# ViInstruct: A Large-Scale Vietnamese Dataset for Instruction-Based Text-to-Speech Synthesis

Thang Ly*, Tri-Nhan Do†, Tu-Anh Nguyen†, Dang-Khoa Mac†
* Faculty of Mathematics and Computer Science, VNUHCM-University of Science, Ho Chi Minh City, Vietnam
† VinFast Joint Stock Company, Hanoi, Vietnam
E-mail: thangquangly0909@gmail.com,v.nhandt23@vinfast.vn,v.anhnct2@vinfast.vn,v.khoamd@vinfast.vn

***Abstract*—Instruction-based text-to-speech (TTS) enables natural-language control of speaking style without reference audio, but no comparable Vietnamese resource exists despite the language's tonal prosody and substantial regional variation. This paper introduces ViInstruct, the first large-scale Vietnamese dataset for instruction-based TTS, comprising approximately 1,718 hours of speech and 1,224,945 utterance-instruction pairs collected from six corpora spanning general speech, regional accent, and emotion settings. We develop an automated pipeline that extracts acoustic features, preserves or infers speaker attributes, maps them to Vietnamese descriptors, and generates one natural-language instruction for each utterance with a large language model. We also provide a standardized benchmark, comprising two evaluation subsets and two baseline systems, to support reproducible evaluation of controllable Vietnamese TTS in a low-resource tonal language.**

***Index Terms*—text-to-speech, instruction-based TTS, Vietnamese, speech dataset, natural language description, controllable speech synthesis**

## I. Introduction

Conventional text-to-speech (TTS) systems typically control speaking style through speaker embeddings or reference recordings, which require users to supply example audio for the desired voice. Instruction-based TTS removes that requirement: target attributes such as gender, age, speaking rate, pitch, accent, and emotion are specified directly in natural language [1], [2].

Substantial progress has been reported for English, from early attribute-conditioned models [1], [3] to large-scale style-captioned resources [4]–[6] and open-vocabulary instruction frameworks [7]. However, these resources remain overwhelmingly English-centric, with only limited instruction-following benchmarks such as InstructTTSEval for English and Chinese [8], leaving low-resource languages without comparable training corpora.

Vietnamese introduces additional difficulties that are not addressed by existing multilingual resources. As a tonal language with six lexical tones, Vietnamese relies on pitch patterns that encode both linguistic content and paralinguistic style. The language also exhibits substantial dialectal variation across three major regions and 63 provinces [9]. Available Vietnamese speech corpora [9]–[14] provide transcripts or discrete categorical labels rather than the natural-language instructions that instruction-based TTS requires.

To address this gap, we present **ViInstruct**, a large-scale Vietnamese instruction-TTS dataset and benchmark designed for controllable speech synthesis research. Our main contributions are as follows:

1) We assemble, to the best of our knowledge, the first large-scale Vietnamese dataset for instruction-based TTS: **1,224,945** utterance-instruction pairs (approximately 1,718 hours) from six source corpora.
2) We design a fully automated annotation pipeline combining acoustic analysis, attribute prediction, quality filtering, quantile binning, and LLM-based instruction generation to produce one Vietnamese instruction per utterance.
3) We establish a reproducible benchmark with adapted Parler-TTS and CapSpeech baselines, showing that controllable Vietnamese TTS is feasible but that robust attribute control—especially for age—remains unsolved.

A public demo page for the dataset is available at https://thangquang09.github.io/viinstruct-demo/.

## II. Related Work

### A. Instruction- and Style-Prompted TTS

Instruction-based speech synthesis has progressed from structured attribute prompting to free-form natural-language control. PromptTTS and PromptTTS 2 [1], [3] showed that style prompts can guide synthesis and be scaled through automatic annotation, while InstructTTS and PromptStyle [2], [15] relaxed prompt structure and improved prompt-to-style alignment. At larger scale, Parler-TTS popularized style-prompted synthesis with descriptive text conditioning, whereas CapSpeech formalized style-captioned TTS datasets and benchmarks [4], [5]. More recent efforts such as OV-InstructTTS and InstructTTSEval [7], [8] move closer to open-vocabulary instruction following.

