Vi-SparkRL: Enhancing Tonal Accuracy and Naturalness in Vietnamese TTS via Multi-Objective Rewards

Xuan-Binh Dinh-Thi*, Tri-Nhan Do†, Tu-Anh Nguyen†, Van-An Chu†, and Dang-Khoa Mac†
* FPT University, Vietnam
binhdtxse170075@fpt.edu.vn
† VinFast, Vietnam
{v.nhandt23, v.anhnct2, v.ancv, v.khoamd}@vinfast.vn

***Abstract*—Recent advances in Large Language Models and neural audio codecs have significantly improved zero-shot Text-to-Speech (TTS) systems; However, maintaining tonal accuracy in languages such as Vietnamese remains a formidable challenge. Existing models often exhibit unstable prosody or phonetic biases that result in a noticeable foreign accent. In this paper, we propose Vi-SparkRL, a Reinforcement Learning enhanced framework designed to preserve Vietnamese tonal integrity in lightweight zero-shot Text-to-Speech models. Our approach introduces Group Sequence Policy Optimization, which performs sequence-level optimization to mitigate training instability and mode-collapse issues common in token-level reinforcement learning for continuous audio signals. Furthermore, we design a comprehensive multi-objective reward mechanism that integrates a specialized VietTone Reward for positional tonal matching, alongside constraints for intelligibility, speaker similarity, and acoustic quality. Evaluated on the PhoAudiobook dataset, Vi-SparkRL achieves a Word Error Rate of 2.85% and a tonal accuracy of 0.96, significantly outperforming traditional Supervised Fine-Tuning and competitive baselines. Our results demonstrate that sequence-level RL effectively aligns TTS models with the complex prosodic requirements of tonal languages while maintaining high naturalness.**

I. Introduction

Recent advances in Large Language Models (LLMs) and neural audio codecs have revolutionized Text-to-Speech (TTS) systems, enabling high-quality zero-shot voice cloning from a short reference audio.Recent LLM-based TTS systems formulate speech synthesis as language modeling over discrete audio tokens [1], [2], [3], [4], [5], achieving high-quality zero-shot cloning for resource-rich languages. Spark-TTS-0.5B [6] simplified this paradigm via BiCodec, decoupling semantic and global tokens in a single-stream framework. Despite this rapid progress, tonal languages such as Vietnamese remain a persistent challenge for zero-shot TTS. Vietnamese possesses six lexical tones where F_0 trajectory dictates meaning [7]; confusing tones such as hỏi and ngã changes semantics entirely. Multilingual systems like XTTS-v2 [8], [9] and GPT-SoVITS [10] produce unstable prosody and phonetic bias on Vietnamese [11], because token-level cross-entropy loss does not enforce tonal constraints.

RL-based post-training has been applied to TTS through preference optimization [12], [13], Group Sequence Policy Optimization (GSPO) for emotional and cross-lingual control [14], [15], and diffusion-model alignment [16], [17]. Although these studies demonstrate the viability of RL for improving TTS quality, none specifically addresses tonal accuracy in tonal languages or incorporates a language-specific prosodic reward into the RL loop.

A further limitation of existing RL approaches for TTS lies in the choice of policy optimization algorithm. Group Relative Policy Optimization (GRPO), adopted by EMORL-TTS and Align2Speak, operates at the token level, computing importance ratios per token independently. As shown in the GSPO paper [18], this effectively yields a single sample per token position, leading to high-variance gradient estimates—particularly severe for speech, where acoustic token sequences far exceed their text counterparts (e.g., $\sim$250 tokens for a five-second utterance at 50 Hz vs. fewer than 20 text tokens). Since prosody is an inherently *sequence-level* phenomenon, Group Sequence Policy Optimization (GSPO) instead computes importance ratios at the sequence level with length normalization, providing stable gra-

dient signals aligned with the holistic nature of speech production [18].

In this work, we propose **Vi-SparkRL**[1], an RL-enhanced framework preserving Vietnamese tonal integrity in lightweight (0.5B) zero-shot TTS. Built on Spark-TTS-0.5B [6], it introduces two innovations: (1) the adaptation of GSPO to acoustic modeling—to our knowledge, the first sequence-level policy optimization for speech synthesis; and (2) a dedicated **VietTone Reward** measuring positional tonal matching. Both are integrated into an online multi-objective loop scoring generated waveforms in real time for intelligibility (WER via Whisper Large V3 [19]), speaker similarity (SIM via ECAPA-TDNN WavLM [20], [21]), and acoustic quality (UTMOS [22]).

