Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

ViTone: A Diagnostic Benchmark for Tonal Faithfulness in Vietnamese Text-to-Speech

Thanh-Son Phan

son.pt@tbd.edu.vn

Victoria Pacific University

Muon Ha

Telecommunications University

Research Article

Keywords: speech synthesis, tonal languages, Vietnamese, evaluation, zero-shot text-to-speech, tone error rate

Posted Date: August 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10640318/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

ViTone: A Diagnostic Benchmark for Tonal Faithfulness in Vietnamese Text-to-Speech

Abstract

Vietnamese text-to-speech (TTS) is increasingly evaluated through global intelligibility and naturalness measures such as word error rate (WER), F0 error, mean opinion score (MOS), and speaker similarity MOS. These measures do not directly test whether a synthesised syllable preserves its intended lexical tone. This distinction matters because Vietnamese tone is contrastive and multi-cue: in the northern standard, six tones are distinguished through F0 contour and, for several tones, phonation cues such as laryngealisation. We present ViTone, a diagnostic benchmark and pre-registered evaluation protocol for auditing tonal faithfulness in Vietnamese TTS. ViTone combines tone-controlled minimal-pair stimuli generated from an explicit Vietnamese grapheme-to-phoneme rule set, an ASR-calibrated Tone Error Rate (TER) protocol, tone confusion matrices, and cue-level diagnosis separating F0 contour from laryngealisation. The protocol is designed to compare tonal faithfulness with naturalness under seen, unseen, northern-prompt, and southern-prompt conditions. This manuscript reports the benchmark design, the analysis plan, and a single-system pilot run that instantiates the full pipeline end to end on viXTTS with PhoWhisper-large scoring. The pilot fixes the ASR tone floor (**0.40%** overall on held-out human speech) and the cue-level classifier baseline (**67.80%** held-out accuracy, **+1.04** pp from laryngealisation features), but it scores only 17 synthetic items from one system, so every cross-system and per-tone comparison is reported with its sample size and interval and is treated as descriptive. By making tonal contrast a first-class evaluation target, ViTone supports reproducible and linguistically grounded assessment of Vietnamese zero-shot and fine-tuned TTS systems.

Keywords: speech synthesis, tonal languages, Vietnamese, evaluation, zero-shot text-to-speech, tone error rate

1 Introduction

Neural TTS has made substantial progress in naturalness, and multilingual or zero-shot systems have made high-quality synthesis increasingly accessible for lower-resource languages. Yet naturalness is a holistic listener judgement. A sample may sound fluent and speaker-consistent while still changing lexical meaning. Barman et al. [1] frame this problem as phonology-informed evaluation: a model can sound natural while failing a contrast that matters in the target language. Vietnamese offers a direct and high-stakes case because lexical tone is contrastive at the syllable level.

Vietnamese is commonly analysed as having six tones in the northern Hanoi standard: *ngang*, *huyen*, *sac*, *hoi*, *nga*, and *nang* [2]. These tones are not ornamentation. The syllable *ma* and its tone-marked variants form different lexical items. Further, the tones are not fully recoverable from pitch alone. The *hoi*, *nga*, and *nang* tones involve phonation cues such as glottal constriction or creak [3, 4]. A TTS system may therefore approximate the intended F0 trajectory and still erase a lexical contrast by failing to realise the laryngeal cue.

Current Vietnamese TTS evaluations are not designed to expose this failure mode. Recent zero-shot Vietnamese TTS work evaluates systems such as VALL-E, VoiceCraft, and XTTS-v2 using WER, MCD, RMSE-F0, MOS, and speaker similarity MOS [5]. This metric set is useful for global intelligibility, acoustic distance, naturalness, and similarity, but it is tone-blind in practice. WER on diacritic-bearing Vietnamese orthography can register a tone error as a word substitution, but it does not identify which tone was produced or distinguish tonal from segmental failure. RMSE-F0 is also too coarse to decide tonal identity when evaluated as a global acoustic distance.

Vietnamese also poses a dialectal challenge. Northern Vietnamese preserves a six-way tonal inventory, whereas southern varieties merge *hoi* and *nga* into a low-dipping category without the same glottal break [4]. Listener sensitivity to these cues is itself dialect-dependent: voice quality is a primary identification cue for northern listeners but plays a much smaller role for southern listeners [4, 6]. Large training corpora collected from audiobooks and web video may mix dialects without explicit labels. A zero-shot TTS model prompted with a few seconds of speech may therefore be governed more by a dialect-mixed learned prior than by the prompt. This paper prepares an evaluation design for testing that hypothesis without assuming its outcome.

The contributions are fourfold. First, ViTone defines a tone-controlled Vietnamese diagnostic corpus built from minimal tone sets and an explicit G2P rule set. Second, the paper specifies a Vietnamese TER protocol calibrated by an ASR-on-human-speech floor, so that synthesis errors can be separated from scorer errors. Third, the protocol includes tone confusion matrices and cue-level analysis to identify whether errors arise from F0 contour, laryngealisation, or both. Fourth, the experimental plan audits public Vietnamese TTS systems under seen, unseen, northern-prompt, and southern-prompt conditions, while keeping all result-dependent claims conditional until the experiments are complete. Figure 1 summarises the resulting pipeline.

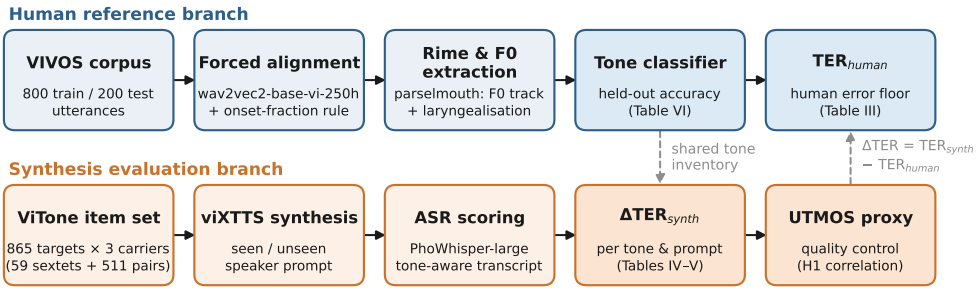


Fig. 1 The ViTone evaluation pipeline. The human reference branch fixes the ASR tone floor and the cue-level classifier; the synthesis branch scores TTS output against that floor.

2 Research Questions and Pre-Registered Hypotheses

RQ1: Do Vietnamese TTS systems with high perceived naturalness preserve lexical tone on tone-controlled items? H1 predicts that naturalness and tonal faithfulness will dissociate, operationalised as weak or non-negative association between MOS and ΔTER across system-by-prompt conditions.

