Preference Optimization for Non-Verbal Vocalization Synthesis

Haoyang Li, Chenglin Xu, Junchuan Zhao, Yuang Cao, Liumeng Xue, Yiwen Guo, Eng Siong Chng

***Abstract*—Non-verbal vocalizations (NVs), such as laughter, coughs, and sighs, are essential for expressive TTS, but the effectiveness of preference optimization for NV generation remains poorly understood. We systematically study preference optimization for NV-capable TTS, focusing on preference signals, preference-pair construction, and DPO-based optimization objectives. We formulate an NV-aware character error rate (NV-CER) by treating NV tags as distinct output symbols and computing a weighted pinyin-based CER over both verbal and non-verbal content, enabling controllable optimization of NV realization without modifying the underlying optimization algorithm. Experiments on Emilia-NV and the augmented NV-Bench covering 18 NV types reveal how different design choices affect NV realization and lexical fidelity, and establish an effective setup using standard DPO. Objective, LLM-based, and human evaluations provide converging evidence for our findings, offering practical insights into NV-aware post-training for expressive TTS.**

***Index Terms*—speech synthesis, text-to-speech, nonverbal, paralinguistic, Direct Preference Optimization**

I. Introduction

Recent advances in neural text-to-speech (TTS) have enabled increasingly natural and expressive speech synthesis [1], [2], [3]. Beyond verbal content, natural communication also relies on non-verbal vocalizations (NVs), including laughter, crying, and hesitation sounds, which convey emotion, discourse structure, and speaker intent [4]. However, faithful NV generation remains challenging and relatively underexplored in TTS. Recent work has advanced NV-aware TTS through dataset construction [5], [6], [7], [8], [9] and benchmark development for standardized evaluation [10], [11]. These resources enable systematic investigation of NV realization in TTS, an area that has received limited attention to date.

Preference optimization [12], [13] has emerged as an effective post-training paradigm for TTS, improving aspects of naturalness and intelligibility [14], [15], [16]. Recent TTS systems, including Fish Audio S2 [17], CosyVoice2 [18], and CosyVoice3 [19], have incorporated NV-aware models into preference optimization pipelines. However, these studies primarily focus on overall synthesis quality, providing limited insight into the specific impact of post-training on NV generation. Fish Audio S2 reports final system performance without isolating the contribution of preference optimization, while CosyVoice2 and CosyVoice3 lack NV evaluation. [9] applied DPO to NV-TTS using preference pairs constructed from manually verified speech with NVs and corresponding speech without NVs, encouraging NV occurrence rather than faithful NV realization, and reported degraded CER performance. Consequently, both the effectiveness of preference optimization for NV generation and the design choices that govern its performance remain unclear, motivating a systematic study of preference signals, preference-pair construction, and optimization objectives for NV-aware TTS.

In this work, we systematically investigate NV-aware preference optimization for LLM-based TTS. We formulate an NV-aware character error rate (NV-CER) by treating non-verbal vocalization tags as distinct output symbols and computing a weighted pinyin-based CER over both verbal and non-verbal content. Building on this metric, we construct preference signals using an NV-capable automatic speech recognition (ASR) model and introduce a simple weighting mechanism to adjust the relative importance of NV tags, enabling explicit control over the trade-off between NV realization and lexical fidelity without modifying the underlying optimization algorithm. We then systematically study preference signals, preference-pair construction strategies, and loss formulations within DPO. Our experiments establish how these design choices affect NV realization, leading to an effective setup using standard DPO. Objective, LLM-based, and human evaluations provide converging evidence, establishing practical principles for effective preference optimization and post-training of NV-aware TTS.