We evaluate Vi-SparkRL on the *PhoAudiobook* dataset [23], a 941-hour professionally narrated Vietnamese corpus with high audio quality (SI-SNR: 4.91 dB) and long segment duration (mean 11.66 s). Unlike earlier resources such as *VIVOS* [11] (15 hours) or *viVoice* [24] (inconsistent quality), it provides the reliable prosodic baseline essential for RL-based tonal alignment, where reward models require high-fidelity reference signals.

The main contributions of this work are as follows:

- We propose Vi-SparkRL, the first RL-based framework designed to optimize Vietnamese tonal accuracy in lightweight (0.5B) zero-shot TTS, addressing a critical gap for tonal languages.
- We adapt GSPO to the acoustic modeling phase—to our knowledge, the first sequence-level RL applied to speech synthesis—stabilizing training and mitigating mode-collapse issues inherent in token-level GRPO for long-form audio.
- We design a multi-objective reward integrating a novel VietTone Reward for positional tonal matching with WER, SIM, and UTMOS constraints, outperforming SFT and competitive baselines on PhoAudiobook.

II. Methodology

The Vi-SparkRL framework integrates an end-to-end online audio evaluation loop into the RL training phase. At each step, the LLM generates G token sequences, decoded into 16kHz waveforms via the frozen BiCodec [5] and scored in real time by the multi-objective reward models to compute group-relative advantages for the policy update.

A. Group Sequence Policy Optimization (GSPO)

To enhance the prosody and naturalness of the baseline SparkTTS model, we employ Group Sequence Policy Optimization (GSPO). Unlike PPO, GSPO eliminates the separate value function via group-relative advantages. For each prompt x, a group of G sequences $\{y_1, y_2, \ldots, y_G\}$ is generated from π_θ, and the advantage of each sequence i is estimated by normalizing its reward R_i within the group:

$$\hat{A}_i = \frac{R_i - \bar{R}}{\sigma_R + \epsilon} \tag{1}$$

where $\bar{R}$ and σ_R denote the mean and standard deviation of the group rewards $\{R_1, \ldots, R_G\}$.

To ensure stable alignment for continuous audio signals, we apply sequence-level importance sampling with length normalization. Unlike token-level methods that compute a separate ratio per token, GSPO optimizes the joint probability of the entire acoustic token sequence, mitigating the high-variance issue in long-form synthesis:

$$\begin{aligned} L_{\text{GSPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \min\big(& s_i(\theta)\,\hat{A}_i, \\ & \text{clip}(s_i(\theta),\,1-\epsilon,\,1+\epsilon)\,\hat{A}_i\big) \\ & -\,\beta\, D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big) \end{aligned} \tag{2}$$

where the sequence-level importance ratio is length-normalized over the $|y_i|$ acoustic tokens:

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)}\right)^{1/|y_i|} \tag{3}$$

Here $\pi_\theta(y_i \mid x)$ denotes the joint probability of the complete acoustic token sequence, $\pi_{\theta_{\text{old}}}$ is the behavior policy used for sampling, and π_{ref} is the frozen reference policy constraining divergence via the KL term ($\beta = 0.01$). The exponent $1/|y_i|$ normalizes for sequence length, which is the key mechanism distinguishing sequence-level GSPO from token-level GRPO. The procedure is summarized in Algorithm 1.