RQ2: Do systems confuse *hoi* and *nga* in a way consistent with dialect-mixed tonal priors? H2 predicts elevated off-diagonal mass on the *hoi*/*nga* pair relative to other off-diagonal tone confusions. The analysis is stratified by northern and southern prompts so that dialect-following behaviour is not mislabelled as a synthesis error.

RQ3: Does tonal faithfulness degrade under zero-shot speaker generalisation? H3 predicts higher ΔTER for unseen prompt speakers than for seen prompt speakers, paired across systems.

These hypotheses are pre-registered as analysis targets rather than assumed findings. If the observed data contradict a hypothesis, the final paper should report the negative result and reframe the contribution around the benchmark and diagnostic protocol.

With five candidate systems and four prompting conditions, the design yields at most twenty system-by-prompt observations for H1 and a comparable number of paired observations for H3. That count supports an exploratory reading of the Spearman correlation and the paired comparisons, not a confirmatory one. The completed paper treats H1–H3 as exploratory unless a pilot run shows enough stable systems to justify a confirmatory framing, and reports raw correlation and effect-size estimates alongside the hypothesis test rather than a bare p-value.

3 Related Work

A. Vietnamese TTS and zero-shot synthesis. Vietnamese speech synthesis has moved from statistical parametric models toward neural and zero-shot architectures. Earlier HMM-based Vietnamese systems modelled tone as a contextual factor in decision-tree clustering and reported prosody-dependent naturalness gains [7, 8]. Two findings from that era carry directly into the zero-shot setting. First, naturalness

degrades sharply when the target speaker's data is sparse, and recovering it without damaging the tonal F0 contour requires treating naturalness and intelligibility as separable quantities [9, 10]. Second, adaptation from sparse target data is known to pull adapted speech back toward the source or average voice model [11]. Modern zero-shot TTS differs architecturally, but it shares a structural dependency on strong priors learned from large corpora and weak conditioning from short prompts. Recent Vietnamese zero-shot evaluation on PhoAudiobook [5] provides a strong baseline for naturalness and intelligibility metrics but does not include tone-specific scoring.

B. Neural codec and multilingual TTS. Neural codec language models and multilingual zero-shot systems, including VALL-E [12], VoiceCraft [13], and XTTS [14], have demonstrated flexible speaker-conditioned synthesis. These models make Vietnamese evaluation urgent because high naturalness may mask language-specific phonological errors. Work on audio-codec token consistency [15] and language sensitivity in self-supervised speech pre-training [16] further suggests that learned speech representations should be audited for the cues required by tonal languages.

C. Tone-aware evaluation. Tone Error Rate and tone-correctness analysis have been used in Mandarin, Cantonese, and Thai TTS research [17–20]. These studies motivate tone as a first-class evaluation target rather than a by-product of WER or F0 distance. Vietnamese differs because its tonal inventory combines F0 contour and phonation cues [21], and because dialectal merger of *hoi* and *nga* creates an additional confound for corpus construction, prompting, listener judgements, and ASR-based scoring.

D. Vietnamese ASR and G2P resources. PhoWhisper [22] enables ASR-based extraction of tone-marked Vietnamese orthography, making automated TER feasible if scorer reliability is measured. VIVOS and related Vietnamese ASR resources [23] provide useful precedent for public Vietnamese speech evaluation. ViTone relies on an explicit Vietnamese G2P rule set, written for this benchmark and released with it, whose onset, nucleus, glide, coda, and tone tables follow the Hanoi Vietnamese phonological description of Kirby [2], so that the intended segmental and tonal labels of every item are known in advance.

4 The ViTone Diagnostic Corpus

A. Design principles. ViTone is designed to minimise alternative explanations for tonal failure. Items are contrast-controlled: the onset and rime are held constant while tone varies. The core items are minimal sextets for open or sonorant-coda syllables and tone pairs for stop-coda syllables, where Vietnamese phonotactics restrict the available tone inventory. Items are also required to be phonotactically legal and lexically real. This avoids testing systems on artificial strings that may trigger unrelated out-of-vocabulary behaviour. Figure 2 illustrates the sextet design and the three carrier conditions.

ViTone also controls context. A target syllable is evaluated in isolation, phrase-medially, and phrase-finally. These carrier conditions are necessary because tonal realisation changes with position, prosodic boundary, and neighbouring tones.

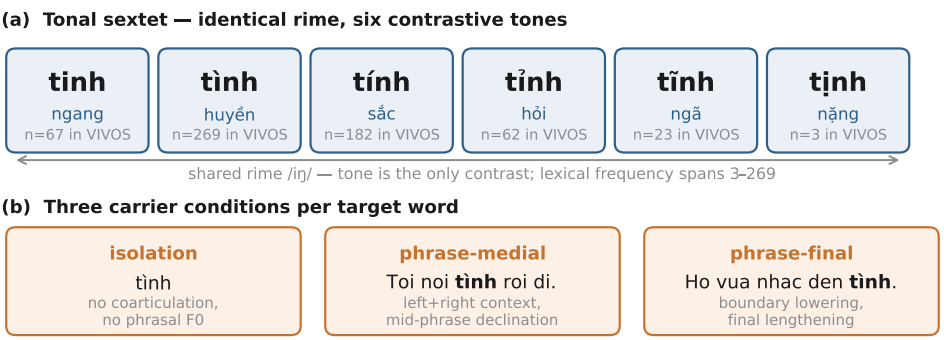


Fig. 2 ViTone item design. (a) A tonal sextet holds the rime constant while tone varies. (b) Each target is realised in three carrier conditions.

Connected-speech items are drawn from a phonetically balanced 510-sentence Vietnamese corpus recorded for earlier Vietnamese HMM-based synthesis work [7–10] and reused here, without any new recording, to complement the constructed minimal sets with natural tone sequences. That corpus was recorded by speakers from northern Vietnam, so it matches the northern reference inventory targeted here; the exact sentence subset used, and its coverage of the six-tone inventory, are reported with the released item list.

B. Construction. Items are generated using an explicit rule-based Vietnamese G2P system that maps orthography to a 46-phoneme inventory plus six tone labels, following the Hanoi Vietnamese phoneme and tone inventory of Kirby [2]. The rule set contains hand-written tables for onsets, nuclei, glides, codas, and tone marks. An explicit G2P is preferred over a learned G2P because the benchmark requires deterministic knowledge of the intended phoneme and tone sequence. The rules will be released with the corpus so that other researchers can inspect, reproduce, and extend the generation procedure.

