Beyond Labeled Datasets: Advancing TTS with Direct Preference Optimization on Unlabeled Speech Dataset

Andrii Zhuravlov[1], Volodymyr Sydorskyi[1],
[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
zhuravlov.andrii@lll.kpi.ua, v.sydorskyi@kpi.ua

Abstract

In our work, we enhance language model-based Text-to-Speech (TTS) training from unlabeled speech data using Direct Preference Optimization (DPO). Given the critical challenges related to the quality and quantity of data required for high-quality speech generation systems, it is essential to develop cost-effective approaches to training such models. We propose a two-stage fine-tuning approach, which extends traditional fine-tuning on texts generated by automatic speech recognition (ASR) models and incorporates direct preference optimization (DPO) along with dataset expansion using texts generated by large language models (LLMs). Experiments and comparisons conducted on two different datasets demonstrate that our approach achieves results comparable to traditional fine-tuning on human-labeled data. The code is publicly available on GitHub[1].

1 Introduction

In recent years, the quality of speech generation has significantly improved, largely due to advancements in high-quality audio quantizers such as HiFiCodec (Yang et al., 2023) and VQ-VAE, which was used in xTTS system (Casanova et al., 2024). These developments have enabled the use of Transformer architectures (Vaswani, 2017), which are known to perform well with large-scale datasets but are prone to overfitting on smaller datasets.

As a result, data collection and the quality of datasets remain critical challenges in the continued advancement of TTS models. One approach to increasing data availability involves using ASR models to automatically annotate raw audio data. However, this method compromises the quality of speech generation, as raw audio data is often of low quality and ASR models introduce recognition errors. To address these issues, a WV-MOS-based filtering method (Ogun et al., 2023) has been proposed to improve data set quality by filtering low-quality samples using WV-MOS models. Additionally, it has been demonstrated that raw audio data quality can be improved using noise-filtering systems (Ni et al., 2023; Hao et al., 2021), which boosts TTS model performance but complicates the preprocessing pipeline.

Fortunately, the Transformer architecture enables the application of techniques from NLP, particularly training pipelines for large language models (LLMs). Training LLMs generally consists of two stages: pretraining and fine-tuning. During pretraining, the model is trained on a large corpus of low-quality data, and in the fine-tuning stage, methods such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) and instruction tuning are used to improve the quality of model outputs. Tian et al. (2024) demonstrated a similar approach and showed that using the Direct Preference Optimization (DPO) algorithm is effective for TTS models. In Tian et al. (2024), preference alignment was guided by three metrics – WER (Whisper-large model (Radford et al., 2023)), SPK_SIM (RawNet (Jung et al., 2024)), and Proxy_MOS (UTMOS (Saeki et al., 2022)) – to evaluate preferences between sample pairs and there was shown a great boost on each metric. Moreover, was shown that improvement does not depend on the exact metric models. In another study (Hussain et al., 2025), preference pairs were constructed using the character error rate (CER) and cosine similarity (SSIM) metrics, along with a modified version of the DPO method – e Reward-aware Preference Optimization (RPO) – to enable more fine-grained preference calibration. The study also demonstrated that employing DPO or RPO for fine-tuning TTS models can lead to improvements in the overall quality of the resulting system.

Inspired by the previous findings (Tian et al., 2024; Hussain et al., 2025), in this work, we

[1]https://github.com/BirdWithDreams/beyond-labeled-datasets-tts

present a semi-supervised training strategy for a language model-based text-to-speech (TTS) system, aiming to reduce reliance on labeled data while maintaining high synthesis quality. We explored six different model training strategies on two distinct datasets. Additionally, we proposed a semi-supervised two-stage training strategy: first, a standard fine-tuning on ASR-labeled data, followed by DPO fine-tuning on a combination of original dataset ASR texts and LLM-generated texts.

Our results demonstrate that the proposed training strategy outperforms traditional fine-tuning on human-labeled data in two out of the three primary evaluation metrics. Additionally, we perform a human evaluation using the Comparative Mean Opinion Score (CMOS) methodology. The results indicate that the proposed approach is statistically comparable to conventional fine-tuning techniques on human labeled data. Furthermore, we adapted the popular xTTS framework to support training using the DPO method.

2 Method

2.1 Semi-supervised training methodology

Since collecting and annotating data for training a TTS model is a complex and resource-intensive process, we developed a fully unsupervised training method. The core idea of our approach consists of two stages: first, a standard fine-tuning of the base model on an ASR-labeled dataset, and second, the creation of a DPO dataset using this model.

Creating the DPO dataset does not require human involvement. All we need is a model checkpoint M for generation, a set of reference audios A, and a set of texts T. The latter initially consisted of ASR-labeled texts from the original datasets, which we further expanded with LLM-generated texts. The generation procedure is detailed in Appendix A.1. Then, for each $(a, t) \subset A \times T$ pair, we generated 10 audio $\{y_{a,t,1}, y_{a,t,2}, \ldots, y_{a,t,10}\}$ variants using the model M. This set of samples was ranked within each pair using three primary evaluation metrics, and the final ranking was determined using a harmonic mean aggregation, sorting the generated audio y from best to worst ($y^f_{a,t,1} \succ y^f_{a,t,2} \succ \ldots \succ y^f_{a,t,10}$) and based on this ranking, win-lose pairs were selected for preference alignment. The full procedure of DPO dataset construction is described in Appendix A.2.