[1]Audio samples and a web-based demonstration are available at: https://anonymous-tts-demo.github.io/gspo_sparktts/

Algorithm 1 GSPO for TTS

Require: Dataset $\mathcal{D}$, policy π_θ, reference π_{ref}, learning rate η

for each prompt $x \in \mathcal{D}$ **do**
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  Sample G sequences $\{y_1, \ldots, y_G\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$
  for $i = 1$ to G **do**
    $\text{wave}_i \leftarrow \text{BiCodec\_Decode}(y_i)$
    $R_i \leftarrow R_{\text{total}}(\text{wave}_i, x)$
  end for
  $\hat{A}_i \leftarrow (R_i - \bar{R})/(\sigma_R + \epsilon)$ **for all** i
  $s_i(\theta) \leftarrow \left(\pi_\theta(y_i \mid x)/\pi_{\theta_{\text{old}}}(y_i \mid x)\right)^{1/|y_i|}$
  $\theta \leftarrow \theta + \eta \cdot \nabla_\theta L_{\text{GSPO}}(\theta)$
end for

B. Multi-objective Hybrid Reward System

A core contribution of our work is the design of a hybrid reward function R_{total} that simultaneously optimizes for acoustic quality, speaker identity, and linguistic accuracy. The total reward is defined as:

$$R_{\text{total}} = 2.0 \times \left(\sum_{k \in \{\text{mos, sim, wer, tone}\}} w_k r_k \right) - 1.0 \quad (4)$$

where $w_{\text{mos}} = 0.4$, $w_{\text{sim}} = 0.3$, $w_{\text{wer}} = 0.2$, and $w_{\text{tone}} = 0.1$, scaled to $[-1, 1]$ for stable gradient updates. The higher weight on acoustic naturalness reflects its role as the primary quality anchor, while the VietTone reward acts as a complementary fine-grained tonal constraint whose impact is validated in the ablation study in the next section.

1) Standard Reward Components: Following established practices in RL-based TTS alignment [12], [13], [15], we adopt three widely-used reward signals. **Acoustic Realism** is measured via UTMOS [22], with the raw MOS score $s_{\text{mos}} \in [1, 5]$ normalized as $r_{\text{mos}} = (\text{clip}(s_{\text{mos}}, 1, 5) - 1)/4$. **Speaker Identity Preservation** is computed using cosine similarity between ECAPA-TDNN WavLM embeddings [20], [21] of the reference and generated audio: $r_{\text{sim}} = (\cos(e_{\text{ref}}, e_{\text{gen}}) + 1)/2$. **Linguistic Faithfulness** is enforced via Whisper Large V3 [19] transcription, with $r_{\text{wer}} = 1.0 - \min(\text{WER}(\text{text}_{\text{ref}}, \text{text}_{\text{hyp}}), 1.0)$.

2) Language-specific Prosody: VietTone Reward: To strictly enforce tonal integrity, we propose a dedicated **VietTone Reward** (r_{tone}) providing a granular linguistic signal during the RL phase, targeting the tonal confusions that standard zero-shot models frequently produce.

Let $\mathcal{T}(s)$ be a transformation function that extracts a sequence of lexical tone marks from a string s using Unicode normalization (mapping tones to a discrete set $\{0, \ldots, 5\}$ representing: *ngang, huyền, sắc, hỏi, ngã, nặng*). For each generated audio sample, we first obtain a hypothesis transcript text_{hyp} using Whisper Large V3. The reward is then calculated as the positional match rate against the reference text text_{ref}:

$$r_{\text{tone}} = \frac{1}{L} \sum_{j=1}^{L} \mathbb{I}\left[\mathcal{T}(\text{text}_{\text{ref}})_j = \mathcal{T}(\text{text}_{\text{hyp}})_j\right] \quad (5)$$

where L denotes the total number of syllables in the reference text and $\mathbb{I}[\cdot]$ is the indicator function. We note that r_{tone} operates on ASR-recognized tone marks and therefore reflects the perceptual recoverability of tones by the ASR model rather than a direct measurement of F_0 contours; a direct acoustic F_0 analysis is discussed as future work in Section III-C. The mechanism of this reward is illustrated in Table I. Although the hypothesis is phonetically similar to the reference, deviations in the third and fifth tones (e.g., "đôi" vs. "đối") yield a penalized score of 0.60. This constraint discourages "tone-deaf" generations, a common SFT failure mode.

Table I
EXAMPLE OF VIETTONE REWARD CALCULATION BASED ON POSITIONAL TONAL MATCHING.