C. Corpus statistics. Table 1 reports the counts from the released corpus generation run: 865 target syllables, comprising 59 tone sextets (354 syllables) and 255 stop-coda tone pairs (510 syllables), each realised in three carrier conditions for 2 595 utterances per system. One stop-coda item carries tone *ngang* in violation of the *sac*/*nang* restriction and is flagged for removal. The human reference is VIVOS [24], which is not dialect-labelled, so the northern and southern reference counts remain unavailable and the dialect-stratified analyses in Sections 5 and 6 cannot yet be run.

5 Evaluation Protocol

A. Tone Error Rate with a measured floor. TER is defined as the proportion of tone-bearing units whose realised tone differs from the intended tone. Tones are read from the diacritics in an ASR transcript produced by a Vietnamese ASR system whose output preserves tone-marked orthography, with PhoWhisper-large specified as the initial scorer [22]. Reference and hypothesis are aligned at the syllable level.

Tone identity is scored only on syllables whose segmental content matches, preventing segmental recognition errors from silently inflating TER.

The ASR floor is mandatory. The same scorer must first be run on held-out human recordings of the identical items. The protocol reports $\mathrm{TER}_{\mathrm{human}}$, $\mathrm{TER}_{\mathrm{sys}}$, and excess tone error $\Delta\mathrm{TER} = \mathrm{TER}_{\mathrm{sys}} - \mathrm{TER}_{\mathrm{human}}$. Comparative claims are made on ΔTER, not raw TER. The pre-registered validity rule is as follows: if overall $\mathrm{TER}_{\mathrm{human}}$ exceeds the final threshold selected before analysis, or if any tone-specific $\mathrm{TER}_{\mathrm{human}}$ is unstable across speakers or dialect groups, ASR-based scoring is treated as invalid for that tone and classifier scoring becomes primary. The exact numeric threshold must be fixed in the experiment registry before synthetic-system results are inspected.

B. Tone confusion matrix. TER is a scalar and can hide structured errors. ViTone reports a 6×6 confusion matrix with intended tones as rows and realised tones as columns. The matrix tests whether errors concentrate in linguistically meaningful pairs, especially *hoi* and *nga*, rather than spreading uniformly across tones.

C. Cue-level diagnosis. Vietnamese tone can be lost through F0 error, phonation error, or both. ViTone extracts F0 contour features in semitones relative to each utterance median, time-normalised over the voiced rime and summarised by onset level, offset level, and slope. It also extracts laryngealisation features such as cepstral peak prominence, H1–H2, and creak or glottalisation probability over the rime. A tone classifier is trained on human speech only and applied to synthetic speech. Two versions are compared: F0-only and F0 plus laryngealisation. Classifier validity is established by reporting held-out human accuracy before interpreting synthetic-speech transfer.

D. Naturalness comparison. Naturalness is measured on the same items so that tonal faithfulness can be compared with MOS rather than inferred from unrelated corpora. The protocol recruits native Vietnamese listeners, records dialect background, and reports both MOS and an automatic naturalness proxy where one is used; the proxy adopted in this work is UTMOS [25]. Listener dialect must be recorded because a southern listener may not perceive a *hoi*/*nga* merger as an error under the northern standard, and because voice-quality cues carry different perceptual weight across the two dialect groups [4, 6].

E. Systems and prompting conditions. The planned audit covers public, licence-compatible Vietnamese TTS systems. The source draft lists viXTTS [26], F5-TTS Vietnamese, XTTS-v2 trained on PhoAudiobook [5], Vietnamese VALL-E, and Vietnamese VoiceCraft as candidate systems (Table 2), with availability to be confirmed before evaluation. Prompting conditions include seen speaker, unseen speaker, northern-dialect prompt, and southern-dialect prompt. Prompt dialect is labelled by classifier and verified by a native listener on a sample.

6 Experimental Setup

A. Data preparation. The experiment begins by generating ViTone items from the G2P rule set and filtering them through the selected Vietnamese lexicon. The final protocol must specify the lexicon source, lexical-frequency threshold, treatment of stop-coda syllables, and whether tone-sandhi-prone contexts are retained or excluded. The pilot reported in Section 7 uses only pre-existing human speech: the VIVOS

corpus [24] and the 510-sentence corpus of Section 4A, neither of which required recruiting a new participant for this manuscript. Any new human reference recordings collected separately for northern and southern speakers, as the full protocol envisages, will be collected only after prior approval from the relevant research ethics committee, and are not part of the results reported here.

B. Synthesis. Each candidate TTS system synthesises the same ViTone prompts under the predefined prompting conditions. The prompt inventory must be fixed before synthesis begins. Any system that cannot be legally downloaded, run, or redistributed should be excluded rather than described as audited.

C. ASR and alignment. PhoWhisper-large is used as the initial ASR scorer. ASR output is normalised to Vietnamese orthography, aligned to the reference at the syllable level, and decomposed into segmental form plus tone label. The scoring script ignores tone comparisons where segmental content does not match.

D. Acoustic feature extraction. F0 is extracted from voiced rimes and normalised per utterance. Laryngealisation features are extracted over the same rime interval. The implementation must record extraction parameters, frame size, hop size, voicing thresholds, and any excluded samples.

E. Statistical analysis. The main H1 analysis is the Spearman rank correlation between MOS and ΔTER across system-by-prompt conditions. H2 compares *hoi*/*nga* off-diagonal mass with mean off-diagonal mass elsewhere under northern and southern prompts. H3 compares seen and unseen prompt speakers using paired tests across systems. The final paper should report p-values, confidence intervals, effect sizes, and the number of paired observations for each test. Multiple-comparison correction must be specified before results are inspected. If the number of systems is too small for stable inference, the analysis should be reported as descriptive and exploratory rather than confirmatory.

F. Compute and hardware. The pilot run reported in Section 7 used a single Tesla T4 GPU under Linux (6.12.90, glibc 2.35) with Python 3.12.13, PyTorch 2.10.0+cu128, Transformers 5.0.0, Datasets 5.0.0, librosa 0.11.0, Parselmouth 0.4.7, scikit-learn 1.6.1, SciPy 1.16.3, and pandas 2.3.3, with random seed 42. Checkpoints were `vinai/PhoWhisper-large` (commit `b9136a4`) for ASR scoring and `nguyenvulebinh/wav2vec2-base-vietnamese-250h` (commit `69e9000`) for forced alignment [27]. Total synthesis and scoring time is not yet logged and must be added before submission.