II. Methodology

A. NV-Aware Preference Signal

1) NV-Aware Speech Recognition: Conventional automatic speech recognition (ASR) systems transcribe only spoken text and disregard non-verbal vocalizations (NVs), making them unsuitable for evaluating expressive speech synthesis. A straightforward alternative is to compare the recognized text against the reference transcription. However, this introduces a fundamental limitation for NV evaluation. Many non-verbal vocalizations, such as *laughter*, *crying*, and *coughing*, do not have deterministic textual representations, making it impossible to establish a one-to-one mapping between speech and text. Although some vocalizations (e.g., hesitation or surprise sounds) may be approximated by lexical words or pinyin, these mappings are inherently ambiguous and context-dependent.

To overcome this limitation, we leverage an NV-aware ASR (NV-ASR) model that explicitly recognizes predefined NV

Haoyang Li, Eng Siong Chng are with Nanyang Technological University, Singapore (e-mail: LI0078NG@e.ntu.edu.sg; ASESChng@ntu.edu.sg;). Chenglin Xu is with LIGHTSPEED, Singapore (e-mail: cxu011@e.ntu.edu.sg;). Junchuan Zhao is with National University of Singapore, Singapore (e-mail: junchuan@u.nus.edu;). Yuang Cao, Liumeng Xue are with Nanjing University, China (e-mail: cya696969@gmail.com; lmxue@nju.edu.cn;). Yiwen Guo is an independent researcher (e-mail: guoyiwen89@gmail.com;). Correspondence to Chenglin Xu. This work was done during Haoyang Li's research internship at LIGHTSPEED.

events alongside spoken text. Given a synthesized speech sample x, the NV-ASR model predicts

$$\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N\}, \tag{1}$$

where each output token corresponds to either a lexical token or an NV tag. Unlike conventional ASR, the resulting transcription preserves explicit non-verbal vocalization information, providing the basis for constructing the proposed NV-aware preference signal.

2) NV-Aware Character Error Rate: Given the reference and predicted transcriptions, we define an **NV-aware Character Error Rate (NV-CER)** to measure both lexical and NV correctness. For Mandarin, lexical tokens are represented using pinyin to reduce sensitivity to homophones and character variations, while NV tags are retained as independent symbols. Edit distance is computed over the combined token sequence,

$$\text{NV-CER} = \frac{D_c(\mathbf{r}, \hat{\mathbf{y}})}{\sum_{t \in \mathbf{r}} c(t)}, \tag{2}$$

where $\mathbf{r}$ and $\hat{\mathbf{y}}$ denote the reference and predicted token sequences, respectively, and $D_c(\cdot, \cdot)$ denotes the weighted edit distance with token-dependent insertion and deletion costs given by $c(t)$ and substitution cost given by the maximum cost of the two tokens. Although instantiated using pinyin for Mandarin, the proposed framework is readily extendable to other languages by replacing the lexical representation with an appropriate language-specific alternative while preserving NV tags as independent symbols.

3) Controllable NV Preference Strength: Different applications may require different trade-offs between NV realization and lexical accuracy. To enable this flexibility, we assign a higher edit cost to NV tags, with w_{NV} denoting the NV weight:

$$c(t) = \begin{cases} w_{\text{NV}}, & \text{if } t \text{ is an NV tag,} \\ 1, & \text{otherwise,} \end{cases} \tag{3}$$

Increasing w_{NV} increases the contribution of NV errors to NV-CER, giving them greater influence when constructing preference signals. Conversely, $w_{\text{NV}} = 1$ assigns equal cost to lexical and NV tokens. This weighting mechanism provides explicit control over the relative importance of NV and lexical accuracy without modifying the underlying preference optimization algorithm.

B. NV-aware Preference Alignment

1) NV-Aware Preference Pair Construction: We adopt Direct Preference Optimization (DPO) as the preference optimization framework due to its simplicity, effectiveness, and widespread adoption in TTS post-training. Given an input text prompt p, the pretrained TTS model independently generates K candidate speech utterances,

$$\mathcal{X} = \{x_1, x_2, \ldots, x_K\}. \tag{4}$$

Each candidate is transcribed using the NV-ASR model described in Section II-A, and its NV-CER score is computed against the reference transcription. The candidate with the lowest NV-CER is selected as the preferred response x^+, while the candidate with the highest NV-CER is selected as the rejected response x^-, forming the preference pair used for optimization. In addition, we investigate several alternative preference construction strategies, including utilizing ground-truth speech during preference selection and manipulating the synthetic speech. These variants are described in Section III-D and evaluated experimentally.

