Zero-Shot Text-to-Speech for Vietnamese

Thi Vu, Linh The Nguyen, Dat Quoc Nguyen
Movian AI, Vietnam
{thivuxy, toank45sphn, datquocnguyen}@gmail.com

Abstract

This paper introduces PhoAudiobook, a newly curated dataset comprising 941 hours of high-quality audio for Vietnamese text-to-speech. Using PhoAudiobook, we conduct experiments on three leading zero-shot TTS models: VALL-E, VoiceCraft, and XTTS-V2. Our findings demonstrate that PhoAudiobook consistently enhances model performance across various metrics. Moreover, VALL-E and VoiceCraft exhibit superior performance in synthesizing short sentences, highlighting their robustness in handling diverse linguistic contexts. We publicly release PhoAudiobook to facilitate further research and development in Vietnamese text-to-speech.

1 Introduction

Text-to-speech (TTS) synthesis has witnessed significant advancements in recent years. State-of-the-art TTS systems typically use a cascaded pipeline that consists of an acoustic model and a vocoder, with mel-spectrograms serving as intermediate representations (Ren et al., 2019; Li et al., 2019; Tan et al., 2024). These advanced TTS systems can synthesize high-quality speech for single or multiple speakers (Kim et al., 2021; Liu et al., 2022).

Zero-shot TTS has emerged as a promising approach to overcome the limitations of traditional TTS systems in generalizing to unseen speakers. By leveraging techniques such as speaker adaptation and speaker encoding, zero-shot TTS aims to synthesize speech for new speakers using only a few seconds of reference audio (Arik et al., 2018; Wang et al., 2020; Cooper et al., 2020; Wu et al., 2022; Casanova et al., 2022, 2024). Recent works have explored the application of language modeling approaches to zero-shot TTS, achieving impressive results. For example, VALL-E (Wang et al., 2023) introduces a text-conditioned language model trained on discrete audio codec tokens, enabling TTS to be treated as a conditional codec language modeling task. VoiceCraft (Peng et al., 2024) casts both sequence infilling-based speech editing and continuation-based zero-shot TTS as a left-to-right language modeling problem by rearranging audio codec tokens.

Despite advancements in zero-shot TTS, its application to low-resource languages remains challenging. These languages often lack the large-scale, high-quality datasets needed to train robust TTS models (Gutkin et al., 2016; Chen et al., 2019; Lux et al., 2022; Ngoc et al., 2023; Huang et al., 2024). Also, linguistic and phonetic differences between languages introduce additional challenges in adapting existing models to new languages. As a result, the performance of zero-shot TTS systems in low-resource languages is often limited, hindering their practical usability.

In this paper, we focus on advancing zero-shot TTS for Vietnamese. Our contributions are:

- **We present PhoAudiobook**, a 941-hour high-quality long-form speech dataset that overcomes the limitations of existing Vietnamese datasets, which usually contain audio samples shorter than 10 seconds. The pipeline to create this dataset can be easily adapted to other languages.
- **We conduct a comprehensive experimental study** to evaluate the performance of three state-of-the-art zero-shot TTS models: VALL-E, VoiceCraft, and XTTS-v2 (Casanova et al., 2024). Using a combination of objective and subjective metrics across multiple benchmark datasets, our results demonstrate that XTTS-v2 trained on PhoAudiobook outperforms its counterpart trained on an existing dataset. Additionally, VALL-E and VoiceCraft exhibit robustness in synthesizing varied input lengths.
- **We publicly release PhoAudiobook** at `https://huggingface.co/datasets/thivux/phoaudiobook` for non-commercial purposes.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1042–1049

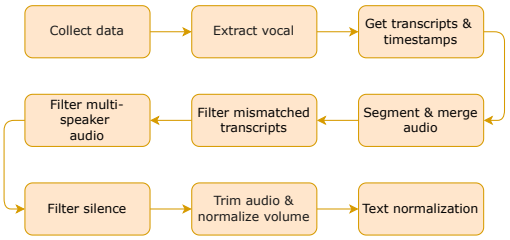

Figure 1: PhoAudiobook creation pipeline.

2 Dataset

2.1 PhoAudiobook

Figure 1 illustrates the creation process of our PhoAudiobook dataset.