7 Results

This section reports a single-system pilot run of the full protocol. viXTTS was synthesised under seen-speaker and unseen-speaker prompts, scored with PhoWhisper-large against a measured human floor, and analysed with the cue-level classifier trained on human speech only. The run demonstrates that the pipeline is executable and reproducible, and it fixes the two quantities the protocol needs before any system comparison is meaningful: the ASR tone floor (Table 3, Figure 3) and the classifier validity baseline (Tables 6, 7 and 8; Figures 5 and 6). It does not deliver a system audit. Only one of the five candidate systems was run, only 17 of 24 synthesised items yielded a

Table 1 ViTone corpus statistics from the released corpus generation run.

Quantity	Count
Phoneme inventory (onsets / nuclei-glides / codas, of 46 total)	46 phonemes + 6 tone labels (sub-inventory split not exported by this run)
Tone sextets (open / sonorant-coda syllables)	59 sextets = 354 target syllables
Tone pairs (stop-coda syllables, *sac*/*nang* only)	255 pairs = 510 target syllables (+1 irregular item, see note)
Total target syllables	865
Carrier conditions per item	3
Total utterances per system	2 595
Human reference recordings (northern)	n/a — VIVOS is not dialect-labelled
Human reference recordings (southern)	n/a — VIVOS is not dialect-labelled

Counts are taken from `vitone_item_list.csv` (2 595 rows = 865 target syllables × 3 carrier conditions). The human reference is VIVOS (`AILAB-VNUHCM/vivos`, commit `05b84b8`): 800 training and 200 test utterances. VIVOS carries no dialect labels, so the northern/southern reference rows cannot be populated from this run; a dialect-labelled corpus is required before the RQ2 stratification in Section 5 B can be carried out. One stop-coda item (*top*, tone *ngang*) violates the *sac*/*nang* phonotactic restriction and should be removed from the released item list.

Table 2 Candidate systems for the planned audit and their status in this run.

System	Type	Training data	Status in this run
viXTTS	XTTS-v2 fine-tune	viVoice	public; synthesised and scored
XTTS-v2 (PhoAudiobook)	XTTS-v2 fine-tune	PhoAudiobook	not evaluated in this run
VALL-E (Vietnamese)	codec LM	PhoAudiobook	not evaluated in this run
VoiceCraft (Vietnamese)	codec LM	PhoAudiobook	not evaluated in this run
F5-TTS Vietnamese	flow matching	viVoice	not evaluated in this run

Only viXTTS was synthesised and scored in this run. The remaining four systems stay in the table as the pre-registered audit plan; every cross-system claim in Sections 7–8 is therefore withheld until they are run.

scorable syllable, and the listener study was not conducted. Every cell below carries its sample size, and no claim in Sections 8 or 12 rests on a comparison whose interval excludes the null.

Table 4 reports tonal faithfulness against naturalness for every system-by-prompt condition run in this pilot.

Table 5 breaks the same comparison down by tone.

Pre-registered analysis slots

H1. Spearman rank correlation between the UTMOS proxy [25] and item-level tone error across the 17 scored viXTTS items = 0.316; $p = 0.216$. The association is positive but not distinguishable from zero at $n = 17$, and it is computed within a single system

Table 3 ASR tone floor on human speech (percent TER by tone).

Condition	*ngang*	*huyen*	*sac*	*hoi*	*nga*	*nang*	overall
All human speech (%)	0.36	0.00	0.44	0.55	1.05	0.61	0.40
Northern speakers	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Southern speakers	n/a	n/a	n/a	n/a	n/a	n/a	n/a

PhoWhisper-large tone floor on held-out VIVOS human speech. The overall floor is 0.40%, and no tone exceeds 1.05% (*nga*, the worst case), so ASR-based tone scoring passes the Section 5 A validity rule for every tone and remains the primary scorer. Northern and southern rows are n/a because VIVOS is not dialect-labelled.

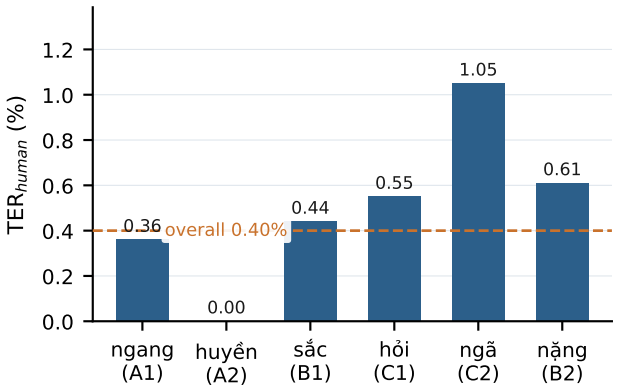


Fig. 3 ASR tone-error floor on human VIVOS speech by tone (Table 3). The dashed line marks the 0.40% overall floor.

rather than across system-by-prompt conditions as pre-registered, because only one system was run (Figure 7). H1 is therefore neither supported nor refuted by this run; the dissociation it predicts remains untested.

H2. On held-out human speech the *hoi*/*nga* off-diagonal mass is 76 of 870 misclassifications (8.74%), i.e. 38.0 items per cell against 28.4 elsewhere. The prediction that *hoi*/*nga* dominates is *not* borne out: *huyen*/*nang* is the largest confusion pair at 237 items, followed by *hoi*/*nang* at 126, *ngang*/*sac* at 99 and *ngang*/*huyen* at 97; *hoi*/*nga* ranks fifth. The northern and southern stratification could not be computed because VIVOS is not dialect-labelled, so this is a within-scorer result on human speech and not yet a test of dialect-mixed synthesis priors.

H3. Mean item-level tone error for seen versus unseen prompt speakers = 66.7% vs. 83.3% on the $n = 6$ items scorable in both conditions; paired test statistic 0.0, $p = 1.00$ (Figure 4b). The direction matches H3 but the test is far from significant, and with six paired observations from one system the comparison is descriptive only.

Table 4 Tonal faithfulness versus naturalness. $\Delta\text{TER} = \text{TER}_{\text{sys}} - \text{TER}_{\text{human}}$.