2) Preference Optimization Objective: Given the constructed preference pairs (x^+, x^-), the TTS model is optimized to directly increase the relative likelihood of the preferred response over the rejected response via DPO:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\beta \left[\log \frac{\pi_\theta(x^+|p)}{\pi_{\text{ref}}(x^+|p)} - \log \frac{\pi_\theta(x^-|p)}{\pi_{\text{ref}}(x^-|p)} \right] \right), \tag{5}$$

where π_θ and π_{ref} denote the trainable and reference TTS models, respectively, and β controls the preference strength.

Besides the standard DPO objective, we investigate several design choices for NV-aware preference optimization, including combining DPO with the conventional supervised fine-tuning (SFT) objective and alternative preference ranking in Section III-D, with their effectiveness evaluated in Section IV.

III. Experiments

A. Dataset

We train on Emilia-NV [8], comprising 573.4 hours of expressive Mandarin speech with transcriptions annotated by 18 predefined NV tags, covering laughter, breathing, hesitation, and other vocal events.

We evaluate on the Mandarin subset of NV-Bench [10], which provides text prompts, speaker-conditioning recordings, and ground-truth speech. It contains a single-label subset with 50 utterances per NV category and a 391-utterance multi-label subset, each containing two or more NV events. As NV-Bench covers only 15 of the 18 training tags, we further construct 50 prompts for each of the three missing categories using `gemini-3.1-pro-preview`, following the NV-Bench annotation style and using its speaker-conditioning utterances. This adds 150 samples to the single-label subset.

B. Evaluation Metrics

Evaluation is based on recent NV benchmarks [10], [11].

1) Objective metrics: **Speech intelligibility.** NV-CER (Eq. 2) evaluates overall transcription accuracy by jointly measuring lexical content and NVs, with the NV weight w_{NV} set to 1. To separately assess verbal and non-verbal generation, we additionally report CER, computed after removing all NV tags, and PCER, computed after excluding all lexical tokens.

Perceptual quality. DNSMOS P.835 [20] predicts speech (SIG), background (BAK), and overall (OVRL) quality, while UTMOS predicts overall MOS for synthesized speech.

Speaker similarity. Speaker similarity is evaluated using speaker embedding cosine similarity (SECS), computed from speaker embeddings extracted with Resemblyzer[1].

[1]https://github.com/resemble-ai/Resemblyzer

TABLE I: Ablation on loss and preference pair construction. GT: ground-truth speech; Syn^{+}/Syn^{-}: best-/worst-scoring synthetic candidates; Hybrid: GT or Syn^{+} selected by NV-CER; Syn^{NoNV}: synthetic speech without NVs. Bold/underline: best/second-best.

Loss	Pair (Pref, Rej)	NV-CER	PCER	CER
DPO	GT, Syn^{-}	93.42	104.89	93.20
DPO+SFT	GT, Syn^{-}	9.43	39.33	8.66
DPO	Hybrid, Syn^{-}	65.18	75.22	64.99
DPO+SFT	Hybrid, Syn^{-}	7.64	33.22	7.00
DPO	Syn^{+}, Syn^{NoNV}	22.95	52.89	22.23
DPO+SFT	Syn^{+},Syn^{NoNV}	3.68	25.56	3.12
DPO	Syn^{+}, Syn^{-}	**2.59**	**21.33**	**2.09**
DPO+SFT	Syn^{+}, Syn^{-}	<u>3.03</u>	<u>23.00</u>	<u>2.52</u>

TABLE II: Effect of training epochs (pref pair: Syn^{+}, Syn^{-}).