### B. Style-Captioned Datasets and Annotation Pipelines

Recent style-caption pipelines usually extract acoustic and speaker attributes with signal processing and classifiers, convert them into structured keywords, and then rely on an LLM to generate natural-language descriptions [3], [5], [6]. ParaSpeechCaps and LibriTTS-P [6], [16] exemplify this recipe for English speech, while WavCaps [17] shows

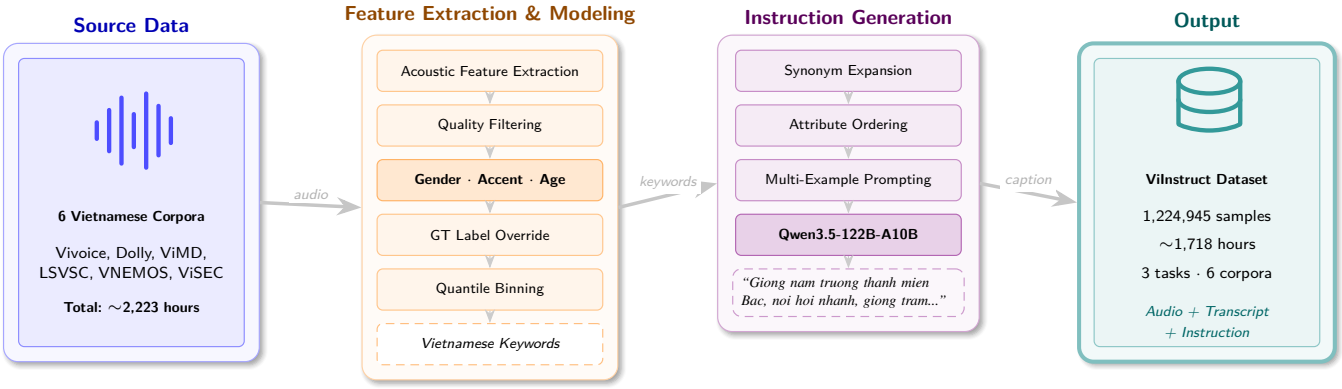


Fig. 1: Overview of the ViInstruct construction pipeline. Raw audio is progressively transformed into instruction–speech pairs through feature extraction, attribute prediction, keyword mapping, and LLM-based instruction generation.

TABLE I: Source corpora and their characteristics. The **Raw** columns report each source corpus as published by its authors; the **Filtered** columns report what ViInstruct retains after quality filtering and deduplication.

| Source | Task | Raw Samples | Filtered Samples | Raw (h) | Filtered (h) |
|---|---|---|---|---|---|
| Vivoice | General | 838,345 | 581,299 | 1,016 | 779 |
| Dolly | General | 664,125 | 584,509 | 1,000 | 808 |
| ViMD | Accent | 18,949 | 12,790 | 103 | 54 |
| LSVSC | Accent | 56,823 | 43,989 | 101 | 75 |
| VNEMOS | Emotion | 250 | 211 | 0.60 | 0.35 |
| ViSEC | Emotion | 5,280 | 2,147 | 3.20 | 1.12 |
| **Total** | | **1,583,772** | **1,224,945** | **2,223** | **1,718** |

that weakly supervised captioning can also scale in general audio. ViInstruct retains this recipe but adapts it to Vietnamese through Vietnamese accent and age prediction, gender-/accent-aware quantile binning, Vietnamese keyword mapping, and Qwen-based instruction generation [18].

### C. Vietnamese Speech Resources

Existing Vietnamese corpora cover ASR, dialect classification, and emotion recognition, including VIVOS and PhoAudiobook for transcription [10], [11], ViMD and LSVSC for dialect research [9], [12], and VNEMOS and ViSEC for emotion [13], [14]. All provide transcripts or discrete metadata; none offers utterance-level free-form Vietnamese instructions jointly encoding gender, age, accent, speaking rate, pitch, loudness, and emotion, which is the gap ViInstruct fills.

## III. Dataset Construction

### A. Source Corpora

ViInstruct integrates six Vietnamese speech corpora spanning three task categories, as summarized in Table I.