Type	Vietnamese Sentence	Tones $T(s)$	Match	Score
Ref.	"Học đi đôi với hành"	[3, 0, 0, 2, 1]	-	-
Hyp.	"Học đi **đối** với **hãnh**"	[3, 0, **2**, 2, **5**]	✓,✓,×,✓,×	-
Final	**Positional Match Rate**		3 / 5	**0.60**

III. EXPERIMENTS

A. Experimental Setup

1) Dataset Selection and Partitioning: From the 941-hour PhoAudiobook corpus [23], we select 14,000 utterances with WER < 0.1 measured by Wav2Vec 2.0 [25], split 80/10/10 into 11,200 for GSPO training, 1,400 for validation, and 1,400 for zero-shot testing with no speaker overlap across partitions.

Audio is tokenized via BiCodec [5] into global style embeddings and temporal semantic tokens, formatted as prompt-only causal LM sequences for TRL

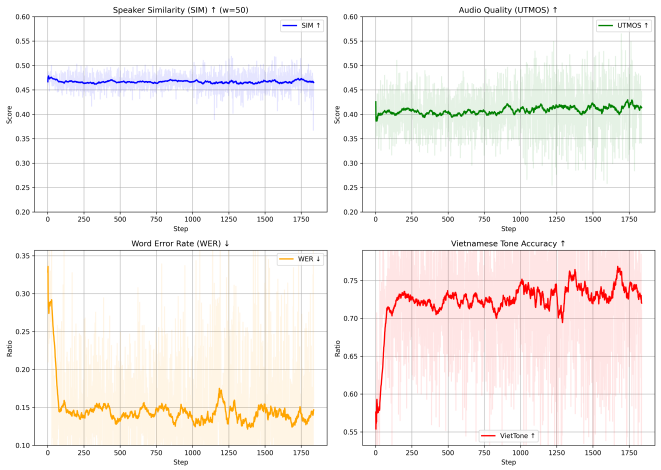

Figure 1. Training dynamics of Vi-SparkRL. Top: SIM (↑) and UTMOS (↑); Bottom: WER (↓) and Tone Accuracy (↑). Smoothed trends (window $w = 50$) are superimposed on raw step-wise values.

[26] compatibility. The baseline **Vi-SparkTTS-0.5B** [6] (0.5B params) shares the identical architecture with Vi-SparkRL, so gains are attributable solely to GSPO and the reward design.

2) Training of the Baseline Model: We fine-tune Vi-SparkTTS-0.5B with Low-Rank Adaptation (LoRA) ($r = 16$, $\alpha = 32$, dropout 0.05) on all attention and feed-forward network (FFN) projections using cross-entropy loss over BiCodec tokens.

3) Reinforcement Learning Setup: The converged SFT LoRA adapters are merged into the base weights as π_{ref}; a fresh LoRA set serves as the active policy π_θ optimized with GSPO.

During training, the active policy generates $K = 8$ candidate acoustic trajectories per prompt using a sampling temperature of 0.9, scored by the multi-objective reward defined in Eq. (4). The Kullback–Leibler (KL)-divergence regularization coefficient is set to $\beta = 0.01$, the peak learning rate to 1×10^{-6}, and the sequence clipping epsilon to $\epsilon = 3 \times 10^{-4}$. All experiments were executed on four NVIDIA A100 (80GB) GPUs in a Distributed Data Parallel (DDP) configuration via the Accelerate library. To ensure a fair comparison, all baseline models in Table II were re-evaluated on our PhoAudiobook test set using identical metric pipelines (Whisper Large V3 for WER, ECAPA-TDNN WavLM for SIM, and UTMOS for acoustic quality). Tone Accuracy is reported only for our models, as no existing baseline provides this metric.

Table II
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART TTS MODELS.

Model	Architecture	WER↓	UTMOS↑	SIM↑	RTF↓
VITS [27]	Variational Autoencoder	0.0521	3.55	0.62	**0.05**
XTTS-v2 [9]	Autoregressive GPT	0.0785	3.82	0.69	0.65
MeloTTS [28]	VITS-based	0.0488	3.75	0.65	0.08
OpenVoice [29]	Flow Matching	0.0550	3.90	0.72	0.25
F5-TTS [30]	Flow Matching	0.0395	4.05	0.79	0.45
Fish-Speech [31], [32]	VQ-VAE + LLM	0.0412	4.02	**0.80**	0.52
Valtec-TTS [33]	Flow-matching/VITS	0.0425	3.92	0.76	0.12
Ours	**Group Sequence RL**	**0.0285**	**4.10**	0.75	0.38