Loss	Epoch	NV-CER	PCER	CER
DPO	1	**2.59**	**21.33**	**2.09**
	2	31.32	56.89	30.83
DPO+SFT	1	<u>3.03</u>	23.00	<u>2.52</u>
	2	3.25	<u>21.78</u>	2.78
	3	3.29	21.89	2.83

2) LLM-based multi-rater: We employ Gemini-2.5-Pro as an LLM-based evaluator to assess NV accuracy (**A**), NV perceptual effect (**P**), overall naturalness (**N**), overall quality (**Q**), and overall expression (**E**), following the Track 2 evaluation rubrics of the NVVSpeech Challenge[2]. All systems for the same utterance are jointly presented with anonymized labels and randomized presentation orders to mitigate system-identity and positional biases in comparative judging. Three independent simulated raters evaluate each sample, with scores assigned according to predefined 1–5 rubrics (0–5 for NV-specific criteria when no NV is audible).

3) Subjective Evaluation: We further conduct an A/B/Tie preference test between the SFT baseline and preference model on the multi-label evaluation set of NV-Bench, providing a human reference for comparison with the objective and LLM-based evaluations. Eleven native-speaking volunteers evaluate four criteria: NV accuracy (NV Acc.); NV naturalness (NV Nat.), with an N/A option when NVs are not comparable (e.g., when an NV is missing); lexical accuracy (Lex. Acc.), assessing text correctness excluding NVs; and overall naturalness (Overall Nat.). The workload is distributed such that each sample is independently evaluated by five volunteers. All participants provided informed consent prior to the evaluation.

C. Implementation Details

We first perform supervised fine-tuning (SFT) on the pre-trained CosyVoice2-0.5B model using the Emilia-NV dataset [8] for 4 epochs with a learning rate of 1×10^{-5}, resulting in a *Base* model that serves as the common initialization for all subsequent post-training experiments. For preference optimization, we construct an approximately 100-hour subset of Emilia-NV by prioritizing utterances containing underrepresented NV classes to obtain a more balanced class distribution.

[2]https://nvvspeech-challenge.github.io/

TABLE III: Effect of w_{NV} on NV-CER preference signals.

w_{NV}	NV-CER	PCER	CER	DNSMOS	Sim
1	**2.59**	21.33	**2.09**	**3.346/3.607/4.091**	**0.891**
5	3.02	<u>18.22</u>	2.60	3.337/3.601/4.086	0.890
10	<u>2.91</u>	**18.11**	<u>2.52</u>	<u>3.339</u>/3.601/<u>4.089</u>	0.890

Using the *Base* model, we independently generate 8 candidate utterances for each training sample in the post-training subset, from which preference pairs are constructed. Candidate generation follows the default Repetition Aware Sampling (RAS) strategy in CosyVoice2, with $top_p = 0.9$, $top_k = 30$, $win\_size = 10$, and $\tau_r = 0.1$ to encourage sampling diversity. During evaluation, top_p and top_k are reduced to 0.8 and 25, respectively, following the default CosyVoice2 configuration.

Unless otherwise specified, all post-training experiments use a learning rate of 1×10^{-6}, dynamic batching with a maximum of 1000 frames per batch, gradient accumulation of 40.

We adopt the NV-ASR model from [10] to compute NV-CER. It fine-tunes SenseVoice-Small [21] on multiple public NV speech corpora [8], [5], [22], [6], [7], [23], and recognizes all 18 NV tags in our training set.

D. Design Choices

Loss formulation: We investigate whether combining DPO with the supervised fine-tuning (SFT) objective improves performance and training stability. The SFT loss is weighted by 0.2 to keep its magnitude comparable to the DPO loss.