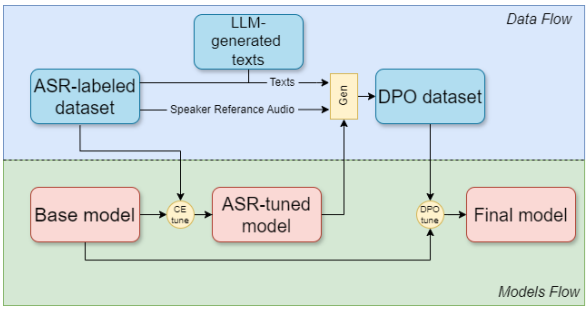

Figure 1: Schematic diagram of our model training pipeline.

2.2 Optimization Objectives

In our work, we used two approaches for training models (Figure 1). The first is standard fine-tuning as it was done in xTTS model (Casanova et al., 2024) with cross-entropy loss. Given an input text x (represented as a sequence of tokens), reference audio r, and target audio y (represented as a sequence of audio tokens), the model is trained to minimize the following loss function (a full derivation and detailed breakdown of its components is provided in Appendix B):

$$\mathcal{L}_{\text{FN}}(\pi_\theta) = \mathcal{L}_{\text{audio}}(\pi_\theta) + \alpha \cdot \mathcal{L}_{\text{text}}(\pi_\theta) \quad (1)$$

The second approach is Direct Preference Optimization, DPO (Rafailov et al., 2024). A win-lose pair dataset (r, x, y_w, y_l) was constructed, where the sequence y_w is preferred over y_l. To optimize the model on such data, we can use the $\mathcal{L}_{\text{DPO}}$ loss as it was described in the original paper (Rafailov et al., 2024).

Instead of using the pure $\mathcal{L}_{\text{DPO}}$ loss, we incorporated an additional $\mathcal{L}_{\text{text}}$ term, as it is known that TTS models perform better when optimized not only with respect to audio but also with respect to the given text itself. Finally, our DPO loss takes the form:

$$\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) = \mathcal{L}^{\text{audio}}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) + \mathcal{L}_{\text{text}}(\pi_\theta) \quad (2)$$

3 Experiment Setup

3.1 Model

We chose xTTSv2 (Casanova et al., 2024) as our base model. We use it because, firstly, it has an LM-based architecture, which is critically important for DPO fine-tuning. Secondly, it achieves state-of-the-art zero-shot performance in multiple

languages, including English. Additionally, it is highly stable during training, which is crucial for the reproducibility and reliability of our results.

3.2 Data

In our work, we used two datasets: The LJ Speech Dataset (Ito and Johnson, 2017) (denoted as $D_{\text{LJ}}^{\text{original}}$), representing a classic single-speaker audiobook-like dataset, and the CSTR VCTK Corpus (Veaux et al., 2017) (denoted as $D_{\text{VCTK}}^{\text{original}}$) as a multi-speaker dataset. Since these datasets contain manually annotated transcriptions, whereas we needed synthetic ones, we generated new transcriptions using Whisper-medium (Radford et al., 2023). We then split the data into a training set and a holdout set in a ratio 80/20 for The LJ Speech Dataset and 90/10 for CSTR VCTK Corpus, resulting in the following two datasets: $D_{\text{LJ}}^{\text{ASR}}$ and $D_{\text{VCTK}}^{\text{ASR}}$.

Additionally, we generated 15,000 texts using Llama 3.2 3b (Dubey et al., 2024). During generation, we employed a specialized text attribute combinator (considering factors such as length, topic, domain, complexity, etc.) to ensure maximum diversity in the generated texts (see Fig. 2). These texts were later used to augment the original datasets, enhancing their variability and robustness.

3.3 Metrics for DPO dataset

To evaluate the models and construct the DPO dataset, we used three main metrics: intelligibility (WER), speaker similarity (SS), and Proxy MOS (PMOS). The following models were used to calculate these metrics: Whisper-Medium (Radford et al., 2023) for WER, ECAPA2 Speaker Embedding Extractor (Thienpondt and Demuynck, 2023) for SS, and UTMOS (Saeki et al., 2022) for PMOS. Model validation was performed on the holdout subsets of our $D_{\text{LJ}}^{\text{original}}$ and $D_{\text{VCTK}}^{\text{original}}$ datasets.

3.4 Experiments

For each dataset group, D_{LJ} and D_{VCTK}, the following fine-tuning experiments were conducted:

1. Fine-tuning (FN) of the base xTTSv2 model on D^{original}.
2. Fine-tuning (FN) of the base xTTSv2 model on D^{ASR}.
3. DPO fine-tuning of the base xTTSv2 model on D^{DPO}.
4. DPO fine-tuning of the base xTTSv2 model on $D^{\text{DPO}} + D^{\text{Generated}}$.
5. DPO fine-tuning of the model from the corresponding checkpoint (LJ-ASR or VCTK-ASR) on D^{DPO}.
6. DPO fine-tuning of the model from the corresponding checkpoint (LJ-ASR or VCTK-ASR) on $D^{\text{DPO}} + D^{\text{Generated}}$.

Let's call these experiments `L1-6` for D_{LJ} group of datasets and `V1-6` for D_{VCTK} group. The validation results for each setup are presented in Table 1 and 2 for the D_{LJ} and D_{VCTK} dataset groups, respectively.