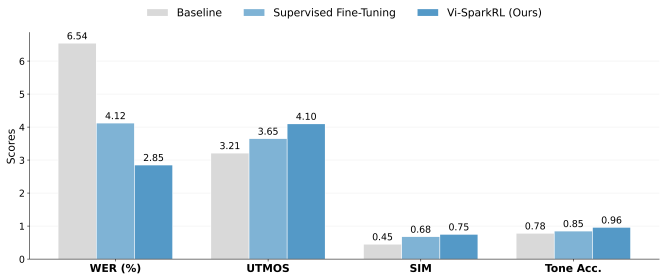

Figure 2. Objective evaluation on PhoAudiobook test set across three training paradigms: zero-shot baseline, SFT, and Vi-SparkRL.

B. Result Analysis and Discussion

Fig. 1 confirms that GSPO training is stable over 1,800 steps with no reward hacking: WER converges rapidly within the first 250 steps, Tonal Accuracy improves gradually throughout, UTMOS trends upward, and SIM remains steady, indicating that KL regularization ($\beta = 0.01$) preserves speaker identity. The final results in Table II show that Vi-SparkRL consistently outperforms all baselines, with each training stage yielding consistent gains across all four metrics.

1) Acoustic Quality and Tonal Integrity: As shown in Fig. 2, Vi-SparkRL achieves the highest UTMOS and lowest WER among all three paradigms, with the most pronounced gain occurring in Tonal Accuracy. Isolating the VietTone reward (GSPO without tone: 0.88 → Vi-SparkRL: 0.96) confirms it directly addresses the tonal confusion that GSPO alone cannot resolve: while GSPO improves general intelligibility and naturalness, it lacks an explicit mechanism to penalize pitch-contour deviations. As both Tonal Accuracy and WER derive from the same Whisper transcription, part of the observed correlation between them is expected; nonetheless, the tonal gain remains consistent across the test set.

2) Speaker Consistency and Training Dynamics: Fig. 2 also reveals that SIM improves substantially from

Table III
ABLATION STUDY ON RL ALGORITHMS AND REWARD COMPONENTS.

Method	WER ↓	UTMOS ↑	SIM ↑	Tone ↑
Pretrained (zero-shot)	0.0654	3.21	0.49	0.78
SFT baseline	0.0412	3.65	0.68	0.85
+ GRPO	0.0350	3.85	0.70	0.90
+ GSPO (no tone)	0.0345	4.08	**0.78**	0.88
Vi-SparkRL	**0.0285**	**4.10**	0.75	**0.96**

the zero-shot baseline to SFT but decreases slightly under Vi-SparkRL, reflecting an inherent trade-off: aggressive pitch alignment via the VietTone reward can shift F_0 contours away from the reference speaker's characteristics. The gap relative to top baselines (F5-TTS 0.79, Fish-Speech 0.80) nonetheless remains modest and stays competitive with flow-matching systems such as OpenVoice (0.72) and Valtec-TTS (0.76).

3) Ablation Study: Table III isolates the contribution of each component.

Token-level GRPO yields marginal gains over SFT due to high-variance advantage estimation on long acoustic sequences, while sequence-level GSPO substantially improves naturalness with stable convergence. GSPO alone leaves tonal accuracy largely unchanged; adding the VietTone reward produces the largest single-component gain, with a minor SIM trade-off. We adopt WER over Phoneme Error Rate as the intelligibility reward because grapheme-to-phoneme conversion bottlenecks multi-generation rollouts.

C. Limitations

Our tonal metric operates on ASR-recognized tone marks and does not directly measure F_0 realization; a correct recognized tone implies perceptual recoverability by the ASR model but does not fully guarantee native-level acoustic realization. Likewise, our evaluation is objective-only. Direct acoustic F_0-contour analysis and native-listener subjective tests (naturalness and tone-perception MOS) would further validate the perceptual tonal gains and are left for future work.

IV. CONCLUSION

Vi-SparkRL demonstrates that sequence-level GSPO with a language-specific VietTone Reward effectively preserves tonal integrity in Vietnamese zero-shot TTS, outperforming existing baselines.

REFERENCES

[1] C. Wang et al., "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[2] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, "Voicecraft: Zero-shot speech editing and text-to-speech in the wild," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 12 442–12 462.