The *general* subset (Vivoice [19] + Dolly [20]) contributes approximately 1,587.5 hours of filtered speech with transcripts. Vivoice is read and spontaneous speech collected from public video channels, whereas Dolly is a text-to-speech synthesized corpus of 484 synthetic voice presets; Dolly accounts for 47.7% of the filtered data, and its synthetic origin should be kept in mind when interpreting the benchmark results. The *accent* subset (ViMD + LSVSC) contributes approximately 128.9 hours with reliable regional annotations, while LSVSC also provides ground-truth gender and age labels; ViMD transcripts are obtained with PhoWhisper [21]. The *emotion* subset (VNEMOS + ViSEC) contributes approximately 1.5 hours with source emotion labels; being two orders of magnitude smaller than the other subsets and carrying no emotion-specific adherence metric in the benchmark, it is offered as a proof-of-concept attribute slot rather than as a controllable dimension this paper can validate.

### B. Feature Extraction Pipeline

After source collection, the construction pipeline proceeds through four main stages, as illustrated in Fig. 1.

**Stage 1: Acoustic Feature Extraction and Quality Filtering.** We extract duration, speaking rate (syllables per second using the transcript), pitch statistics (mean, standard deviation, range, coefficient of variation) via the CREPE algorithm [22], and loudness in LUFS. Voice activity detection (Silero VAD [23]) segments speech from silence to compute speech proportion. We then apply four main quality thresholds: duration $\geq 2.0$ s, signal-to-noise ratio (SNR) $\geq 15$ dB, clipping ratio $< 0.01$, and speech proportion $\geq 0.30$. For ViSEC alone, the minimum-duration threshold is relaxed to 1.0 s, since its short acted utterances would otherwise lose a disproportionate share of the already small emotion subset.

**Stage 2: Speaker Attribute Prediction and Ground-Truth Override.** For samples without ground-truth labels, we employ three specialized models: a wav2vec2-based [24] gender classifier[1], a ViMD-based wav2vec2 accent classifier[2], and a publicly released finetuned MERaLiON-3-

[1]https://huggingface.co/thangtrungnguyen/vietnamese-female-male-voice-classification-model

[2]https://huggingface.co/thangquang09/wav2vec2-base-vi-accent-classification

10B age classifier [25], namely `lewiswoncy/m_test_9`[3]. The released age model card reports 91.96% accuracy on CommonVoice Vietnamese age recognition and uses three coarse classes: *teens* (10–19), *adults* (20–59), and *seniors* (60–100). For accent, ViMD's province-level labels (63 provinces) are mapped to North/Central/South, reproducing ViMD's three-region dialect-identification setting; the resulting `wav2vec2-base-vi` classifier is our accent predictor, reaching 94.60% macro-F1 against the 91.02% ViMD reports. Where source corpora provide ground-truth labels (e.g., ViMD accent, LSVSC gender/region/age, VNEMOS and ViSEC emotion), we preserve these annotations to avoid prediction noise. In particular, LSVSC contributes a *children* category absent from the age model's output inventory, resulting in four age groups in the final dataset.

**Stage 3: Feature Binning.** Continuous acoustic features are discretized into Vietnamese keywords via quantile binning, chosen over equal-width binning because the feature distributions are strongly skewed. Pitch mean is binned per gender-accent group, since prosodic ranges differ systematically across those groups; speaking rate, loudness and pitch variation use a single global edge set. Pitch mean and speaking rate are mapped to seven ordinal levels while loudness and pitch variation use five. Bin edges are computed on the largest corpus (Vivoice) and reused for the smaller corpora to keep keywords comparable across sources, so bins are near-uniform within Vivoice but imbalanced in the pooled release.

### C. LLM-Based Instruction Generation

**Stage 4: LLM-Based Instruction Generation.** Following prior work [3], [5], Qwen3.5-122B-A10B [18] is used to generate natural-language instructions from the binned keywords. Each prompt contains the transcript, shuffled attribute tags, and three few-shot examples. To increase lexical variation, four lightweight mechanisms are applied: synonym substitution for Vietnamese keywords, random attribute ordering, diverse few-shot examples, and temperature uniformly sampled per utterance between 0.3 and 0.7. The resulting instruction is stored in the `caption_full` field and serves as the sole natural-language conditioning text in this paper.