Table 1: Model Performance on D_{LJ} dataset group

Model	WER ↓	SS ↑	PMOS ↑
Base xTTSv2	0.071 ± 0.008	0.423 ± 0.003	3.68 ± 0.016
L1 (Original)	0.056 ± 0.014	$\mathbf{0.481 \pm 0.003}$	3.816 ± 0.013
L2 (ASR)	0.064 ± 0.010	0.478 ± 0.003	3.79 ± 0.013
L3 (DPO)	$\mathbf{0.043 \pm 0.003}$	0.445 ± 0.003	3.733 ± 0.012
L4 (DPO)	0.064 ± 0.011	0.465 ± 0.002	$\mathbf{3.959 \pm 0.010}$
L5 (DPO)	0.110 ± 0.012	0.432 ± 0.003	2.821 ± 0.011
L6 (DPO)	0.224 ± 0.035	0.417 ± 0.003	2.392 ± 0.012

Table 2: Model Performance on D_{VCTK} dataset group

Model	WER ↓	SS ↑	PMOS ↑
Base xTTSv2	0.020 ± 0.004	0.481 ± 0.014	3.895 ± 0.026
V1 (Original)	0.041 ± 0.007	$\mathbf{0.500 \pm 0.014}$	3.685 ± 0.029
V2 (ASR)	0.055 ± 0.009	0.494 ± 0.014	3.630 ± 0.030
V3 (DPO)	0.014 ± 0.003	0.471 ± 0.015	4.009 ± 0.022
V4 (DPO)	$\mathbf{0.013 \pm 0.003}$	0.482 ± 0.016	$\mathbf{4.108 \pm 0.019}$
V5 (DPO)	0.273 ± 0.037	0.412 ± 0.014	2.662 ± 0.047
V6 (DPO)	0.087 ± 0.013	0.453 ± 0.014	3.324 ± 0.043

3.5 Fine-Tuning vs. DPO Fine-Tuning

When comparing these two training methods, the first noticeable trend is that standard fine-tuning achieves the best SS metric across both datasets: 0.481 for D_{LJ} (`L1`) and 0.5 for D_{VCTK} (`V1`). As expected, training on ASR-labeled data (`L2`, `V2`) performs worse than training on human-labeled data (`L1`, `V1`). However, DPO training on ASR-labeled data (`L3`, `V3`) either outperforms or at least matches traditional fine-tuning with a cross-entropy objective on human-labeled data (`L1`, `V1`).

Interestingly, the best results in speech naturalness (PMOS metric) are achieved when the dataset is expanded with LLM-generated data (`L4`, `V4`), even surpassing a PMOS score of 4. Regarding intelligibility (WER metric), the best performance in the D_{VCTK} dataset group (WER 0.013) is also obtained with DPO tuning on the expanded

dataset (V4), outperforming both classical fine-tuning on human-labeled data (V1, WER 0.04) and the xTTSv2 baseline (WER 0.02).

For the D_{LJ} dataset group, the best WER score is achieved by the L3 model (WER 0.041), outperforming both L1 (WER 0.053) and the baseline (WER 0.071). However, the DPO fine-tune on the expanded dataset (L4) achieves results similar to traditional fine-tuning on human-labeled data and worse than DPO tune on unexpanded dataset (L3). This behavior can be explained by the nature of the The LJ Speech Dataset (Ito and Johnson, 2017). This dataset consists of audiobook recordings where audio is sometimes segmented inaccurately, resulting in partial sentences, such as only the beginning or end of a sentence like "`According to Secretary Dillon,`" or "`iron and the like in combination with phosphoric, sulphuric and other acids.`". Expanding the D_{LJ} dataset with LLM-generated texts, which consist of fully formed sentences, does not necessarily improve model performance within the D_{LJ} dataset.

In contrast, the CSTR VCTK Corpus (Veaux et al., 2017) dataset, which was created by reading newspaper sentences rather than slicing pre-existing audio, is more aligned with the way LLM-generated texts are structured. This explains why in the D_{VCTK} dataset group, fine-tuning on the expanded dataset (V4) yields better results (WER 0.013, SS 0.481, PMOS 4.108) than fine-tuning on standard texts (V3) (WER 0.014, SS 0.471, PMOS 4.009).

3.6 Effects of ASR Checkpoint Initialization

Comparing experiments 5-6 with 3-4, we observe that fine-tuning from the ASR checkpoint consistently yields worse results than fine-tuning from the base model on the same data. L5 and V5 show much higher WER (0.109 and 0.269, respectively) and lower PMOS (2.821 and 2.665). This can be explained by the fact that standard fine-tuning narrows the generation space, whereas DPO fine-tuning only adjusts the probability distribution within that space without altering it. In other words, the "softer" DPO fine-tuning from a checkpoint with greater generation variability leads to better results than fine-tuning from a checkpoint with lower variability. This holds true even though, in the latter case, the model was explicitly trained to reproduce the distribution of a specific dataset.

However, we observe that in both cases (L4, L6 and V4, V6) adding AI-generated texts improves model performance across almost all metrics (except for WER in the L3-4 cases), supporting our hypothesis that expanding the dataset with AI-generated texts positively impacts model quality.

3.7 CMOS validation

To further evaluate the proposed approach, we conducted a CMOS (Comparative Mean Opinion Score) validation following the methodology detailed in Appendix C.1. CMOS evaluation was conducted on four experimental pairs: (1 vs. 4), (2 vs. 4), (1 vs. 6), and (2 vs. 6). The evaluation considers two criteria: speaker similarity (SS), and a combined metric reflecting both naturalness and intelligibility (CM). Results are presented in Table 3.