First, we collect raw Vietnamese audiobook data from the publicly accessible website `https://sachnoiviet.net`. This raw dataset includes 23K hours of content from 2,697 audiobooks, narrated by 735 distinct speakers. Next, we use `demucs` to extract the vocal track, effectively removing any background music or sound effects (Défossez, 2021; Rouard et al., 2023). We then employ the multilingual `Whisper-large-v3` model to generate transcriptions and corresponding timestamps for the audio data (Radford et al., 2023). The output from `Whisper-large-v3` includes transcripts for short audio segments, usually a few seconds long, along with their corresponding timestamps. These segments are often aligned with natural pauses in speech. We then concatenate *successive* audio segments and their corresponding transcripts to create longer audio samples and transcriptions, each lasting between 10 and 20 seconds. To ensure the quality of the transcriptions, we process these merged samples using the state-of-the-art Vietnamese ASR model, `PhoWhisper-large` (Le et al., 2024). We then retain only the samples where the `Whisper-large-v3`-based transcription matches exactly with the transcription output from `PhoWhisper-large`. Furthermore, we tackle the challenge of multi-speaker audio. We use the `wav2vec2-bartpho` model to identify and filter out short audio samples containing multiple speakers, ensuring that all audio segments associated with a particular speaker are indeed spoken by that individual.[1]

To reduce excessive silence in the audio data, we exclude samples with transcripts shorter than 25 words and trim silence from the beginning and end of each sample. Additionally, we use the `sox` library to normalize audio volume levels for maintaining consistency and avoiding abrupt loudness throughout the dataset.[2] Finally, we standardize the transcriptions through a text normalization step, which includes converting text to lowercase, adding appropriate punctuation, and normalizing numerical expressions into their text form (e.g., "43" becomes "forty three"). We carry out this text normalization step using a sequence-to-sequence model, which we develop by fine-tuning the pre-trained `mbart-large-50` model (Liu et al., 2020) on a Vietnamese dataset consisting of unnormalized input and normalized output text pairs.[3]

The data creation process described above results in a refined 1,400-hour audio corpus. To ensure balanced speaker representation, we limit each speaker to a maximum of 4 hours of audio. This results in a high-quality dataset comprising **941** hours of audio from **735** speakers. From the remaining $1,400 - 941 = 559$ hours of audio, we sample 0.8 hour of audio from 20 speakers to construct a "seen" speaker test set. Additionally, we split the 941 hours of audio from 735 speakers into three sets (on speaker level): a training set containing 940 hours from 710 speakers, a validation set with 0.5 hours from 5 speakers, and an "unseen" speaker test set comprising 0.4 hours from 20 speakers who have the shortest total audio durations. Here, the 20 speakers in the "seen" speaker test set are part of the 710 speakers used for training.

We conduct a post-processing step to manually inspect each audio sample and its corresponding transcription from both the "seen" and "unseen" speaker test sets. This process results in all correct transcriptions in the test sets of PhoAudiobook.

2.2 Dataset analysis

Table 1 presents the characteristics of our dataset – PhoAudiobook, in comparison to other Vietnamese speech datasets, including VinBigData (VinBigData, 2023), VietnamCeleb (Pham et al., 2023), the VLSP 2020 ASR Challenge,[4] BUD500 (Pham et al., 2024), and viVoice (Gia et al., 2024).

Duration: PhoAudiobook, with 941 hours, is the second-largest dataset, closely following viVoice,

[1] `https://huggingface.co/nguyenvulebinh/wav2vec2-bartpho`

[2] `https://sourceforge.net/projects/sox`

[3] `https://huggingface.co/datasets/nguyenvulebinh/spoken_norm_pattern`

[4] `https://vlsp.org.vn/vlsp2020/eval/asr`

Dataset	Duration (h)	Mean Dur. (s)	25% Dur. (s)	75% Dur. (s)	Domain	SI-SNR (dB)	# Speakers	Rate (wpm)	Fs (Hz)
VinBigData	101	6.47	3.54	8.09	General-purpose	4.77	Unknown	229	16000
VietnamCeleb	187	7.74	2.84	9.64	Unknown	3.89	Unknown	No transcripts	16000
VLSP	243	4.37	2.43	5.21	Unknown	4.35	Unknown	242	16000
BUD500	462	2.56	2.11	2.94	General-purpose	4.22	Unknown	224	16000
viVoice	1016	4.12	1.96	5.55	General-purpose	4.81	Unknown	243	24000
PhoAudiobook	941	11.66	10.63	12.18	Audiobooks	4.91	735	201	16000

Table 1: Characteristics of PhoAudiobook and other speech datasets for Vietnamese.