**Instruction Quality Audit.** Qwen3.5 is multilingual and supports Vietnamese, but its Vietnamese competence is not separately documented, so instruction quality is audited rather than assumed. Because both the attribute tags and the instruction are text, the faithfulness of the composition step can be measured without listening: every one of the 1,224,945 instructions is matched against its own tags using a Vietnamese descriptor lexicon that covers the generation-time synonyms plus paraphrases observed in the corpus. Some lexical realization of every conditioning tag is recoverable in 81.4% of instructions, the remainder being paraphrase outside the lexicon rather than omission. Where a slot is realized in words, it names the conditioning bin in 96.6–100.0% of cases depending on the slot: gender and accent are at or above 99.99%, loudness is lowest at 96.6%, and 99.7% of all acoustic disagreements are one bin apart, that is, intensity slips rather than category errors. No emotion word occurs in any of the 1,222,587 instructions that carry no emotion tag, so the pipeline does not invent affect. The one systematic defect is age: 4.7% of instructions render the tag người trưởng thành (adult) as lớn tuổi (elderly), two semantically adjacent Vietnamese age expressions, which compounds the coarse age supervision discussed in Section VI. The audit bounds tag-to-text faithfulness only; whether the tags match the audio rests on the ground-truth labels and classifier accuracies above, with perceptual validation left to future work.

[3]https://huggingface.co/lewiswoncy/m_test_9

TABLE II: Example generated instructions for each task category.

| Task | Example Instruction |
|---|---|
| General | Giọng nam trưởng thành miền Nam, nói với nhịp gấp, âm lượng rất to, giọng hơi trầm và lối diễn đạt sinh động. |
| Accent | Cô gái trẻ Bắc Kạn nói rất chậm rãi, giọng cao, âm lượng rất nhỏ nhẹ và đều đều. |
| Emotion | Giọng nam trưởng thành miền Bắc, nói rất to và thong thả, giọng rất cao, lối diễn đạt rất biểu cảm với cảm xúc vui vẻ. |

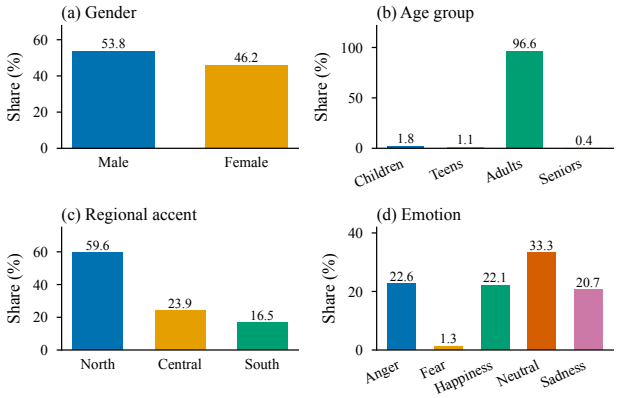


Fig. 2: Distribution of (a) gender, (b) age group, (c) regional accent, and (d) emotion across the dataset. In panel (c), ViMD's province-level labels are mapped to North/Central/South so that all six corpora are covered.

## IV. Dataset Analysis

### A. Dataset Statistics

We retain predefined train/validation/test partitions whenever they are available in the source corpus, which applies to ViMD and LSVSC. Dolly does not provide an official split but includes `voice_id`, so we construct a speaker-disjoint split at the voice level. The remaining corpora without predefined splits or stable speaker identifiers for disjoint partitioning (Vivoice, VNEMOS, and ViSEC) are divided with stratified

TABLE III: Dataset split distribution by source corpus.

| Source | Train | Val | Test | Total |
|---|---|---|---|---|
| Vivoice | 523,169 | 29,065 | 29,065 | 581,299 |
| Dolly | 526,067 | 29,210 | 29,232 | 584,509 |
| ViMD | 10,087 | 1,293 | 1,410 | 12,790 |
| LSVSC | 37,181 | 3,800 | 3,008 | 43,989 |
| VNEMOS | 189 | 11 | 11 | 211 |
| ViSEC | 1,932 | 107 | 108 | 2,147 |
| **Total** | **1,098,625** | **63,486** | **62,834** | **1,224,945** |

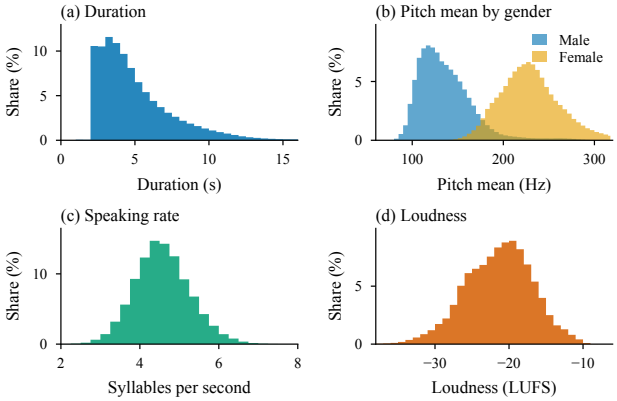


Fig. 3: Distribution of continuous acoustic features: (a) duration, (b) pitch mean stratified by gender, (c) speaking rate, and (d) loudness.