Table 3: Method Pair Comparison Data

Comparison	SS	CM
Exp. 1 vs Exp. 4	-0.12	-0.06
Exp. 1 vs Exp. 6	0.88	0.21
Exp. 2 vs Exp. 4	-0.21	-0.17
Exp. 2 vs Exp. 6	0.75	0.14

Positive values in Table 3 indicate that the first experiment in the pair is preferred over the second, while negative values indicate the opposite.

In the comparison between Exp. 1 and Exp. 4, the SS metric marginally favors Exp. 4 (-0.12), while the combined metric indicates near equivalence (-0.06). Similarly, for the Exp. 2 vs. Exp. 4 comparison, both metrics slightly favor Exp. 4, with -0.21 for SS and -0.17 for CM.

More substantial differences are observed with Exp. 6. In both (1 vs. 6) and (2 vs. 6) comparisons, the metrics are positive (e.g., 0.88 and 0.75 for SS and combined in 1 vs. 6), indicating that the classical tuning baselines were preferred. These results support our conclusions based on automatic validation metrics (WER, SS, PMOS). Additionally, we performed a statistical significance analysis of the results, detailed in Appendix C.2. We also provide a detailed subgroup CMOS analysis in Appendix C.3. Overall, the CMOS evaluation indicates that the proposed method, particularly in Exp. 4, achieves quality comparable to conventional fine-tuning using human-labeled data.

4 Conclusion

We have developed a two-stage training strategy for TTS models based on DPO fine-tuning. We proposed a fully unsupervised training pipeline for TTS models and demonstrated that it can achieve results comparable to traditional supervised fine-tuning on human-labeled data. This approach significantly reduces costs, as manual annotation requires substantial resources and time. Therefore, our method is more efficient without sacrificing model quality.

Additionally, we showed that expanding original datasets with LLM-generated texts substantially improves the naturalness (PMOS) of generated audio while having a mixed impact on intelligibility (WER), which requires further investigation across different data types and datasets.

5 Limitations

Given the limitations of our work, we used the high-quality xTTSv2 model as our baseline. For future research, it would be valuable to train several models from scratch – one on ASR-labeled data and one on human-labeled data and compare how DPO fine-tuning affects their quality. Another interesting direction is to compare our method of constructing win-lose pairs for DPO with human-based pair selection.

Our pipeline involves components that may introduce or amplify societal biases:

1. ASR-Induced Bias: we rely on an Automatic Speech Recognition (ASR) model (Whisper-medium) to generate transcripts for unlabeled audio. It is well-documented that ASR systems can have higher error rates for speakers with non-native accents, certain dialects, or speech impediments. Such transcription errors may degrade the quality of synthesized speech for already underrepresented groups, potentially reinforcing existing biases in the system.

2. LLM-Induced Bias: The use of a Large Language Model (Llama 3) to generate supplementary text for training introduces the risk of inheriting its intrinsic biases. While we employed an attribute combinator to encourage text diversity (Appendix A.1), the generated content may still reflect dominant cultural viewpoints or stereotypes present in the LLM's training data.

Future work should involve auditing the model's performance across more diverse demographic groups and developing methods to mitigate any identified biases.

6 Ethical concerns

The development of advanced Text-to-Speech (TTS) technologies, such as the one presented in this paper, carries significant societal implications that warrant careful consideration. We are committed to the responsible advancement of AI and outline the primary ethical concerns related to our work below.

The most significant risk associated with high-fidelity TTS is the potential for misuse in creating synthetic audio, often referred to as "deepfakes."

- Using unauthorized voice synthesis to impersonate someone for fraudulent purposes, such as deceiving individuals or bypassing voice authentication systems.

- Disinformation and propaganda involve fabricating audio evidence to spread misinformation, defame individuals, or manipulate public opinion.

- Generating non-consensual audio content to harass or bully.

While our research aims to advance machine learning methodology, we recognize this dual-use nature. We advocate for the development and adoption of robust safeguards, such as audio watermarking techniques and detection models for synthetic speech, which should accompany any deployment of this technology in real-world applications.

Acknowledgments

We are especially grateful to the Armed Forces of Ukraine – without their resilience and protection, this work would not have been possible.

This work is part of the bachelor's research conducted by Andrii Zhuravlov under the supervision of Volodymyr Sydorskyi at the Institute for Applied System Analysis, Department of Artificial Intelligence, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute."

During the preparation of this work, the authors used OpenAI's ChatGPT in order to improve clarity, coherence, and LaTeX formatting. After using

this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xiang Hao, Xiangdong Su, Radu Horaud, and Xiaofei Li. 2021. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6633–6637. IEEE.

Shehzeen Hussain, Paarth Neekhara, Xuesong Yang, Edresson Casanova, Subhankar Ghosh, Mikyas T Desta, Roy Fejgin, Rafael Valle, and Jason Li. 2025. Koel-tts: Enhancing llm based speech generation with preference alignment and classifier free guidance. *arXiv preprint arXiv:2502.05236*.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Jee-weon Jung, Wangyou Zhang, Jiatong Shi, Zakaria Aldeneh, Takuya Higuchi, Barry-John Theobald, Ahmed Hussen Abdelaziz, and Shinji Watanabe. 2024. Espnet-spk: full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models. *arXiv preprint arXiv:2401.17230*.