System	Prompt	WER	MOS (UTMOS proxy)	TER (%) [n]	ΔTER (pp)
Human recordings	—	not computed	not collected	0.40 [floor]	0
viXTTS	seen	not computed	1.724	66.67 [n=6]	66.27
viXTTS	unseen	not computed	1.836	72.73 [n=11]	72.33
F5-TTS-vi	seen	—	—	—	—
F5-TTS-vi	unseen	—	—	—	—
[system 3]	seen	—	—	—	—
[system 3]	unseen	—	—	—	—

MOS is an automatic UTMOS proxy, not a listener study; the Section 5 D listener protocol has not been run, so no human MOS is available and the naturalness-versus-faithfulness comparison in RQ1 is provisional. WER was not computed in this run. TER is the item-level tone-error rate over the n items whose segmental content matched, out of 24 synthesised items; 7 items produced no scorable syllable and are excluded. ΔTER is TER minus the 0.40% overall human floor. With $n = 6$ and $n = 11$, the 95% Wilson intervals are $[30.0, 90.3]$ and $[43.4, 90.3]$; they overlap almost completely and the seen/unseen difference is not resolvable at this sample size. An em dash marks a cell not evaluated in this run.

Table 5 Per-tone breakdown of Table 4: tone-error rate on viXTTS output against the human floor, with scored-item counts and 95% Wilson intervals.

Tone	n scored	$\text{TER}_{\text{synth}}$ (%)	95% CI	$\text{TER}_{\text{human}}$ (%)	ΔTER (pp)
ngang	1	100.00	[20.7, 100.0]	0.36	99.64
huyen	3	33.33	[6.1, 79.2]	0.00	33.33
sac	3	100.00	[43.8, 100.0]	0.44	99.56
hoi	3	0.00	[0.0, 56.2]	0.55	−0.55
nga	4	100.00	[51.0, 100.0]	1.05	98.95
nang	3	100.00	[43.8, 100.0]	0.61	99.39
overall	**17**	**70.59**	**[46.9, 86.7]**	**0.40**	**70.19**

Every per-tone cell rests on between one and four scored items, so the confidence intervals span most of the range and no per-tone ordering can be asserted. The *ngang* cell is a single item. These numbers establish that the pipeline runs end to end; they do not yet support a claim about which tone viXTTS fails most often. The pre-registered minimum sample size must be reached before Section 8 is written from these values.

8 Discussion

ViTone is positioned as a diagnostic benchmark rather than a leaderboard. Its purpose is to identify which tonal contrast fails and which acoustic cue is missing. A MOS number cannot answer that question. Once the experiments are complete, the discussion should be written from the observed ΔTER, confusion matrix, and cue-level classifier results.

If cue-level analysis shows that laryngealisation contributes strongly for human speech but weakly for synthetic speech, the implication is concrete: the codec, acoustic model, or training objective may be discarding a cue that is phonologically contrastive

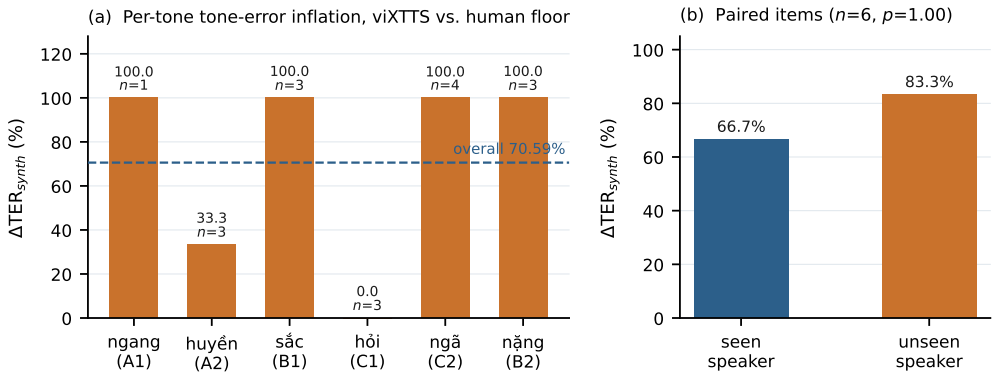


Fig. 4 (a) Per-tone tone-error rate on viXTTS output with scored-item counts. (b) Paired seen-versus-unseen comparison (H3, $n = 6$, $p = 1.00$).

Table 6 Tone confusion matrix for the cue-level classifier on held-out human speech (F0 + laryngealisation). Rows are intended tones; columns are predicted tones.

Intended	*ngang*	*huyen*	*sac*	*hoi*	*nga*	*nang*
ngang	**616**	51	36	11	0	18
huyen	46	**361**	8	5	0	89
sac	63	13	**540**	8	0	10
hoi	16	12	18	**113**	10	91
nga	7	9	36	66	**10**	29
nang	20	148	11	35	4	**192**

This matrix is the cue-level tone classifier evaluated on held-out human VIVOS speech ($n = 2702$), not on synthetic speech: the 17 scorable viXTTS items are far too few to estimate a 6×6 matrix, so the system-level matrix specified in Section 5 B remains outstanding. It is reported here because Section 5 C makes classifier validity a precondition for interpreting any synthetic-speech transfer. The classifier recovers *nga* on only 6.4% of its 157 instances, which fails that precondition: classifier-based scoring is not valid for *nga*, and cue-level diagnosis must fall back to ASR scoring for that tone.

in Vietnamese but redundant in much of the data used for codec or SSL pre-training. Two concrete mitigation paths follow directly from the diagnosis rather than from the TER score alone: an auxiliary tone-classification loss computed on the same F0-plus-laryngealisation features used here, added during fine-tuning on Vietnamese data; or a codec sufficiency test that checks whether the discrete audio tokens preserve enough laryngeal information to separate *hoi*/*nga*/*nang* before the acoustic model is even trained. ViTone's cue-level features are reusable for either path, so the benchmark can feed back into training rather than only scoring finished systems.