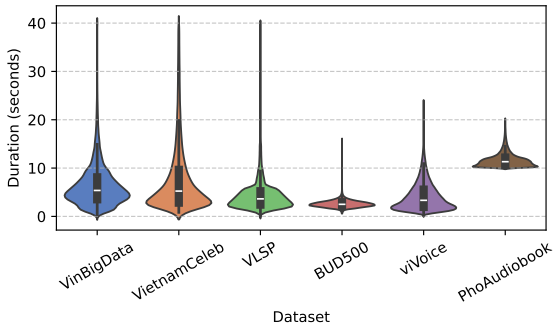

Figure 2: Duration distributions of datasets. Audio samples are capped at 40 seconds for visualization purposes.

which has 1,016 hours. The other datasets are considerably smaller, ranging from 101 to 462 hours. Figure 2 shows that previous datasets primarily consist of audio segments shorter than 10 seconds. PhoAudiobook addresses this limitation by providing audio samples ranging from 10 to 20 seconds.

Domain: PhoAudiobook is derived from audiobooks, typically recorded with professional equipment in controlled environments, ensuring high-quality audio. In contrast, other datasets are general-purpose (e.g., news, YouTube videos, conversations) and may include audio recorded on consumer devices in uncontrolled settings, often with background noise. However, general-purpose datasets have the advantage of covering diverse topics and speaking styles.

Signal-to-Noise Ratio (SI-SNR): Using an SI-SNR estimator from the speechbrain toolkit (Ravanelli et al., 2024), we calculated SI-SNR across 1000 randomly sampled audio from each dataset. PhoAudiobook achieves the highest SI-SNR, surpassing all other datasets, including viVoice.

Speaker Information: PhoAudiobook is the only dataset with explicit speaker identity, and therefore, number of speakers (735).

Speaking Rate (wpm): Among the four datasets with transcripts, PhoAudiobook has the lowest speaking rate of words per minute. This reflects the nature of the dataset, which features long-form audio where speakers naturally pause and rest.

Sampling Rate: All datasets except viVoice use a standard sampling rate of 16000 Hz, which is a widely used sampling rate for speech data.

PhoAudiobook is comparable in total duration to viVoice, however, it offers several advantages:

- **Text Normalization:** viVoice lacks text normalization, which limits its suitability for certain TTS models. In contrast, PhoAudiobook offers normalized transcripts, enhancing compatibility with these models.
- **Audio Quality:** The unnormalized audio waveforms in viVoice may cause quality issues like distortion and inconsistent volume. In contrast, PhoAudiobook ensures audio waveforms are normalized for consistent quality.
- **Speaker ID:** viVoice does not provide speaker IDs for individual audio samples, but uses YouTube channel names as a proxy. This approach can be problematic when a YouTube channel features multiple speakers, limiting the use of this dataset to models that do not require speaker identification. In contrast, PhoAudiobook provides distinct speaker IDs for each audio sample, ensuring its broader applicability for speaker-dependent tasks.

3 Empirical approach

3.1 Models & Training data augmentation

We conduct experiments using 3 state-of-the-art zero-shot TTS models: VALL-E (Wang et al., 2023), VoiceCraft (Peng et al., 2024), and XTTS-v2 (Casanova et al., 2024). (i) VALL-E, a pioneering language model-based approach, treats text-

to-speech (TTS) as a conditional language modeling task. It utilizes discrete acoustic tokens derived from a neural audio codec and leverages massive datasets to achieve impressive zero-shot, in-context learning capabilities. (ii) VoiceCraft, a token-infilling neural codec language model, excels in both speech editing and zero-shot TTS. It employs a Transformer decoder architecture with a novel token rearrangement procedure to generate high-quality speech. (iii) XTTS-v2 builds upon the Tortoise model (Betker, 2023), incorporating modifications for multilingual training and enhanced voice cloning. It excels in synthesizing speech for numerous languages, including low-resource ones.

To enhance data distribution and ensure our TTS models effectively handle shorter input text, we augment the PhoAudiobook training set with shorter audio clips. Specifically, we treat the PhoAudiobook training set, which consists of 940 hours, as a new raw dataset and apply our dataset creation process as detailed in Section 2.1. However, we omit (i) the step of merging short segments into longer ones and (ii) the step of excluding short samples. This augmentation phase results in an additional 554 hours of short audio, bringing the total to 940 + 554 = 1494 hours of audio for training. See implementation details on how we train VALL-E, VoiceCraft, and XTTS-v2 on this 1494-hour training set in Appendix A.

