VALL-E Revisited: A Replication Study Exploring Efficient Text-to-Speech Model Training with Limited Resources

Murad Bozik[1,2], Suzan Verberne[2], Joost Borekens[1,2]

[1]Daisys AI, [2]Leiden University

1 Introduction

Text-to-Speech (TTS) is a complex task in speech processing that aims to generate natural-sounding speech from written text. It is fundamentally a one-to-many sequence prediction problem, as a single text input can correspond to multiple valid speech outputs with variations in prosody, emphasis, and speaker characteristics.

In order to facilitate the transfer of speaking style, most models incorporate reference speech as additional input to the model architecture [1]. This approach allows the system to capture and reproduce specific voice characteristics and speaking styles. Such use of reference speech is also used in the context of transformer-based TTS architectures such as VALL-E [2]. In transformer-based TTS systems, the model creates a continuation of the reference speech by vocalizing the target text while maintaining style, acoustic environment, and speaker characteristics throughout the generated speech. This capability enables voice cloning, as the system can generate new speech in the voice of the reference speaker, resulting in natural-sounding output that closely matches the qualities of the reference audio.

The VALL-E [2] model, proposed by Wang et al., represents a significant advancement in transformer-based text-to-speech (TTS) technology. VALL-E's architecture is composed of two decoder-only transformer models, operating on discrete audio codes produced by the EnCodec model [3]. This design choice allows VALL-E to demonstrate zero-shot voice cloning capabilities using just 3 seconds of audio prompt. The VALL-E model utilizes EnCodec codes as an intermediate representation and employs the EnCodec decoder as its vocoder. These EnCodec codes consists of 8 codebooks such that speech audio is represented by a $(T, 8)$ matrix, where T denotes the time dimension. In VALL-E's two-stage architecture, the first transformer autoregressively predicts the audio represented as a sequence of codes from the first EnCodec codebook. Subsequently, the second Non-Autoregressive (NAR) transformer consecutively predicts the remaining 7 sequences. The use of discrete audio codes and the two-stage prediction process contribute to VALL-E's abbilitiy to perform zero-shot voice cloning with minimal reference audio, marking significant step forward in the field of TTS technology.

2 Motivation and Challenges

Our study is driven by two key factors. First, the original VALL-E implementation was not open-sourced, limiting the research community's ability to verify its capabilities and build upon its architecture. Second, we observe a trend in recent TTS models with similar architectures (operating on discrete audio codes) using increasingly larger datasets. For instance, the original VALL-E used 60,000 hours of LibriLight [4] data, BaseTTS [5] report using 100,000 hours of public domain speech data not accessible by research community, and Seed-TTS [6] claim to use datasets orders of magnitude larger. In contrast, our motivation is to create models that can achieve comparable performance using datasets that are orders of magnitude smaller and thereby support low-resource languages, language varieties, or domains. This approach also aims to develop more environmentally friendly models for research, reducing the computational resources required while maintaining high-quality speech synthesis capabilities. By doing so, we seek to make advanced TTS technology more accessible and sustainable for a broader range of applications and research contexts.

During our effort to create an open-source replication of VALL-e, we encountered several significant challenges: First, a major difficulty was accurately assessing the model's training status. To address this, we developed a window-based accuracy metric to evaluate whether the model was learning to predict the correct codes with potential delays. This metric is computed in quantiles (25%, 50%, 75%, and 100%) from the beginning of the sequence. Window-based accuracy metric provided insights that differed from traditional accuracy measures, making it challenging to evaluate the model's learning progress. Second, we faced issues with instability in the Non-Autoregressive (NAR) transformer during mixed-precision training. We addressed this issue with adjustments to model size and training parameters. Third, balancing inference sampling parameters for optimal speech generation proved to be a complex task, requiring careful tuning and experimentation.

3 Methodology and Findings

We implemented the VALL-E architecture from the initial paper description, and adapted it for smaller datasets by reducing the NAR model size without significant performance loss. We developed a flexible pipeline capable of processing diverse data. Key findings from our experiments are:

1. Data Efficiency: Experiments shows significant improvements in data efficiency compared to the original VALL-E implementation. Our model was trained on just 221 hours of clean bilingual audio, yet still produced high-quality synthesized speech. This finding suggests that the model architecture is more data-efficient than previously thought, potentially making it suitable for low-resource languages or domains.
2. Importance of Inference Mask: We discovered that applying the same mask calculation during inference as in training is essential for producing intelligible speech. Without the mask, the output becomes unintelligible, highlighting the sensitivity of the model to implementation details.
3. Sampler Parameter Sensitivity: We discovered that the model is sensitive to temperature and top-k settings during generation. Lower temperature and top-k values effectively scale the logits, resulting in generated speech with a slower speaking style and frequent pauses. These parameters significantly influence the generated speech's pace and fluency, proves the need for careful tuning to achieve natural sounding output.
4. Effect of EnCodec Code Dynamics: Our window-based accuracy metric revealed a consistent pattern of slightly decreasing accuracy across sequence length. Additionally, while window-based accuracy metric demonstrates a decreasing pattern, traditional accuracy metrics such as top-1 and top-10 continue to improve. This explains the model's evolving ability to balance between predicting repeated codes and anticipating changes in the audio sequence, despite increased complexity in longer sequences. This proves assessing model performance based on token prediction accuracy remains a challenging problem.

4 Conclusion

Our study provides valuable insights into the difficulty of implementing advanced TTS systems. We successfully adapted the architecture for smaller datasets, and gained deeper understanding of the model's learning progress through various sampling methods and the implementation of a window-based accuracy metric. The audio samples generated by our model can be compared to original VALL-E model output at our demo page: `https://valle-samples.speechtechlabs.com`.

References

[1] Y. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, “StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models,” *arXiv.org*, 2023.

[2] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, “Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers,” Jan. 2023, arXiv:2301.02111 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2301.02111

[3] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High Fidelity Neural Audio Compression,” Oct. 2022, arXiv:2210.13438 [cs, eess, stat]. [Online]. Available: http://arxiv.org/abs/2210.13438

[4] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, “Libri-Light: A Benchmark for ASR with Limited or No Supervision,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7669–7673, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/9052942/?arnumber=9052942

[5] M. Łajszczak, G. Cámbara, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Á. Martín-Cortinas, A. Abbas, A. Michalski, A. Moinet, S. Karlapati, E. Muszyńska, H. Guo, B. Putrycz, S. L. Gambino, K. Yoo, E. Sokolova, and T. Drugman, “BASE TTS: Lessons from building a billion-parameter Text-to-Speech model on 100K hours of data,” Feb. 2024, arXiv:2402.08093 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2402.08093

[6] P. Anastassiou, J. Chen, J. Chen, Y. Chen, Z. Chen, Z. Chen, J. Cong, L. Deng, C. Ding, L. Gao, M. Gong, P. Huang, Q. Huang, Z. Huang, Y. Huo, D. Jia, C. Li, F. Li, H. Li, J. Li, X. Li, X. Li, L. Liu, S. Liu, S. Liu, X. Liu, Y. Liu, Z. Liu, L. Lu, J. Pan, X. Wang, Y. Wang, Y. Wang, Z. Wei, J. Wu, C. Yao, Y. Yang, Y. Yi, J. Zhang, Q. Zhang, S. Zhang, W. Zhang, Y. Zhang, Z. Zhao, D. Zhong, and X. Zhuang, “Seed-TTS: A Family of High-Quality Versatile Speech Generation Models,” Jun. 2024, arXiv:2406.02430 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2406.02430