Preference pair construction: We investigate how preference-pair construction affects optimization. For preferred responses, we compare three strategies: (1) the default, selecting the top-ranked synthetic candidate (Section II-B1); (2) always selecting ground-truth speech; and (3) selecting ground-truth speech when its NV-CER is lower than the top-ranked synthetic candidate, otherwise using the synthetic candidate. For rejected responses, we additionally consider synthesized speech without NVs, following [9].

Preference signal: Unless otherwise stated, preference scores are computed using NV-CER (Eq. 2) with $w_{\mathrm{NV}} = 1$. We investigate three preference-signal alternatives: (1) conventional pinyin CER from the general-purpose SenseVoice-Small ASR model [21]; (2) a composite score combining NV recognition and perceptual quality,

$$s = \mathrm{CER}_{\mathrm{NV}} - \lambda \left(\frac{\mathrm{UTMOS} - 1}{4} \right), \tag{6}$$

where $\lambda = 0.4$; and (3) varying w_{NV} in Eq 3 to explicitly control the relative importance of NV errors.

IV. Results and Discussion

A. Loss and Preference Pair Ablation

We investigate loss formulation and preference-pair construction on the single tag testset of NV-Bench. As shown in

TABLE IV: Comparison of loss and preference-signal variants on the single-tag test set. Preference pairs use Syn^{+}/Syn^{-}.

Loss	Preference Signal	NV-CER	PCER	CER	UTMOS	DNSMOS	Sim
Base	–	3.47	32.78	2.74	2.01	3.356/3.616/4.099	0.890
SFT	–	3.62	30.78	2.93	2.00	3.353/3.613/4.098	0.890
DPO	ASR CER	2.95	25.78	2.36	2.03	3.350/3.609/4.095	0.891
DPO	NV-CER	2.59	**21.33**	2.09	2.01	3.346/3.607/4.091	0.891
DPO	NV-CER + UTMOS	**2.58**	22.44	2.09	**2.11**	**3.377/3.629**/4.110	0.891
DPO+SFT	ASR CER	3.35	26.00	2.78	2.01	3.353/3.612/4.099	0.889
DPO+SFT	NV-CER	3.03	23.00	2.52	2.01	3.348/3.608/4.097	0.890
DPO+SFT	NV-CER + UTMOS	3.07	24.89	2.56	2.04	3.366/3.620/**4.111**	0.889

TABLE V: Comparison of loss and preference-signal variants on the multi-tag test set. Preference pairs use Syn^{+}/Syn^{-}.

Loss	Preference Signal	NV-CER	PCER	CER	UTMOS	DNSMOS	Sim	A	P	N	Q	E
Base	–	4.98	39.28	3.69	1.92	3.355/3.617/4.099	0.889	3.28	2.34	2.43	3.28	2.28
SFT	–	5.11	35.58	3.94	1.90	3.341/3.609/4.086	0.890	3.35	2.31	2.37	3.23	2.22
DPO	ASR CER	4.94	37.71	3.64	1.95	3.342/3.605/4.089	0.891	3.45	2.73	2.85	3.59	2.67
DPO	NV-CER	4.29	28.40	3.52	1.93	3.346/3.610/4.091	0.891	3.70	2.85	2.92	3.61	2.75
DPO	NV-CER ($w_{\mathrm{NV}} = 10$)	4.57	**24.58**	3.91	1.94	3.352/3.613/4.096	0.890	**3.73**	**2.91**	2.95	3.62	**2.80**
DPO	NV-CER + UTMOS	**4.25**	31.31	**3.27**	**2.05**	**3.392/3.643/4.119**	0.891	3.61	2.90	**2.97**	**3.72**	2.79

TABLE VI: Human Preference on the multi-tag test set.