90/5/5 sample-level splits. The released dataset therefore contains 1,098,625 training samples, 63,486 validation samples, and 62,834 test samples. Table III summarizes the split distribution.

### B. Attribute Distribution

Fig. 2 shows the distribution of key categorical attributes across the dataset. Gender is approximately balanced (53.8% male, 46.2% female). The age distribution comprises four groups: adults (96.6%), children (1.8%), teens (1.1%), and seniors (0.4%), the children category coming from ground-truth LSVSC annotations. Regional accent is dominated by Northern speech (59.6%), followed by Central (23.9%) and Southern (16.5%). The emotion subset covers five categories (anger, fear, happiness, neutral, sadness); ViSEC contributes four categories while VNEMOS provides all five, with fear sourced exclusively from VNEMOS, whose fifth class is published as *anxiety* and mapped to fear here.

Fig. 3 presents the continuous acoustic feature distributions. Duration follows a right-skewed distribution (median 4.31 s), consistent with the 2 s minimum threshold. The pitch distribution in (b) reveals the expected bimodal pattern when stratified by gender, with male speakers centered around 137 Hz and female speakers around 231 Hz. Speaking rate clusters around 4.57 syllables per second, and loudness follows a roughly normal distribution centered at $-21.4$ LUFS.

Each utterance is paired with exactly one `caption_full` instruction, averaging 22.1 words across the dataset.

## V. Baseline Systems and Evaluation Protocol

### A. Experimental Setup

To validate the utility of ViInstruct for controllable Vietnamese TTS, two baseline families were adapted to Vietnamese: Parler-TTS [4] and CapSpeech NAR [5]. Both systems follow a two-stage adaptation protocol. *Stage 1* is trained only on the general-task training split of ViInstruct (1,049,236 samples from Vivoice and Dolly), so the models first learn Vietnamese transcript-to-speech generation before specialized control data are introduced.

*Stage 2* then specializes the stage-1 checkpoints toward accent, age, and emotion control. The CapSpeech recipe uses replay-based continual finetuning in the spirit of experience replay [26]: each update draws 50% of the batch from a 100k-utterance general replay buffer and the remainder from targeted control pools—ViMD (20%), region-balanced LSVSC (10%), an age-focused pool (15%), and ground-truth emotion data (5%). The reported Parler-TTS baseline is its released stage-2 checkpoint, which likewise adds specialized control data but uses its own task-rebalancing recipe. The comparison between the two models therefore reflects system-level performance rather than controlled architectural differences.

**Parler-TTS** is an autoregressive system that predicts discrete audio codes conditioned on the transcript and the instruction description. The pretrained English model (`parler-tts-mini-v1`) was adapted by replacing the original Flan-T5-large caption encoder with a frozen ViT5-large [27], keeping T5 subword text units and the DAC-based audio tokenizer and decoder at 44 kHz [28].

**CapSpeech NAR** is a non-autoregressive CrossDiT-based system trained with flow matching. The Vietnamese adaptation replaces the original English grapheme-to-phoneme front-end with 176 character-level text units and the caption encoder with ViT5-large. Mel-spectrograms are synthesized at 24 kHz and decoded with BigVGAN-v2 [29].

### B. Evaluation Setup

Two evaluation subsets are derived from held-out data using a fixed random seed:

- **General Full-Tag Set** (1,228 samples): 1,000 samples from Dolly test, stratified by gender and three-region accent, plus 228 age-balanced samples (76 each of teens/adults/seniors, the maximum non-overlapping seniors available) drawn from the remaining Dolly+Vivoice test pool without ID overlap. Since Dolly is synthesized, this subset is majority-synthetic.
- **Accent Set** (1,710 samples): full ViMD test (1,410) plus downsampled LSVSC test (300; 100 per region) for balanced three-region accent evaluation.