Zhaoheng Ni, Sravya Popuri, Ning Dong, Kohei Saijo, Xiaohui Zhang, Gael Le Lan, Yangyang Shi, Vikas Chandra, and Changhan Wang. 2023. Exploring speech enhancement for low-resource speech synthesis. *arXiv preprint arXiv:2309.10795*.

Sewade Ogun, Vincent Colotte, and Emmanuel Vincent. 2023. Can we use common voice to train a multi-speaker tts system? In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 900–905. IEEE.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Jenthe Thienpondt and Kris Demuynck. 2023. Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Jinchuan Tian, Chunlei Zhang, Jiatong Shi, Hao Zhang, Jianwei Yu, Shinji Watanabe, and Dong Yu. 2024. Preference alignment improves language model-based tts. *arXiv preprint arXiv:2409.12403*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 6:15.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*.

A Experiment Details

A.1 Generation of LLM texts

To enhance the diversity of LLM-generated texts, we used a specialized attribute combinator to construct prompts for the LLM. Fig. 2 shows what attributes, sub-attributes and constraints were used to create high variability of generated texts.

All of this enables the creation of a highly diverse vocabulary, addressing one of the key challenges in TTS model training – bias between overrepresented and underrepresented words. At the same time, using LLM-generated texts helps to fill in gaps in the vocabulary and allows for fine control over text types and formats. This can be particularly useful when fine-tuning a model for a highly specialized domain with limited original data.

A.2 Creating DPO datasets

In the first stage, the LJ-ASR and VCTK-ASR models were trained using standard fine-tuning on ASR-generated texts from the base xTTSv2 model. Then, using the latest checkpoints of these models, the datasets $D_{\text{LJ}}^{\text{DPO}}$, $D_{\text{LJ}}^{\text{Generated}}$, $D_{\text{VCTK}}^{\text{DPO}}$, and $D_{\text{VCTK}}^{\text{Generated}}$ were constructed.

Method for Constructing $D_{\text{LJ}}^{\text{Generated}}$: a selection of audio samples was taken from the original dataset $D_{\text{LJ}}^{original}$, and AI-generated texts were evenly distributed among them. For each (audio, text) pair, 10 samples $(y_{a,t,1}, y_{a,t,2}, \ldots, y_{a,t,10})$ were generated using the LJ-ASR model, and evaluation metrics were computed for each sample using our evaluation models. Notice, that each y is not a generated audio, but a sequence of audio codes produced by LM head (see Casanova et al. (2024)).

Next, these samples were ranked from best to worst according to each metric $\left(y_{a,t,1}^{\text{wer}} \succ y_{a,t,2}^{\text{wer}} \succ \ldots \succ y_{a,t,10}^{\text{wer}}\right)$. Based on their ranking, a normalized score between 0 and 1 was assigned to each sample.

$$metric\_rank = place/10$$

To determine the final ranking, we calculated harmonic mean of our metrics' ranks:

$$f\_rank = \frac{3}{\frac{1}{wer\_rank} + \frac{1}{ss\_rank} + \frac{1}{mos\_rank}}.$$

Then, based on its values, the preferred (y_w) and less preferred (y_l) samples were selected. We choose them as the second sample from each edge, mean $y_w = y_{a,t,2}^{\text{f\_rank}}$ and $y_l = y_{a,t,9}^{\text{f\_rank}}$. The most extreme samples, the absolute best $(y_1^{\text{f\_rank}})$ and worst $(y_{10}^{\text{f\_rank}})$, were excluded to ensure that the preference optimization for the model was not overly obvious. Following this process, the $D_{\text{LJ}}^{\text{Generated}}$ dataset was constructed: (a, t, y_w, y_l), where a is reference audio sample, t - reference text, y_w - preferred sequence of audio codes and y_l - non-preferred sequence of audio codes.

Method for Constructing $D_{\text{VCTK}}^{\text{Generated}}$: Since $D_{\text{VCTK}}^{\text{original}}$ contains 108 unique speakers and 13,000 unique texts—where different speakers may read the same text—the dataset includes a total of 44,000 (speaker, text) pairs. Each speaker has between 200 and 500 recordings. We decided to construct the DPO version of this dataset in a similar manner. Our 15,000 LLM-generated texts were evenly distributed among all speakers, with repetitions, ensuring that each speaker had an average of 500 unique texts. The subsequent sample generation, ranking, and win-lose pair selection followed the same approach as for $D_{\text{LJ}}^{\text{Generated}}$, with the VCTK-ASR model used during sample generation.

Construction of $D_{\text{LJ}}^{\text{DPO}}$ **and** $D_{\text{VCTK}}^{\text{DPO}}$: The datasets $D_{\text{LJ}}^{\text{DPO}}$ and $D_{\text{VCTK}}^{\text{DPO}}$ were constructed similarly to $D_{\text{LJ}}^{\text{Generated}}$ and $D_{\text{VCTK}}^{\text{Generated}}$, with the key difference that the texts were taken from the original $D_{\text{LJ}}^{\text{original}}$ and $D_{\text{VCTK}}^{\text{original}}$ datasets.

B Objectives definitions

Classical cross-entropy (CE) loss on text and audio tokens:

$$\begin{aligned}\mathcal{L}_{\text{FN}}(\pi_\theta) &= \mathcal{L}_{\text{audio}}(\pi_\theta) + \alpha \cdot \mathcal{L}_{\text{text}}(\pi_\theta) \\ &= -\mathbb{E}_{(x,r,y)\sim\mathcal{D}} \log \pi_\theta(y \mid x, r) \\ &\quad - \alpha \cdot \mathbb{E}_{(x,r,y)\sim\mathcal{D}} \log \pi_\theta\left(x^t \mid x^{t-1}, r\right)\end{aligned} \tag{3}$$

DPO loss from the original paper (Rafailov et al., 2024):

$$\begin{aligned}&\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = \\ &= -\mathbb{E}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]\end{aligned} \tag{4}$$

where π_θ is the model that is being optimized and π_{ref} is the original model.