[3] P. Anastassiou et al., "Seed-tts: A family of high-quality versatile speech generation models," *arXiv preprint arXiv:2406.02430*, 2024.

[4] Z. Du et al., "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[5] X. Wang et al., "Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens," *arXiv preprint arXiv:2503.01710*, 2025.

[6] SparkAudio, *Spark-tts-0.5b*, https://huggingface.co/SparkAudio/Spark-TTS-0.5B, Hugging Face Model Hub, 2025.

[7] A. H. Pham, *Vietnamese tone*. Routledge, 2005.

[8] E. Casanova et al., "XTTS: A massively multilingual zero-shot text-to-speech model," in *Proc. INTERSPEECH 2024*, 2024, pp. 4978–4982. DOI: 10.21437/Interspeech.2024-2016

[9] C. AI, *XTTS-v2: A multilingual zero-shot TTS model*, https://huggingface.co/coqui/XTTS-v2, Accessed: 2025-03-22, 2023.

[10] RVC-Boss, *GPT-SoVITS: Few-shot voice conversion and text-to-speech*, https://github.com/RVC-Boss/GPT-SoVITS, 2024.

[11] H.-T. Luong and H.-Q. Vu, "A non-expert kaldi recipe for vietnamese speech recognition system," in *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, 2016, pp. 51–55.

[12] D. Zhang et al., "Speechalign: Aligning speech generation to human preferences," *Advances in Neural Information Processing Systems*, vol. 37, pp. 50 343–50 360, 2024.

[13] S. S. Hussain et al., "Koel-tts: Enhancing llm based speech generation with preference alignment and classifier free guidance," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 21 230–21 245.

[14] H. Li, Y. Liu, Y. Sun, H. Shi, L. Qu, and T. Li, "Emorl-tts: Reinforcement learning for fine-grained emotion control in llm-based tts," *ArXiv*, vol. abs/2510.05758, 2025. [Online]. Available: https : / / api . semanticscholar . org / CorpusID : 281886724

[15] S. Hussain et al., *Align2speak: Improving tts for low resource languages via asr-guided online preference optimization*, 2025. arXiv: 2509.21718 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2509.21718

[16] J. Chen, J. S. Byun, M. Elsner, P. Wang, and A. Perrault, "Fine-tuning text-to-speech diffusion models using reinforcement learning with human feedback," *arXiv preprint arXiv:2508.03123*, 2025.

[17] C. Gao et al., "Explore the reinforcement learning for the LLM based ASR and TTS system," *arXiv preprint arXiv:2509.18569*, 2025.

[18] C. Zheng et al., "Group sequence policy optimization," *arXiv preprint arXiv:2507.18071*, 2025.

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, PMLR, 2023, pp. 28 492–28 518.

[20] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, 2020, pp. 3830–3834.

[21] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[22] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.

[23] T. Vu, D. Q. Nguyen, et al., "Zero-shot text-to-speech for vietnamese," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2025, pp. 1042–1049.

[24] Capleaf, *viVoice: A vietnamese text-to-speech dataset*, https://huggingface.co/datasets/capleaf/viVoice, Hugging Face Datasets, 2024.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[26] L. von Werra et al., *Trl: Transformer reinforcement learning*, https://github.com/huggingface/trl, 2020.

[27] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International conference on machine learning*, PMLR, 2021, pp. 5530–5540.

[28] W. Zhao, X. Yu, and Z. Qin, *Melotts: High-quality multi-lingual multi-accent text-to-speech*, 2023. [Online]. Available: https : / / github . com / myshell-ai/MeloTTS

[29] Z. Qin, W. Zhao, X. Yu, and X. Sun, "Openvoice: Versatile instant voice cloning," *arXiv preprint arXiv:2312.01479*, 2023.

[30] Y. Chen et al., "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 6255–6271.

[31] S. Liao et al., *Fish audio s2 technical report*, 2026. arXiv: 2603 . 08823 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2603.08823

[32] S. Liao et al., "Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis," *arXiv preprint arXiv:2411.01156*, 2024.

[33] V. Team, *Valtec vietnamese tts with zero-shot voice cloning*, 2026. [Online]. Available: https : //github.com/tronghieuit/valtec-tts