### C. Evaluation Metrics

**Layer 1: Speech Quality.** Intelligibility is measured by character error rate (CER) and word error rate (WER), computed by transcribing synthesized audio with PhoWhisper-large [21] and comparing against reference transcripts after

lowercase normalization and punctuation removal. Naturalness is assessed using predicted mean opinion score (UTMOSv2) [30]. Failure rate is defined as the union of four practical failure modes: unintelligible output (CER $>$ 80%), duration anomaly (predicted-to-reference duration ratio $> 3\times$ or $< 0.3\times$), silent output (RMS $< -60$ dB), and missing generated audio. In the reported benchmark runs, missing-audio rate is 0.0% for both systems; Parler-TTS's 0.57% failure rate therefore stems from the other three modes, not from missing audio.

**Layer 2: Instruction Adherence.** For categorical tags (gender, age, accent), we report *macro-F1* as the primary classification metric. Gender and age are evaluated on the general full-tag set; accent is reported on both the general full-tag set and the dedicated accent set. Accent is predicted by our reproduced ViMD-based wav2vec2 classifier and age by the MERaLiON-based classifier [9], [25], whose three-class inventory excludes the *children* class from evaluation.

**Acoustic Slots.** Speaking rate, pitch, loudness, and pitch variation are re-predicted from synthesized audio. For acoustic tags, we report average bin accuracy across the four acoustic slots.

### D. Results

Table IV reports the main automatic results with reference-audio calibration for Layer-1 quality metrics. The reference audio provides a practical calibration point for the evaluation pipeline (0.96% CER, 2.15% WER, 2.95 UTMOSv2, 0.0% failure), rather than an oracle upper or lower bound. CapSpeech NAR achieves lower CER and WER than this reference while remaining below it on UTMOSv2: because the reference mixes synthesized Dolly with recorded Vivoice audio, this indicates outputs that are easier for the ASR model to transcribe than that mixture, not more natural ones, as the UTMOSv2 gap (2.65 vs. 2.95) shows. Parler-TTS records 3.19% CER and 4.80% WER and is weaker on all Layer-1 metrics.

For categorical adherence, macro-F1 gives the same ranking while being more robust to class imbalance: CapSpeech leads on every tag, and age remains the hardest for both systems, with both strongest on adult speech and weaker on the teen and senior classes. Accent adherence is region-dependent, and CapSpeech improves all three regions over Parler-TTS, especially for Southern speech. CapSpeech also shows stronger acoustic bin control, reaching 66.55% average bin accuracy against 36.46%.

For reference, chance-level macro-F1 on the three-way age and accent tasks is approximately 33%, and chance-level bin accuracy is 14.3% for the seven-level slots and 20% for the five-level slots. Parler-TTS therefore stays close to chance on accent adherence, and both systems remain near chance on age, indicating that coarse and imbalanced age supervision is an open problem rather than a solved control dimension.

TABLE IV: Main automatic evaluation results. ↓/↑: lower/higher is better; categorical adherence uses macro-F1. Reference audio is the held-out source recording per item; it calibrates the measurement pipeline rather than providing a natural-speech upper bound, since the subset is majority-synthetic. Layer-2 entries are N/A because reference instructions come from the annotation pipeline, not independent prompts.

| Metric | Ref. Audio | Parler-TTS | CapSpeech NAR |
|---|---|---|---|
| *Layer 1: Speech Quality* | | | |
| CER (%) ↓ | 0.96 | 3.19 | 0.85 |
| WER (%) ↓ | 2.15 | 4.80 | 1.55 |
| UTMOSv2 ↑ | 2.95 | 2.54 | 2.65 |
| Fail. Rate (%) ↓ | 0.0 | 0.57 | 0.0 |
| *Layer 2: Instruction Adherence* | | | |
| Gender Macro-F1 (%) ↑ | N/A | 84.83 | 100.00 |
| Age Macro-F1 (%) ↑ | N/A | 49.94 | 52.82 |
| Accent F1@General (%) ↑ | N/A | 48.56 | 77.48 |
| Accent F1@Accent Set (%) ↑ | N/A | 38.82 | 72.37 |
| Bin Acc Avg (%) ↑ | N/A | 36.46 | 66.55 |

## VI. Conclusion

This paper presented ViInstruct, the first large-scale Vietnamese dataset for instruction-based TTS, comprising 1,224,945 utterance-instruction pairs from six corpora, together with an automated pipeline that converts heterogeneous Vietnamese speech resources into a unified instruction-TTS benchmark. Benchmark results with two adapted baselines show that the task is already tractable for Vietnamese, with CapSpeech providing a strong reference point for intelligibility and categorical adherence, while also confirming that controllability is not yet solved. The two baselines differ in architecture, training recipe, and pretraining data simultaneously, so their gap should be read as a system-level reference point rather than a controlled architectural comparison. More broadly, ViInstruct positions instruction-style TTS research beyond English and Mandarin and provides an early benchmark for a low-resource tonal language.