Final objective for second stage of proposed method:

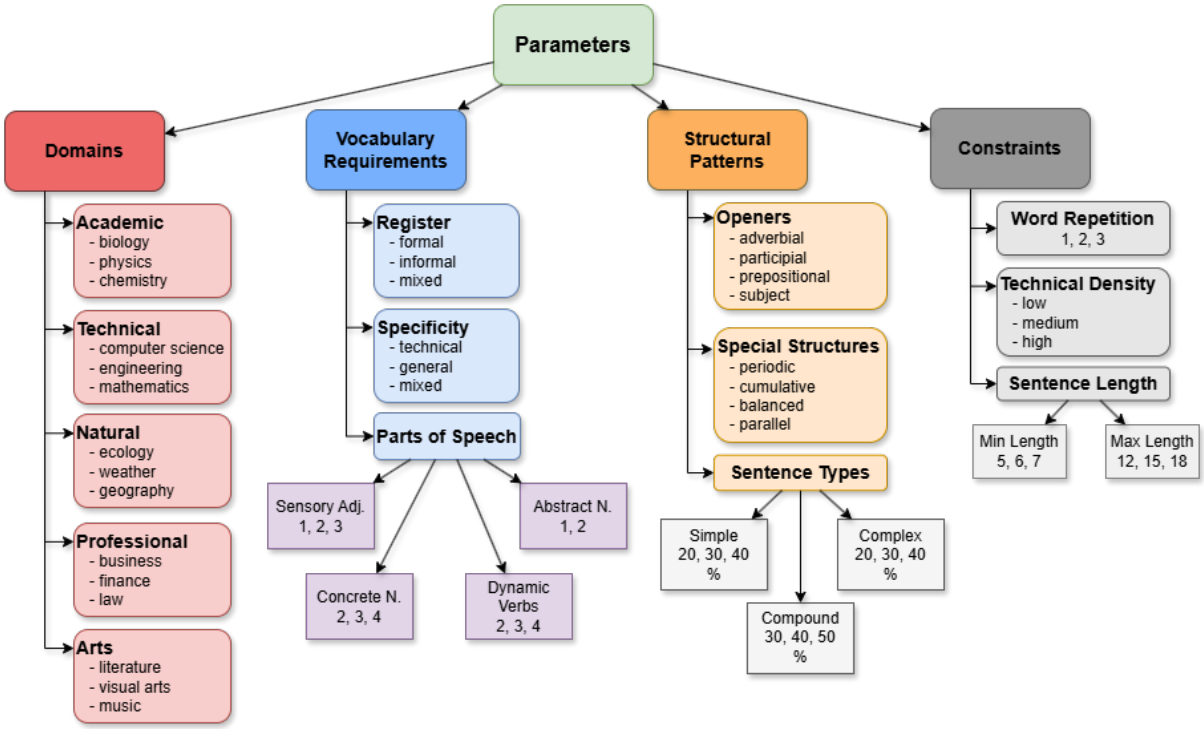

Figure 2: Combinator's attributes diagram

$$\begin{aligned}\mathcal{L}\left(\pi_\theta;\pi_{\text{ref}}\right) &= \mathcal{L}_{\text{DPO}}^{\text{audio}}\left(\pi_\theta;\pi_{\text{ref}}\right)+\mathcal{L}_{\text{text}}\left(\pi_\theta\right)\\ &= -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(y_w\mid x\right)}{\pi_{\text{ref}}\left(y_w\mid x\right)}-\beta\log\frac{\pi_\theta\left(y_l\mid x\right)}{\pi_{\text{ref}}\left(y_l\mid x\right)}\right)\right]\\ &\quad -\alpha\mathbb{E}\log\pi_\theta\left(x^t\mid x^{t-1},r\right)\end{aligned} \tag{5}$$

C CMOS validation details

C.1 Validation methodology

To facilitate the CMOS validation process, an automated service was developed to efficiently and conveniently collect user feedback. This service operates in a fully automated mode and presents evaluation tasks in a user-friendly format. The selection of speakers and texts for CMOS validation was intentionally made diverse: from the VCTK dataset, two speakers (one male and one female) were chosen for each of five distinct accents—American, British, Indian, Irish, and Scottish—resulting in a total of ten speakers, with an additional speaker selected from the LJ-Speech dataset. For each speaker, four different texts were selected from the validation subsets of the original datasets ($D_{\text{LJ}}^{\text{original}}$ and $D_{\text{VCTK}}^{\text{original}}$), and synthetic audio was generated for these texts using the corresponding TTS model.

The survey methodology is as follows: for each participant, the service randomly selects three speakers from the ten available in $D_{\text{VCTK}}^{\text{original}}$ and adds one speaker from $D_{\text{LJ}}^{\text{original}}$. The participant is then presented with 32 evaluation items: four method pairs × four speakers × four comparisons per speaker. Each evaluation item is structured as follows: "Please assess which of the two audio samples better corresponds to the reference recording according to a specific criterion, using a scale from -3 to 3, where 3 indicates that the first sample is significantly better, 0 means both are approximately equal, and -3 indicates that the second sample is significantly better."

