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation. arXiv preprint (2025), https://arxiv.org/abs/2510.06961
20. Yao, Z., Wu, D., Wang, X., Zhang, B., Yu, F., Yang, C., Peng, Z., Chen, X., Xie, L., Lei, X.: Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In: Proc. Interspeech (2021), https://arxiv.org/abs/2102.01547
21. Zhang, H., Sproat, R., Ng, A.H., Stahlberg, F., Peng, X., Gorman, K., Roark, B.: Neural models of text normalization for speech applications. Computational Lin-

guistics **45**(2), 293–337 (2019). https://doi.org/10.1162/COLI_a_00349, https://doi.org/10.1162/COLI_a_00349

22. Zhang, X., et al.: Benchmarking lf-mmi, ctc and rnn-t criteria for streaming asr. arXiv preprint (2020), https://arxiv.org/abs/2011.04785
23. Zhuo, J., Yang, Y., Shao, Y., Xu, Y., Yu, D., Yu, K., Chen, X.: Vietasr: Achieving industry-level vietnamese asr with 50-hour labeled data and large-scale speech pretraining. In: Proc. Interspeech 2025. pp. 1163–1167 (2025). https://doi.org/10.21437/Interspeech.2025-398, https://www.isca-archive.org/interspeech_2025/zhuo25_interspeech.html