**Limitations.** Several limitations remain. First, age annotations are highly imbalanced toward adult speech and partially rely on a coarse three-way automatic classifier, which likely contributes to the near-chance age adherence observed for both baselines. Second, the emotion subset is still small (2,358 samples, 1.5 hours) and is not accompanied by reliable automatic evaluation. Third, the accent subset is regionally imbalanced, with Northern speech dominating the available source data. Fourth, pitch is currently summarized by scalar statistics, which do not explicitly separate lexical tone from paralinguistic prosody. Fifth, nearly half of the pairs come from synthesized speech, so models trained on ViInstruct partly learn the characteristics of the upstream TTS system rather than of natural Vietnamese speech. Finally, the benchmark is entirely automatic: UTMOSv2 is trained on predominantly English listening-test data and is unvalidated for Vietnamese, so its values indicate relative ordering rather than calibrated perceptual scores, and adherence is judged by

classifiers rather than listeners. A human listening study of naturalness, accent identity, and perceived adherence is the evaluation's most important missing piece. Addressing these issues will require better age supervision, broader emotional speech coverage, more recorded speech, tone-aware control representations tailored to Vietnamese, and Vietnamese-validated perceptual evaluation.

## Data Availability and License

We release ViInstruct as an annotation layer, not a repackaged audio corpus: the generated instructions, attribute annotations, and split definitions are keyed to source-corpus identifiers, and users obtain the audio from those corpora directly. The upstream terms are mutually incompatible or unstated — viVoice and Dolly-Audio are gated CC BY-NC-SA 4.0, ViMD is CC BY-NC-ND 4.0, ViSEC is CC BY 4.0, and LSVSC and VNEMOS publish no license — so no combined artifact can satisfy ShareAlike and NoDerivatives at once, and two subsets cannot be redistributed at all. The annotation layer is for non-commercial research use; each subset remains governed by its upstream terms.

## Ethics Statement

No new speech was recorded and no attempt was made to identify speakers. The source material comprises public video channels (Vivoice, ViSEC), film and television recordings (VNEMOS), read speech (ViMD, LSVSC), and synthesized speech (Dolly). Where ground-truth labels are absent, gender, age, and accent are model predictions rather than self-reported identity and should not be read as factual claims about individuals; the instructions inherit that uncertainty.

## References


[1] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "PromptTTS: Controllable text-to-speech with text descriptions," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[2] D. Yang, S. Liu, R. Huang, G. Lei, C. Weng, H. Meng, and D. Yu, "InstructTTS: Modelling expressive TTS in discrete latent space with natural language style prompt," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 2913–2925, 2024.

[3] Y. Leng, Z. Guo, K. Shen, X. Tan, Z. Ju, Y. Liu, Y. Liu, D. Yang, L. Zhang, K. Song, L. He, X.-Y. Li, S. Zhao, T. Qin, and J. Bian, "PromptTTS 2: Describing and generating voices with text prompt," in *Proc. ICLR*, 2024.

[4] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," 2024.

[5] H. Wang, J. Hai, D. Chong, K. Thakkar, T. Feng, D. Yang, J. Lee, L. Moro-Velázquez, J. Villalba, Z. Qin, S. Narayanan, M. Elhilali, and N. Dehak, "CapSpeech: Enabling downstream applications in style-captioned text-to-speech," 2025.

[6] A. Diwan, Z. Zheng, D. Harwath, and E. Choi, "Scaling rich style-prompted text-to-speech datasets," in *Proc. EMNLP*, 2025, pp. 3133–3152.

[7] Y. Ren, J. Yi, J. Tao, H. Sun, Z. Wen, H. Gu, L. Xu, and Y. Bai, "OV-InstructTTS: Towards open-vocabulary instruct text-to-speech," 2026.