C.2 Statistical significance analysis

To assess the reliability of the CMOS evaluation, we conducted a statistical significance test by evaluating the null hypothesis that the mean CMOS score is zero. A p-value above the significance threshold (0.05) indicates that the compared models are statistically equivalent, whereas a value below the threshold suggests a significant preference for one model over the other. The outcomes of this analysis for the SS and CM metrics are presented in Table 4 and Table 5, respectively.

The statistical analysis of the CMOS metrics (Table 4 and Table 5) reveals that Experiment 6, which implements the proposed method, significantly underperforms in speaker similarity compared to tra-

Table 4: Statistical significance of the SS metric

	1–4	**1–6**	**2–4**	**2–6**
Mean	-0.12	0.88	-0.21	0.75
t-test p-val	0.4	$\mathbf{0.001}$	0.22	$\mathbf{0.007}$
Wilcoxon p-val	0.5	$\mathbf{0.001}$	0.25	$\mathbf{0.015}$
Sample size	83	80	76	73

Table 5: Statistical significance of the CM metric

	1–4	**1–6**	**2–4**	**2–6**
Mean	-0.06	0.21	-0.17	0.14
t-test p-val	0.74	0.34	0.39	0.52
Wilcoxon p-val	0.74	0.34	0.35	0.49
Sample size	83	80	76	73

ditional approaches. In particular, Experiments 1 and 2 show strong and statistically significant advantages over Experiment 6 in the SS metric, with p-values well below 0.001. In contrast, comparisons involving Experiment 4 do not exhibit significant differences, indicating that this configuration achieves perceptual speaker similarity comparable to traditional fine-tuning.

For the Combined Metric (CM), none of the comparisons across experimental conditions yield statistically significant differences (all $p > 0.3$), suggesting that all methods perform similarly in terms of overall speech naturalness and text accuracy. These findings indicate that the proposed method, especially in Experiment 4, maintains competitive perceptual quality, while Experiment 6 demonstrates limited effectiveness in preserving vocal identity.

C.3 CMOS on specific groups

To further investigate the behavior of the proposed models across different speaker characteristics, we conducted a stratified CMOS analysis by accent and gender. As described in Appendix C.1, we examined the same four experimental method pairs: (1 vs. 4), (2 vs. 4), (1 vs. 6), and (2 vs. 6). For each pair, two criteria were evaluated: speaker similarity (SS) and a composite metric capturing clarity, naturalness, and intelligibility (CM). Results are presented in Table 6 and Table 7.

Overall, positive values in Tables 6 and 7 indicate a preference for the first method in each comparison, while negative values indicate preference for the second.

The proposed Method 4 (DPO training with generated data) demonstrates advantages in perceptual quality (CM) across several accents, with moderate gains in speaker similarity (SS). Compared

Table 6: CMOS Results by Accent

Method Pair	**Accent**	**SS**	**CM**	**N**
1–4	American	-0.545	0.818	11
	English	-0.235	-0.706	17
	Indian	0.300	0.500	10
	Irish	-0.533	-0.400	15
	Scottish	0.444	-0.111	9
	lj	0.048	-0.095	21
1–6	American	1.300	0.800	5
	English	1.400	-1.200	5
	Indian	0.933	0.333	15
	Irish	1.231	0.000	13
	Scottish	0.400	-0.050	20
	lj	1.000	0.647	17
2–4	American	0.833	-0.333	6
	English	-0.600	0.000	10
	Indian	-0.263	-0.526	19
	Irish	-0.333	0.000	12
	Scottish	0.545	0.364	11
	lj	-0.667	-0.278	18
2–6	American	1.417	0.417	12
	English	0.909	0.091	11
	Indian	-0.071	-0.214	14
	Irish	0.667	0.111	9
	Scottish	0.444	0.222	9
	lj	1.056	0.222	18

to the baseline trained on human-annotated data (Method 1), Method 4 achieves better CM scores for the LJ speaker (-0.095), English (-0.706), and Irish (-0.400), and also shows improved SS for English (-0.235) and Irish (-0.533), suggesting enhanced or preserved speaker identity. Relative to the ASR-supervised baseline (Method 2), Method 4 again receives more favorable CM values for American-accented speech (-0.333) and Scottish (-0.364), along with strong SS improvements for English (-0.600) and the LJ speaker (-0.667), highlighting its robustness on several accent groups. However, performance on some accents, such as Indian and American in the 1–4 comparison, remains challenging. By contrast, Method 6 (DPO with generated data initialized from a pretrained checkpoint) mostly underperforms relative to both baseline methods across individual accent groups, showing less consistent gains in either CM or SS. It is also important to note the variability in group sizes (N), with some accent groups containing relatively few samples. This limits the statistical robustness of per-accent conclusions and calls for caution when interpreting fine-grained differences.