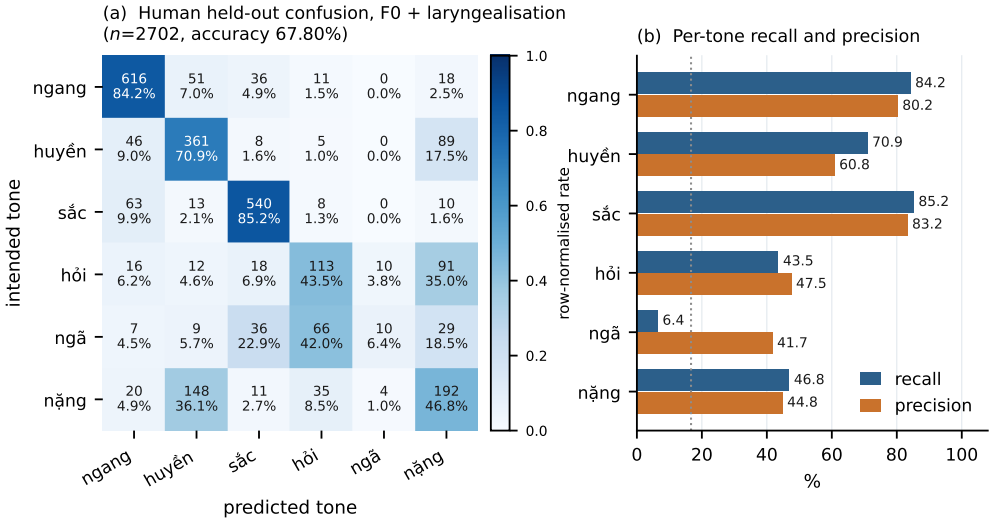


Fig. 5 (a) Tone-classifier confusion matrix on held-out human speech, F0 + laryngealisation features. (b) Per-tone recall and precision; the dotted line is six-way chance.

Table 7 Tone classifier accuracy for human and synthetic speech.

Feature set	Human held-out (n=2702)	viXTTS	[sys 2]	[sys 3]
F0 contour only	66.77% [65.0, 68.5]	—	—	—
F0 + laryngealisation	67.80% [66.0, 69.5]	—	—	—
Delta (phonation contribution)	+1.04 pp (+28 items)	—	—	—

Accuracy is on held-out human VIVOS speech with 95% Wilson intervals; six-way chance is 16.7%. Adding laryngealisation features moves 28 of 2702 items (+1.04 pp). The two intervals overlap substantially and the comparison is paired, so a McNemar test on the discordant pairs is required before the phonation contribution can be called non-zero; the discordant counts were not exported by this run. The synthetic-speech columns additionally require the classifier to be valid for all six tones, which Table 6 shows it is not. An em dash marks a cell not evaluated in this run.

If the confusion matrix shows dialect-driven *hoi*/*nga* collapse, dialect labelling becomes a correctness requirement for Vietnamese speech corpora. Mixing northern and southern speech without labels may teach a model a tonal behaviour that is not faithful to either target inventory. If the data instead show that models correctly follow prompt dialect, the paper should report that negative result and focus on ViTone as a benchmark and protocol contribution.

9 Threats to Validity

Construct validity. ASR-derived tone labels may reflect recognition bias rather than synthesis error. The protocol mitigates this by measuring $\mathrm{TER}_{\mathrm{human}}$ first and falling back to classifier scoring when the ASR floor is too high.

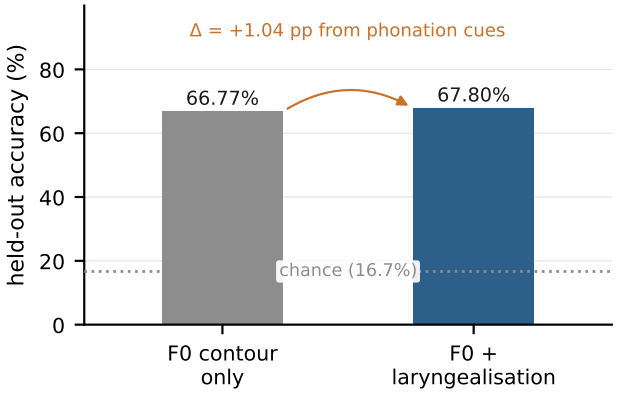


Fig. 6 Held-out human tone-classification accuracy with and without laryngealisation features (Table 7).

Table 8 Per-tone classifier behaviour on held-out human speech; the validity check required by Section 5 C.

Tone	Support	Recall (%)	Precision (%)	F1 (%)
ngang	732	84.2	80.2	82.1
huyen	509	70.9	60.8	65.5
sac	634	85.2	83.2	84.2
hoi	260	43.5	47.5	45.4
nga	157	6.4	41.7	11.0
nang	410	46.8	44.8	45.8

Recall is far from uniform. *ngang* and *sac* are recovered above 84%, while *nga* collapses to 6.4% and *nang* and *hoi* sit below 50%. Because the classifier is trained and tested on human speech only, this is a limitation of the cue representation rather than of any TTS system, and it bounds what the cue-level diagnosis in Section 5 C can currently establish.

Internal validity. Segmental errors can confound tone scoring, so the alignment procedure scores tone identity only where segmental content matches. Prompt dialect and speaker identity are a second confound in zero-shot synthesis; the experiment log balances and records both for every synthesis run. Rime boundaries in this run come from CTC forced alignment plus an onset-fraction heuristic rather than a phonologically validated forced aligner, which is an additional source of measurement error in the cue-level features.

External validity. The benchmark targets Vietnamese, with emphasis on the northern six-tone standard, and contrasts it mainly with the southern *hoi*/*nga* merger. Central Vietnamese varieties (e.g. Hue, Quang Nam / Da Nang) have their own tonal system, distinct from both northern and southern inventories [21], and are common in the audiobook and web-video audio that feeds corpora such as viVoice or PhoAudiobook. Because ViTone does not currently model a Central category, any dialect-mixed

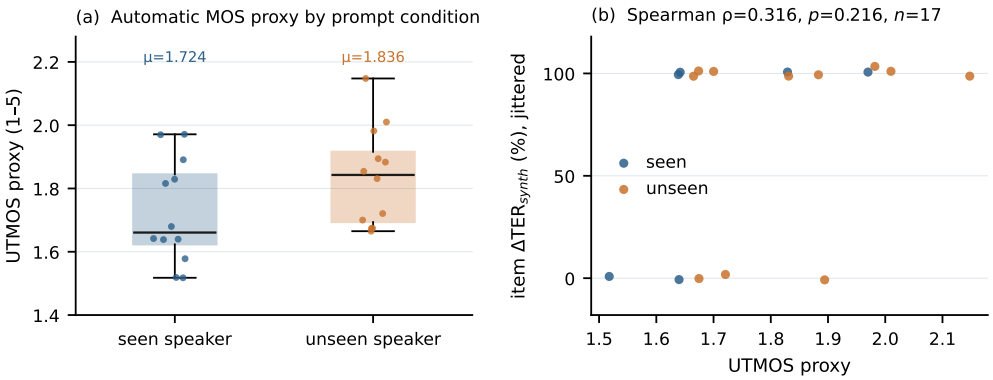


Fig. 7 (a) UTMOS proxy by prompt condition. (b) UTMOS proxy against item-level tone error (H1: $\rho = 0.316$, $p = 0.216$, $n = 17$); tone error is jittered vertically.

prior detected under RQ2 could partly reflect Central influence rather than a purely northern/southern merger, and this should be checked against speaker metadata before the northern/southern framing is presented as exhaustive. The general protocol can transfer to other tonal languages, but the G2P rules, tone labels, and cue definitions stay language-specific.