[8] K. Huang, Q. Tu, L. Fan, C. Yang, D. Zhang, S. Li, Z. Fei, Q. Cheng, and X. Qiu, "InstructTTSEval: Benchmarking complex natural-language instruction following in text-to-speech systems," 2025.

[9] N. V. Dinh, T. C. Dang, L. T. Nguyen, and K. V. Nguyen, "Multi-dialect vietnamese: Task, dataset, baseline models and challenges," in *Proc. EMNLP*, 2024, pp. 7476–7498.

[10] H.-T. Luong and H.-Q. Vu, "A non-expert Kaldi recipe for vietnamese speech recognition system," in *Proc. WLSP*, 2016, pp. 51–55.

[11] T. Vu, L. T. Nguyen, and D. Q. Nguyen, "Zero-shot text-to-speech for vietnamese," in *Proc. ACL (Volume 2: Short Papers)*, 2025, pp. 1042–1049.

[12] L. T. T. Tran, H.-G. Kim, H. M. La, and S. V. Pham, "Automatic speech recognition of vietnamese for a new large-scale corpus," *Electronics*, vol. 13, no. 5, p. 977, 2024.

[13] N. D. Q. Anh, M.-H. Ha, Q. C. Nguyen, T. H. N. Thi, Q. Vu, D. X. Minh-Duc, D.-C. Nguyen, and T. K. Dinh, "VNEMOS: Vietnamese speech emotion inference using deep neural networks," in *Proc. Int. Conf. Integrated Circuits, Design, Verification (ICDV)*, 2024, pp. 97–101.

[14] P. V. Thanh, N. T. T. Huyen, P. N. Quan, and N. T. T. Trang, "A robust pitch-fusion model for speech emotion recognition in tonal languages," in *Proc. IEEE ICASSP*, 2024, pp. 12 386–12 390.

[15] G. Liu, Y. Zhang, Y. Lei, Y. Chen, R. Wang, Z. Li, and L. Xie, "PromptStyle: Controllable style transfer for text-to-speech with natural language descriptions," in *Proc. INTERSPEECH*, 2023, pp. 4888–4892.

[16] M. Kawamura, R. Yamamoto, Y. Shirahata, T. Hasumi, and K. Tachibana, "LibriTTS-P: A corpus with speaking style and speaker identity prompts for text-to-speech and style captioning," in *Proc. INTERSPEECH*, 2024, pp. 1850–1854.

[17] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 3339–3354, 2024.

[18] Qwen Team, "Qwen3.5: Accelerating productivity with native multimodal agents," Feb. 2026, [Online]. Available: https://qwen.ai/blog?id=qwen3.5.

[19] T. L. P. Gia, T. P. Minh, H. N. Quoc, T. N. Quoc, and V. T. Hoang, "viVoice: Enabling vietnamese multi-speaker speech synthesis," 2024, cC BY-NC-SA 4.0. [Online]. Available: https://hf.co/datasets/capleaf/viVoice.

[20] V. H. Nguyen and D. T. Nguyen, "Dolly-Audio: Vietnamese multi-speaker high-quality speech corpus," 2025, dolly AI Team. CC BY-NC-SA 4.0. [Online]. Available: https://hf.co/datasets/dolly-vn/dolly-audio-1000h-vietnamese.

[21] T.-T. Le, L. T. Nguyen, and D. Q. Nguyen, "PhoWhisper: Automatic speech recognition for vietnamese," in *Proc. ICLR Tiny Papers*, 2024.

[22] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A convolutional representation for pitch estimation," in *Proc. IEEE ICASSP*, 2018, pp. 161–165.

[23] Silero Team, "Silero VAD: Pre-trained enterprise-grade voice activity detector," 2024, [Online]. Available: https://github.com/snakers4/silero-vad.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[25] MERaLiON Team, "MERaLiON-AudioLLM: Bridging audio and language with large language models," 2024.

[26] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," in *Proc. NeurIPS*, 2019.

[27] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh, "ViT5: Pretrained text-to-text transformer for vietnamese language generation," in *Proc. NAACL SRW*, 2022, pp. 136–142.

[28] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Proc. NeurIPS*, 2023, pp. 27 980–27 993.

[29] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," in *Proc. ICLR*, 2023.

[30] K. Baba, W. Nakata, Y. Saito, and H. Saruwatari, "The T05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech," in *Proc. IEEE SLT*, 2024, pp. 818–824.