Gender-based analysis further supports the effectiveness of the proposed Method 4, particularly for female speakers. Compared to the human-

Table 7: CMOS Results by Gender

Method Pair	Gender	SS	CM	N
1–4	F	−0.292	−0.250	24
	M	−0.105	0.079	38
	lj (F)	0.048	−0.095	21
1–6	F	0.656	0.031	32
	M	1.032	0.161	31
	lj (F)	1.000	0.647	17
2–4	F	−0.379	−0.069	29
	M	0.241	−0.207	29
	lj (F)	−0.667	−0.278	18
2–6	F	0.300	0.150	30
	M	0.857	0.086	35
	lj (F)	1.056	0.222	18

annotated baseline (Method 1), Method 4 achieves better CM scores for female speakers (-0.250) and the LJ speaker (-0.095), while also improving SS for females (-0.292), indicating that the proposed approach is preferred in terms of both perceptual quality and speaker similarity. For male speakers, results are more mixed: while SS is slightly better (-0.105), the CM score (0.079) indicates a mild preference for the baseline. In comparison to the ASR-supervised baseline (Method 2), Method 4 again shows lower CM for female (-0.069) and male (-0.207) speakers, and achieves a strong improvement for the LJ speaker (-0.278), with consistent SS gains for females (-0.379) and the LJ speaker (-0.667), reinforcing the robustness of Method 4 for female voices. In contrast, Method 6 performs worse than both baselines across all gender groups. CM scores are consistently positive when compared to both Method 1 and Method 2, indicating that listeners preferred the baseline systems in terms of clarity, naturalness, and intelligibility. SS values also show degradation, with all comparisons yielding positive scores, suggesting less accurate speaker identity preservation. As with the accent-based analysis, these observations should be interpreted with caution due to relatively small group sizes (N), especially for the LJ speaker.

To complement the CMOS evaluation, we conducted statistical significance testing on the speaker similarity (SS) and clarity/naturalness (CM) scores within each subgroup. For every experimental method pair and demographic subgroup (by gender and accent), we applied one-sample t-tests and Wilcoxon signed-rank tests against a null hypothesis of zero (i.e., no perceived difference between systems). The resulting p-values are presented in Table 8 (gender) and Table 9 (accent).

The statistical significance analysis supports the earlier observations (Table 4 and Table 5). For Method 4, p-values across most gender groups are above the 0.05 threshold in both SS and CM comparisons against baseline Methods 1 and 2, indicating no statistically significant difference and suggesting that the proposed method performs comparably to the baselines. In contrast, Method 6 consistently shows statistically significant differences in SS when compared to both baselines (e.g., $p < 0.05$ for both males and females), pointing to a degradation in speaker similarity. For CM, however, most comparisons yield p-values above 0.05, implying that the perceptual quality of Method 6 is not significantly different from the baselines despite the SS drop.

For accent-based comparisons, the majority of p-values also exceed 0.05, which may reflect a lack of statistical power due to small sample sizes within individual accent groups. Nevertheless, a few accents (e.g., Irish and American in 1–6 and 2–6 pairs) show marginal or significant effects, particularly in SS, indicating that accent-specific behavior may warrant closer examination in future studies with larger cohorts.

These findings underscore the importance of subgroup-level analysis in evaluating TTS systems. Listener demographics—such as gender and accent—can influence judgments of speaker similarity and perceptual quality, and adequate subgroup representation is crucial to draw robust, generalizable conclusions.

Table 8: Statistical Significance of CMOS Scores by Gender

Method Pair	Gender	SS t-p	SS w-p	CM t-p	CM w-p
1–4	M	0.612	0.618	0.761	0.766
	F	0.307	0.349	0.434	0.532
	lj (F)	0.867	0.769	0.820	0.744
1–6	M	0.0001	0.0006	0.643	0.603
	F	0.0001	0.0006	0.662	0.606
	lj (F)	0.063	0.078	0.287	0.223
2–4	M	0.452	0.504	0.546	0.552
	F	0.163	0.189	0.828	0.729
	lj (F)	0.014	0.023	0.508	0.520
2–6	M	0.0001	0.0006	0.782	0.776
	F	0.410	0.392	0.679	0.622
	lj (F)	0.015	0.023	0.664	0.668

Table 9: Statistical Significance of CMOS Scores by Accent

Method Pair	Accent	SS t-p	SS w-p	CM t-p	CM w-p
1–4	Scottish	0.498	0.531	0.834	1.000
	Indian	0.343	0.531	0.363	0.424
	Irish	0.072	0.072	0.361	0.404
	lj	0.867	0.769	0.820	0.744
	American	0.216	0.030	0.156	0.055
	English	0.431	0.463	0.055	0.054
1–6	Scottish	0.237	0.227	0.878	0.975
	Indian	0.025	0.034	0.559	0.523
	Irish	0.009	0.004	1.000	1.000
	lj	0.063	0.078	0.287	0.223
	American	0.004	0.008	0.236	0.254
	English	0.478	0.750	0.109	0.188
2–4	Scottish	0.327	0.336	0.420	0.539
	Indian	0.426	0.365	0.213	0.169
	Irish	0.529	0.624	1.000	1.000
	lj	0.014	0.023	0.508	0.520
	American	0.259	0.375	0.679	0.750
	English	0.193	0.219	1.000	1.000
2–6	Scottish	0.447	0.516	0.708	0.844
	Indian	0.856	0.917	0.609	0.667
	Irish	0.169	0.250	0.824	1.000
	lj	0.015	0.023	0.664	0.668
	American	0.002	0.008	0.499	0.550
	English	0.074	0.110	0.884	0.902

Figure 3: DPO training loss

D Additional Experimental Results

D.1 DPO optimzation

Figure 3 illustrates the average training loss across several model variants using DPO. Most models demonstrate a smooth and consistent decrease in loss, indicating stable convergence behavior. While some variance exists across configurations, there are no signs of divergence or abrupt fluctuations. Overall, these results suggest that training with DPO is stable under the tested conditions.