3.2 Evaluation setup

Baseline: The baseline model is viXTTS (Gia et al., 2024),[5] which is fine-tuned from the pre-trained XTTS-v2 on the viVoice dataset.

Test sets: In addition to using our PhoAudiobook "seen" and "unseen" speaker test sets, we also compare our models with viXTTS on the VIVOS test set (Luong and Vu, 2016), which contains 0.75 hours of short audio data from 19 speakers. Furthermore, we randomly select 8 speakers, totaling 0.5 hours of audio, from the viVoice dataset for testing. It is important to note that viVoice is available only as a single training dataset, without a predefined training/validation/test split. Consequently, this 0.5-hour viVoice audio set is in fact used for training the baseline viXTTS.

Metrics: To compare our models and the baseline, we use objective metrics including Word Error Rate (WER), Mel-Cepstral Distortion (MCD) and F0 Root Mean Square Error (RMSE_{F0}), as well as subjective metrics Mean Opinion Score (MOS) and Similarity MOS (SMOS).

Objective metrics provide quantifiable measures of specific aspects of synthesized speech:

- **Word Error Rate (WER):** This metric assesses the intelligibility of synthesized speech by calculating the edit distance between the transcription of the synthesized speech and the ground truth transcription. Specifically, it counts the number of insertions, deletions, and substitutions needed to turn one into the other. A lower WER indicates higher intelligibility. Here, we employ the ASR model `PhoWhisper-large` (Le et al., 2024) to generate the transcription of the synthesized speech.
- **Mel-Cepstral Distortion (MCD):** This metric quantifies the spectral difference between synthesized speech and the ground truth audio. A lower MCD value indicates higher spectral similarity and better quality. We use the `pymcd`[6] package to compute the MCD.
- **F0 Root Mean Square Error (RMSE_{F0}):** This metric measures the difference in fundamental frequency (F0) between synthesized speech and the ground truth audio. A lower RMSE_{F0} suggests a better matching of intonation and prosody. We use the `Amphion` (Li et al., 2025) toolkit to compute this value.

Subjective Metrics rely on human judgments to evaluate the overall quality and naturalness of synthesized speech.

- **Mean Opinion Score (MOS):** This metric assesses the overall quality of synthesized speech, taking into account factors such as naturalness, clarity, and listening effort. Human listeners rate the speech on a scale from 1 (very poor) to 5 (excellent).
- **Similarity MOS (SMOS):** This metric evaluates the perceived speaker similarity between the speech prompt and the generated speech. Listeners rate the similarity on a scale from 1 (completely different) to 5 (identical).

To conduct the subjective evaluation, we first randomly sample one audio file from each speaker in the test set. We then hire 10 native speakers

[5] https://huggingface.co/capleaf/viXTTS

[6] https://pypi.org/project/pymcd/

	Model	PAB-S	PAB-U	VIVOS	viVoice
WER ↓	Original	0.88	0.83	5.14	4.97
	VALL-E$_{PAB}$	24.96	12.90	**12.63**	13.58
	VoiceCraft$_{PAB}$	7.53	15.14	13.53	21.70
	XTTS-v2$_{PAB}$	**4.16**	**4.31**	37.81	**8.32**
	viXTTS	4.23	5.17	37.81	12.54
MCD ↓	VALL-E$_{PAB}$	7.50	8.28	10.13	8.70
	VoiceCraft$_{PAB}$	6.69	7.98	10.27	9.15
	XTTS-v2$_{PAB}$	**6.30**	**7.81**	**9.85**	**8.34**
	viXTTS	7.47	8.48	10.54	8.71
RMSE$_{F0}$ ↓	VALL-E$_{PAB}$	226.55	246.88	267.80	**223.56**
	VoiceCraft$_{PAB}$	**214.66**	247.54	**259.46**	233.68
	XTTS-v2$_{PAB}$	216.44	**242.51**	290.77	228.81
	viXTTS	249.54	271.70	338.59	238.05
MOS ↑	Original	4.61 ± 0.17	4.63 ± 0.16	4.41 ± 0.14	4.66 ± 0.20
	VALL-E$_{PAB}$	3.96 ± 0.29	**4.04** ± 0.28	3.44 ± 0.21	3.75 ± 0.38
	VoiceCraft$_{PAB}$	4.16 ± 0.21	3.75 ± 0.29	**3.85** ± 0.22	**3.98** ± 0.22
	XTTS-v2$_{PAB}$	**4.20** ± 0.20	3.89 ± 0.21	2.79 ± 0.21	3.98 ± 0.29
	viXTTS	4.05 ± 0.23	3.85 ± 0.25	2.37 ± 0.24	3.48 ± 0.44
SMOS ↑	Original	4.23 ± 0.23	3.90 ± 0.32	3.87 ± 0.24	3.34 ± 0.47
	VALL-E$_{PAB}$	**3.77** ± 0.24	3.46 ± 0.29	**3.35** ± 0.25	3.20 ± 0.38
	VoiceCraft$_{PAB}$	3.64 ± 0.30	3.32 ± 0.35	3.25 ± 0.25	**3.41** ± 0.36
	XTTS-v2$_{PAB}$	3.55 ± 0.27	**3.56** ± 0.29	3.03 ± 0.23	3.39 ± 0.41
	viXTTS	2.88 ± 0.28	2.63 ± 0.32	2.48 ± 0.23	3.11 ± 0.43