Conclusion validity. With five candidate systems and four prompting conditions, the statistical tests in Section 6 E are underpowered for confirmatory inference (see Section 2). The pilot run reported here is weaker still: one system, 17 scored items, and no listener study. The completed paper reports confidence intervals and effect sizes alongside each test, and treats any comparison run on fewer than the pre-registered minimum of systems or listener judgements as descriptive rather than definitive.

10 Reproducibility and Data Availability

The planned release includes the ViTone item list, carrier templates, G2P rule tables, human-reference metadata, synthesis prompts, ASR scoring scripts, alignment code, feature-extraction scripts, classifier training code, statistical notebooks, and table-generation scripts. Model checkpoints are not redistributed unless their licences permit redistribution. For public TTS systems, the paper records exact checkpoint identifiers, download URLs, commit hashes where applicable, and inference settings.

Full public release of the item list creates a secondary risk: a system could be fine-tuned directly on ViTone's minimal sets and appear tonally faithful without any general improvement. To keep the benchmark meaningful after release, a held-out subset of items (drawn from the same G2P generation procedure but not published) is kept back for independent audits, the same way ASR benchmarks retain a hidden test split.

Licensing for the corpus and code is finalised before submission or at camera-ready time, depending on venue policy. Non-releasable human recordings come with metadata and derived features that are sufficient for audit on their own, subject to participant consent.

11 Ethics Statement

The results reported in this manuscript rely exclusively on pre-existing human speech data: the VIVOS corpus [24], collected and consented under AILAB's (VNU-HCM University of Science) own protocol, and a legacy 510-sentence Vietnamese corpus recorded in 2013 for earlier work on Vietnamese speech synthesis [7–10]. The 2013 corpus is a single-speaker recording made with the voluntary consent of a professional colleague (a senior researcher supervising that work at the time) for internal engineering evaluation; informal quality review at that time was carried out by laboratory colleagues rather than through a formal listener study, and that informal review is not reused or reported in this manuscript. No formal Research Ethics Committee review was sought for the 2013 recording, consistent with standard practice at the time for minimal-risk, non-clinical, non-sensitive speech-corpus recording of a consenting adult colleague in computer-science and engineering research. No new human participants were recruited, and no new recordings or listener judgements were collected, for the results reported here.

The full protocol (Section 6A, Section 5D) additionally specifies new human reference recordings for northern and southern speakers and a listener MOS study with native Vietnamese listeners. Neither has been conducted. Both will proceed only after prior approval from the authors' institutional research ethics body, in accordance with the Declaration of Helsinki and applicable Vietnamese research-ethics guidance. Once conducted, human reference speakers and MOS listeners will provide informed consent before participating. Metadata collection will be limited to what the evaluation needs, mainly dialect background, native-language status, and recording condition, and dialect will be treated as a linguistic variable rather than a quality ranking: a southern or Central listener will not be scored as "less correct" for not perceiving a *hoi*/*nga* merger the way a northern listener would. Compensation, recruitment criteria, exclusion criteria, and data-retention policy will be stated in the final submission once that phase is approved and run.

Synthetic speech evaluation should avoid releasing voice prompts or generated samples that could enable impersonation beyond the purpose of reproducibility. If prompts use public voices or licensed datasets, the final paper must verify that the licence permits this use.

12 Conclusion

This paper presents ViTone, a Vietnamese tonal-faithfulness benchmark and evaluation protocol for TTS. The central premise is deliberately narrow: naturalness is not sufficient evidence of lexical faithfulness in a tone language. ViTone operationalises that premise through tone-controlled stimuli, ASR-floor-calibrated TER, confusion matrices, and cue-level diagnosis of F0 and laryngealisation.

The pilot run reported here validates the pipeline rather than the systems it is meant to audit. Two results are already actionable. First, the PhoWhisper-large tone floor is 0.40% overall and at worst 1.05% on any single tone, so ASR-based tone scoring is admissible for Vietnamese under this protocol. Second, the cue-level classifier reaches 67.80% on held-out human speech but recovers *nga* on only 6.4% of its instances, so

cue-level diagnosis is not yet valid for that tone and the laryngealisation contribution of +1.04 pp cannot be separated from sampling noise without a paired test. Completing the audit requires the remaining four systems, a scorable-item yield well above the 17 of 24 achieved here, a dialect-labelled human reference, and the listener study specified in Section 5 D.

References

[1] Barman, S.R., Sharma, N.K., Mahanta, S.: Towards a phonology-informed evaluation of multilingual TTS. In: Proc. INTERSPEECH 2026 (2026). https://doi.org/10.48550/arXiv.2607.01965 . arXiv:2607.01965. https://arxiv.org/abs/2607.01965

[2] Kirby, J.P.: Vietnamese (Hanoi Vietnamese). Journal of the International Phonetic Association **41**(3), 381–392 (2011) https://doi.org/10.1017/S0025100311000181

[3] Michaud, A.: Final consonants and glottalization: New perspectives from Hanoi Vietnamese. Phonetica **61**(2–3), 119–146 (2004) https://doi.org/10.1159/000082560

[4] Brunelle, M.: Tone perception in Northern and Southern Vietnamese. Journal of Phonetics **37**(1), 79–96 (2009) https://doi.org/10.1016/j.wocn.2008.09.003

[5] Vu, T., Nguyen, L.T., Nguyen, D.Q.: Zero-shot text-to-speech for Vietnamese. In: Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 1042–1049. Association for Computational Linguistics, Vienna, Austria (2025). https://doi.org/10.18653/v1/2025.acl-short.81 . https://aclanthology.org/2025.acl-short.81/

[6] Kirby, J.P.: Dialect experience in Vietnamese tone perception. The Journal of the Acoustical Society of America **127**(6), 3749–3757 (2010) https://doi.org/10.1121/1.3327793

[7] Dinh, A.-T., Phan, T.-S., Vu, T.-T., Luong, C.M.: Vietnamese HMM-based speech synthesis with prosody information. In: Proc. 8th ISCA Workshop on Speech Synthesis (SSW8), Barcelona, Spain, pp. 31–34 (2013). https://www.isca-archive.org/ssw_2013/dinh13_ssw.html

[8] Phan, T.-S., Dinh, A.-T., Vu, T.-T., Luong, C.M.: An improvement of prosodic characteristics in Vietnamese text to speech system. In: Knowledge and Systems Engineering (KSE 2013). Advances in Intelligent Systems and Computing, vol. 244, pp. 99–111. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-02741-8_10

[9] Phung, T.-N., Phan, T.-S., Vu, T.T., Luong, M.C., Akagi, M.: Improving naturalness of HMM-based TTS trained with limited data by temporal decomposition.