Criterion	Preference	Tie	SFT	N/A
NV Acc.	340 (17.39%)	1342 (68.64%)	273 (13.96%)	–
NV Nat.	594 (30.38%)	669 (34.22%)	584 (29.87%)	108 (5.52%)
Lexical Acc.	186 (9.51%)	1630 (83.38%)	139 (7.11%)	–
Overall Nat.	674 (34.48%)	598 (30.59%)	683 (34.94%)	–

Table I, the Syn^{+}/Syn^{-} pair achieves the best performance and is adopted in subsequent experiments. Using GT/Hybrid in the preferred response reduces performance, while Syn^{NoNV} also underperforms Syn^{-} because the base model rarely omits NVs, making NV-free speech weak negative examples.

Table I and II show that DPO alone is unstable, with performance deteriorating after the first or second epoch. Adding SFT substantially improves training stability but weakens the DPO effect for Syn^{+}/Syn^{-}. We therefore use DPO in subsequent experiments and include DPO+SFT in selected experiments as a robustness check.

B. Investigation on the impact of NV-Weight tuning

Table III evaluates the effect of the NV tag weight on the single tag testset. Increasing w_{NV} improves PCER, with a modest increase in CER. Across all settings, CER remains below that of the Base and SFT models in Table IV. This result suggests that weighted NV-CER can modulate the relative emphasis on NV-related and lexical errors without modifying the underlying preference optimization framework.

C. Comparison with Baselines

Table IV compares NV-aware preference signals with conventional ASR-CER preference signal, the pretrained (*Base*), and SFT models on the single-tag test set. SFT slightly degrades performance from Base, whereas NV-CER consistently improves both lexical and NV-related metrics and outperforms conventional ASR-CER. Combining NV-CER with UTMOS achieves similar transcription accuracy while slightly improves UTMOS and DNSMOS. The same trends hold for both DPO and DPO+SFT, demonstrating robustness across loss formulations. Speaker similarity remains largely unchanged.

Table V further evaluates the methods on the multi-tag test set using objective and LLM-based metrics. NV-CER-based preference signals consistently outperform the baselines.

D. Human Evaluation

Table VI compares the preference model (NV-CER+UTMOS, Table V) with the SFT baseline via subjective listening tests. High tie rates for NV Acc. and Lex. Acc. are consistent with the strong baseline performance reflected by the CER-based metrics. Among non-tied judgments, the preference model is favored 55.5% vs. 44.5% for NV Acc. and 57.2% vs. 42.8% for Lex. Acc., consistent with the objective and LLM evaluations. No significant preference is observed for NV Nat. or Overall Nat., suggesting that the accuracy gains do not compromise perceived naturalness. This aligns with the preference signal, which targets NV and lexical correctness through NV-CER, while UTMOS does not explicitly assess NV naturalness. The discrepancy with LLM evaluation further highlights the need for better metrics to distinguish NV naturalness when perceived differences are small. Overall, NV-aware preference signals improve targeted accuracy while preserving perceived naturalness.

V. Conclusion

We presented an NV-aware preference optimization framework for expressive TTS, leveraging an NV-capable ASR model to construct preference signals over lexical and NV content. Weighted NV-CER enables control over their relative importance without modifying the optimization algorithm. Systematic analysis of preference signals, preference-pair construction, and loss formulations identifies effective DPO configurations and reveals how these choices affect target objectives. Objective, LLM-based, and human evaluations consistently support improvements over baselines. Together, these results provide a practical foundation and guidance for effective NV-aware post-training.

REFERENCES

[1] K. Shen, Z. Ju, X. Tan, E. Liu, Y. Leng, L. He, T. Qin, J. Bian *et al.*, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *International conference on learning representations*, vol. 2024, 2024, pp. 698–722.

[2] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, and X. Chen, "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.

[3] H. Hu, X. Zhu, T. He, D. Guo, B. Zhang, X. Wang, Z. Guo, Z. Jiang, H. Hao, Z. Guo *et al.*, "Qwen3-tts technical report," *arXiv preprint arXiv:2601.15621*, 2026.

[4] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[5] M. Borisov, E. Spirin, and D. Diatlova, "Nonverbaltts: A public english corpus of text-aligned nonverbal vocalizations with emotion annotations for text-to-speech," *arXiv preprint arXiv:2507.13155*, 2025.