Table 2: Test results of different TTS models. Our models, "VALL-E$_{PAB}$", "VoiceCraft$_{PAB}$" and "XTTS-v2$_{PAB}$" are obtained by training VALL-E, VoiceCraft, and XTTS-v2 on our PhoAudiobook training data, respectively. "PAB-S" and "PAB-U" refer to the PhoAudiobook "seen" and "unseen" speaker test sets, respectively. The viXTTS model is fine-tuned from the pretrained XTTS-v2 using the entire viVoice dataset.

to rate the outputs for the in-distribution test sets (PAB-S, PAB-U) and 20 native speakers for the out-of-distribution test sets (VIVOS, viVoice), with all ratings on a scale from 1 to 5, using 0.5-point increments. To ensure fairness, we shuffle and anonymize the model names so that each listener is unaware of which model produces each sample.

4 Results

Table 2 presents the results obtained for our trained models and the baseline. It is clear that our XTTS-v2$_{PAB}$ consistently outperforms viXTTS across all metrics and test sets. For instance, on the viVoice set, XTTS-v2$_{PAB}$ achieves the best WER of 8.32, which is substantially lower than the 12.54 WER of viXTTS, even though viXTTS is tested on its own training data. Additionally, XTTS-v2$_{PAB}$ also produces substantially higher SMOS and RMSE$_{F0}$ scores compared to viXTTS in all test sets, indicating that the speech it generates more closely resembles the reference speaker. These results suggest that XTTS-v2$_{PAB}$ outputs more intelligible and natural-sounding speech that better captures the nuances of the target speaker's voice, for both long (PhoAudiobook and viVoice) and shorter (VIVOS) text inputs.

We observe a variation in the performance of different models across test sets. While VoiceCraft$_{PAB}$ and VALL-E$_{PAB}$ are less competent than XTTS-v2$_{PAB}$ on the test sets PAB-S, PAB-U and viVoice, they outperform XTTS-v2$_{PAB}$ on the VIVOS test set. Specifically, for PAB-S, PAB-U, and viVoice test sets, VALL-E$_{PAB}$ and VoiceCraft$_{PAB}$ underperform compared to XTTS-v2$_{PAB}$ in terms of WER, while achieving comparable results on other metrics such as MCD, RMSE$_{F0}$, MOS, and SMOS. However, on the VIVOS test set, XTTS-v2$_{PAB}$ and viXTTS perform significantly worse than VALL-E$_{PAB}$ and VoiceCraft$_{PAB}$ across all evaluation metrics. This indicates that VALL-E$_{PAB}$ and VoiceCraft$_{PAB}$ are more adept at handling short sentences, which are characteristic of the VIVOS test set. Upon manual inspection, we found that for short text inputs, XTTS-v2-based models – XTTS-v2$_{PAB}$ and viXTTS – often generate redundant or rambling speech at the end of the output. This suggests a potential architectural issue within the XTTS-v2 model itself, rather than a data-related problem, as both the viVoice dataset and the "augmented" PhoAudiobook training set contain short audio samples.