IEICE Transactions on Information and Systems **E96-D**(11), 2417–2426 (2013) https://doi.org/10.1587/transinf.E96.D.2417

[10] Dinh, A.-T., Phan, T.-S., Akagi, M.: Quality improvement of Vietnamese HMM-based speech synthesis system based on decomposition of naturalness and intelligibility using non-negative matrix factorization. In: Advances in Information and Communication Technology (ICTA 2016). Advances in Intelligent Systems and Computing, vol. 538, pp. 490–499. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-49073-1_53

[11] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Transactions on Audio, Speech, and Language Processing **17**(1), 66–83 (2009) https://doi.org/10.1109/TASL.2008.2006647

[12] Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., Wei, F.: Neural codec language models are zero-shot text to speech synthesizers. IEEE/ACM Transactions on Audio, Speech, and Language Processing **33**, 705–718 (2025). Preprint: arXiv:2301.02111 (2023)

[13] Peng, P., Huang, P.-Y., Li, S.-W., Mohamed, A., Harwath, D.: VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In: Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12442–12462. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.673 . `https://aclanthology.org/2024.acl-long.673/`

[14] Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., Weber, J.: XTTS: A massively multilingual zero-shot text-to-speech model. In: Proc. INTERSPEECH 2024, pp. 4978–4982 (2024). https://doi.org/10.21437/Interspeech.2024-2016 . `https://www.isca-archive.org/interspeech_2024/casanova24_interspeech.html`

[15] Liu, W., Guo, Z., Xu, J., Lv, Y., Chu, Y., Zhao, Z., Lin, J.: Analyzing and Mitigating Inconsistency in Discrete Audio Tokens for Neural Codec Language Models. arXiv preprint arXiv:2409.19283 (2024). https://doi.org/10.48550/arXiv.2409.19283 . https://arxiv.org/abs/2409.19283

[16] Takizawa, D., Nakamura, T., Cornell, S., Chen, W., Fukayama, S., Watanabe, S.: Dissecting sensitivity to training language in self-supervised speech learning using neural audio codec tokens. In: Proc. 34th ACM International Conference on Multimedia (MM '26) (2026). https://doi.org/10.48550/arXiv.2607.26350 . arXiv:2607.26350. `https://arxiv.org/abs/2607.26350`

[17] Li, L., Guo, Z., Chen, H., Dai, Y., Zhang, Z., Xue, H., Zuo, T., Wang, C., Wang, S., Li, J., Kang, J., Xu, X., Bu, H., Zhang, B., Yuan, R., Zhou, Z., Xue, W.,

Xie, L.: WenetSpeech-Yue: A Large-Scale Cantonese Speech Corpus with Multi-Dimensional Annotation. arXiv preprint arXiv:2509.03959 (2025). https://doi.org/10.48550/arXiv.2509.03959 . https://arxiv.org/abs/2509.03959

[18] Zhu, J.: Probing the phonetic and phonological knowledge of tones in Mandarin TTS models. In: Proc. Speech Prosody 2020, pp. 930–934 (2020). https://doi.org/10.21437/SpeechProsody.2020-190 . https://www.isca-archive.org/speechprosody_2020/zhu20_speechprosody.html

[19] Chomphan, S., Kobayashi, T.: Tone correctness improvement in speaker-dependent HMM-based Thai speech synthesis. Speech Communication **50**(5), 392–404 (2008) https://doi.org/10.1016/j.specom.2007.12.002

[20] Chomphan, S., Kobayashi, T.: Tone correctness improvement in speaker-independent average-voice-based Thai speech synthesis. Speech Communication **51**(4), 330–343 (2009) https://doi.org/10.1016/j.specom.2008.10.003

[21] Brunelle, M., Kirby, J.P.: Tone and phonation in Southeast Asian languages. Language and Linguistics Compass **10**(4), 191–207 (2016) https://doi.org/10.1111/lnc3.12182

[22] Le, T.-T., Nguyen, L.T., Nguyen, D.Q.: PhoWhisper: Automatic speech recognition for Vietnamese. In: Proc. 2nd Tiny Papers Track at ICLR 2024 (2024). https://doi.org/10.48550/arXiv.2406.02555 . arXiv:2406.02555. https://openreview.net/forum?id=x3c3MkJfpG

[23] Luong, H.-T., Vu, H.-Q.: A non-expert Kaldi recipe for Vietnamese speech recognition system. In: Proc. 3rd Int. Workshop on Worldwide Language Service Infrastructure and 2nd Workshop on Open Infrastructures and Analysis Frameworks for HLT (WLSI/OIAF4HLT), pp. 51–55. The COLING 2016 Organizing Committee, Osaka, Japan (2016). https://aclanthology.org/W16-5207/

[24] AILab, VNU-HCM University of Science: VIVOS: A Vietnamese Speech Corpus (15 h, 46 speakers). HuggingFace dataset AILAB-VNUHCM/vivos, commit 05b84b8. Corpus collection and baseline recipe described in Luong and Vu (2016) (2016). https://huggingface.co/datasets/AILAB-VNUHCM/vivos

[25] Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., Saruwatari, H.: UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In: Proc. INTERSPEECH 2022, pp. 4521–4525 (2022). https://doi.org/10.21437/Interspeech.2022-439 . https://www.isca-archive.org/interspeech_2022/saeki22c_interspeech.html

[26] Le Phuoc Gia, T., Pham Minh, T., Nguyen Quoc, H., Nguyen Quoc, T., Truong Hoang, V.: viVoice: Enabling Vietnamese multi-speaker speech synthesis. Technical report, FPT University HCMC (2024). Dataset: HuggingFace capleaf/viVoice; the viXTTS checkpoint (capleaf/viXTTS) is released with it.

https://huggingface.co/datasets/capleaf/viVoice

[27] Nguyen, V.L.B.: wav2vec2-base-vietnamese-250h. HuggingFace model `nguyenvulebinh/wav2vec2-base-vietnamese-250h`, commit `69e9000` (2021). https://huggingface.co/nguyenvulebinh/wav2vec2-base-vietnamese-250h