[6] R. Ye, Y. Zhou, R. Yu, Z. Lin, K. Li, X. Li, X. Liu, G. Zeng, and Z. Wu, "A scalable pipeline for enabling non-verbal speech generation and understanding," *arXiv preprint arXiv:2508.05385*, 2025.

[7] Z. Wu, D. Liu, J. Liu, Y. Wang, L. Li, L. Jin, H. Bu, P. Zhang, and M. Li, "Smiip-nv: A multi-annotation non-verbal expressive speech corpus in mandarin for llm-based speech synthesis," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 12 564–12 570.

[8] H. Liao, Q. Ni, Y. Wang, Y. Lu, H. Zhan, P. Xie, Q. Zhang, and Z. Wu, "Emilia-nv: A non-verbal speech dataset with word-level annotation for human-like speech modeling," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 17 587–17 591.

[9] B. Bai, Q. Lu, W. Yang, Z. Sun, Y. Hou, P. Jia, S. Pu, R. Fu, Y. Gao, Y. Li *et al.*, "Synparaspeech: Automated synthesis of paralinguistic datasets for speech generation and understanding," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 15 527–15 531.

[10] Q. Ni, H. Liao, D. Chen, Y. Wang, and Z. Wu, "Nv-bench: Benchmark of nonverbal vocalization synthesis for expressive text-to-speech generation," *arXiv preprint arXiv:2603.15352*, 2026.

[11] L. Xue, W. Bian, J. Pan, W. Wang, Y. Ren, B. Kang, J. Hu, Z. Ma, S. Wang, X. Qian *et al.*, "Nvbench: A benchmark for speech synthesis with non-verbal vocalizations," *arXiv preprint arXiv:2604.16211*, 2026.

[12] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in neural information processing systems*, vol. 36, pp. 53 728–53 741, 2023.

[13] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[14] X. Gao, C. Zhang, Y. Chen, H. Zhang, and N. F. Chen, "Emo-dpo: Controllable emotional speech synthesis through direct preference optimization," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[15] J. Tian, C. Zhang, J. Shi, H. Zhang, J. Yu, S. Watanabe, and D. Yu, "Preference alignment improves language model-based tts," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[16] Y. H. Yeo, H. Li, Y. Peng, S. Gopal, H. Liu, L. P. Garcia-Perera, H. B. Sailor, J. H. Wong, and E. S. Chng, "Improving code-switching asr with code-mixing guided synthetic speech," *arXiv preprint arXiv:2606.19381*, 2026.

[17] S. Liao, Y. Wang, S. Liu, Y. Cheng, R. Zhang, T. Li, S. Li, Y. Zheng, X. Liu, Q. Wang *et al.*, "Fish audio s2 technical report," *arXiv preprint arXiv:2603.08823*, 2026.

[18] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[19] Z. Du, C. Gao, Y. Wang, F. Yu, T. Zhao, H. Wang, X. Lv, H. Wang, C. Ni, X. Shi *et al.*, "Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training," *arXiv preprint arXiv:2505.17589*, 2025.

[20] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 886–890.

[21] K. An, Q. Chen, C. Deng, Z. Du, C. Gao, Z. Gao, Y. Gu, T. He, H. Hu, K. Hu *et al.*, "Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms," *arXiv preprint arXiv:2407.04051*, 2024.

[22] K. Wang and D. Herremans, "Disfluencyspeech–single-speaker conversational speech dataset with paralanguage," in *TENCON 2024-2024 IEEE Region 10 Conference (TENCON)*. IEEE, 2024, pp. 469–472.

[23] J. Mai, J. Ji, X. Xing, C. Yang, W. Chen, J. Xing, and X. Xu, "Mnv-17: A high-quality performative mandarin dataset for nonverbal vocalization recognition in speech," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 18 312–18 316.