5 Conclusion

We have introduced PhoAudiobook, a comprehensive 941-hour high-quality dataset designed for Vietnamese text-to-speech (TTS) synthesis. Using this dataset, we conducted experiments with three leading zero-shot TTS models: VALL-E, VoiceCraft, and XTTS-v2. Our findings show that XTTS-v2 consistently outperforms its counterpart trained on the viVoice dataset across all metrics, highlighting the superiority of PhoAudiobook in enhancing model performance. Additionally, VALL-E and VoiceCraft demonstrate exceptional capability in handling short sentences.

Limitations

While models trained on PhoAudiobook show high performance on purely Vietnamese datasets, we have not evaluated their performance in code-switching scenarios where the input text includes both Vietnamese and English. Future research should investigate the models' ability to handle multilingual inputs to enhance their applicability in more diverse linguistic contexts.

Acknowledgments

This work was completed while all authors were at Movian AI, Vietnam. All datasets and models were downloaded, trained, and evaluated using Movian AI's resources.

We would like to express our sincere gratitude to Mr. Nguyen Nguyen for providing a toolkit that facilitated the process of downloading raw data from https://sachnoiviet.net.

References

Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural Voice Cloning with a Few Samples. In *Proceedings of NeurIPS*.

James Betker. 2023. Better speech synthesis through scaling.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. In *Proceedings of INTERSPEECH*.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. In *Proceedings of ICML*.

Yuan-Jui Chen, Tao Tu, Cheng chieh Yeh, and Hung-Yi Lee. 2019. End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning. In *Proceedings of INTERSPEECH*.

Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. 2020. Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings. In *Proceedings of ICASSP*.

Alexandre Défossez. 2021. Hybrid Spectrogram and Waveform Source Separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*.

Thinh Le Phuoc Gia, Tuan Pham Minh, Hung Nguyen Quoc, Trung Nguyen Quoc, and Vinh Truong Hoang. 2024. viVoice: Enabling Vietnamese Multi-Speaker Speech Synthesis.

Alexander Gutkin, Linne Ha, Martin Jansche, Knot Pipatsrisawat, and Richard Sproat. 2016. TTS for Low Resource Languages: A Bangla Synthesizer. In *Proceedings of LREC*.

Rongjie Huang, Chunlei Zhang, Yongqi Wang, Dongchao Yang, Jinchuan Tian, Zhenhui Ye, Luping Liu, Zehan Wang, Ziyue Jiang, Xuankai Chang, Jiatong Shi, Chao Weng, Zhou Zhao, and Dong Yu. 2024. Make-A-Voice: Revisiting Voice Large Language Models as Scalable Multilingual and Multitask Learners. In *Proceedings of ACL*.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *Proceedings of ICML*.

Thanh-Thien Le, Linh The Nguyen, and Dat Quoc Nguyen. 2024. PhoWhisper: Automatic Speech Recognition for Vietnamese. In *Proceedings of the ICLR 2024 Tiny Papers track*.

Jiaqi Li, Xueyao Zhang, Yuancheng Wang, Haorui He, Chaoren Wang, Li Wang, Huan Liao, Junyi Ao, Zeyu Xie, Yiqiao Huang, Junan Zhang, and Zhizheng Wu. 2025. Overview of the Amphion Toolkit (v0.2). *arXiv preprint*, arXiv:2501.15442.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In *Proceedings of AAAI*.

Yanqing Liu, Ruiqing Xue, Lei He, Xu Tan, and Sheng Zhao. 2022. DelightfulTTS 2: End-to-End Speech Synthesis with Adversarial Vector-Quantized Auto-Encoders. In *Proceedings of INTERSPEECH*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Hieu-Thi Luong and Hai-Quan Vu. 2016. A non-expert Kaldi recipe for Vietnamese Speech Recognition System. In *Proceedings of WLSI/OIAF4HLT*, pages 51–55.

Florian Lux, Julia Koch, and Ngoc Thang Vu. 2022. Low-Resource Multilingual and Zero-Shot Multispeaker TTS. In *Proceedings of AACL-IJCNLP*.

Phuong Pham Ngoc, Chung Tran Quang, and Mai Luong Chi. 2023. ADAPT-TTS: HIGH-QUALITY ZERO-SHOT MULTI-SPEAKER TEXT-TO-SPEECH ADAPTIVE-BASED FOR VIETNAMESE. *Journal of Computer Science and Cybernetics*, 39(2):159–173.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. 2024. VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. In *Proceedings of ACL*.

Anh Pham, Khanh Linh Tran, Linh Nguyen, Thanh Duy Cao, Phuc Phan, and Duong A. Nguyen. 2024. Bud500: A Comprehensive Vietnamese ASR Dataset.

Viet Thanh Pham, Xuan Thai Hoa Nguyen, Vu Hoang, and Thi Thu Trang Nguyen. 2023. Vietnam-Celeb: a large-scale dataset for Vietnamese speaker recognition. In *Proceedings of INTERSPEECH*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Ha Nguyen, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaëlle Laperrière, Mickael Rouvier, Renato De Mori, and Yannick Estève. 2024. Open-Source Conversational AI with SpeechBrain 1.0. *Journal of Machine Learning Research*, 25(333).

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. In *Proceedings of NeurIPS*.

Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid Transformers for Music Source Separation. In *Proceedings of ICASSP*.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Sheng Zhao, Tao Qin, Frank Soong, and Tie-Yan Liu. 2024. NaturalSpeech: End-to-End Text-to-Speech Synthesis With Human-Level Quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4234–4245.

VinBigData. 2023. VinBigData Shares 100-Hour Data for the Community.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv preprint*, arXiv:2301.02111.

Tao Wang, Jianhua Tao, Ruibo Fu, Jiangyan Yi, Zhengqi Wen, and Rongxiu Zhong. 2020. Spoken Content and Voice Factorization for Few-Shot Speaker Adaptation. In *Proceedings of INTERSPEECH*.

Yihan Wu, Xu Tan, Bohan Li, Lei He, Sheng Zhao, Ruihua Song, Tao Qin, and Tie-Yan Liu. 2022. AdaSpeech 4: Adaptive Text to Speech in Zero-Shot Scenarios. In *Proceedings of INTERSPEECH*.

A Implementation details

For the VALL-E implementation, the first phase is the data processing stage. We fine-tune a model based on the Vietnamese Wav2Vec 2.0 large model[7] for the Vietnamese dialect recognition task using our in-house data. The model achieves an accuracy of approximately 95% for the dialect recognition task. We then apply this model to the PhoAudiobook. To accurately determine the regional dialect for each speaker and reduce inference time, we sample about 20 audios for each speaker and feed them into the model. We assign the dialect of each speaker based on the region with the highest number of predicted audios. Because the VALL-E model is trained based on the phoneme level, we use the "phonemizer" package to convert the text into phonemes based on the dialect of each speaker.[8] For the audio, we also use the Audio CodeC encoder to compress raw audio into discrete tokens. The second phase is model training. We employ 12 Transformer-decoder layers, each with 1024 hidden units and 16 attention heads. We use a batch size corresponding to a maximum duration of 40 audio seconds, a base learning rate of 0.05, and 4 gradient accumulation steps. Our model is trained on 8 A100-40GB GPUs. Our implementation is based on the customized GitHub repository that reproduces the idea from the VALL-E paper.[9] We make modifications to this repository for our specific language and data.

Since VoiceCraft takes phoneme representations as input, we first convert our data to phonemes using the "phonemizer" package. We then append the derived Vietnamese phonemes to the existing English vocabulary and expand the text embedding layer to accommodate Vietnamese phonemes. The 830M_TTSEnhanced checkpoint is a public VoiceCraft model fine-tuned with a text-to-speech objective and serves as our starting point. Following the author's implementation,[10] we fine-tune this model using the AdamW optimizer with a learning rate of $1e^{-5}$ and a batch size of $25,000$ tokens, which corresponds to approximately 8.3 minutes of audio. We train the model on 4 A100-40GB GPUs for 16 epochs.

XTTS-v2 employs BPE for text encoding. To

[7] https://huggingface.co/nguyenvulebinh/wav2vec2-large-vi

[8] https://github.com/bootphon/phonemizer

[9] https://github.com/lifeiteng/vall-e/tree/main

[10] https://github.com/jasonppy/VoiceCraft

adapt the training to Vietnamese data, we use the same Vietnamese token list employed by Gia et al. (2024). We follow the training recipes provided in the coqui's TTS repository and fine-tune the public XTTS-v2 checkpoint trained for 16 languages.[11] We extend the character and audio length limits to accommodate audio segments up to 20 seconds in duration for training data. We use the AdamW optimizer with a learning rate of $5e^{-6}$, a batch size of 4, and fine-tune the model on a single A100-40GB GPU for 18 epochs.

For all these 3 models, we select the model checkpoint that obtains the best loss on the PhoAudiobook validation set.

[11] https://github.com/coqui-ai/TTS