Best-of-N TTS Evaluation is Confounded by ASR Family Alignment

Taehyung Yu [1] **Seongjae Kang** [1]

Abstract

Best-of-N (BoN) inference improves content consistency in zero-shot text-to-speech by selecting from N candidates with an automatic speech recognition (ASR) verifier. We identify an underexplored evaluation confound: a verifier's apparent quality depends strongly on which ASR family judges it. On LibriSpeech-PC test-clean (Meister et al., 2023) with F5-TTS (Chen et al., 2025), verifier rankings reverse across Whisper, wav2vec 2.0, and HuBERT evaluators, and same-family verifier–evaluator pairs recover 2–3$\times$ more oracle headroom than cross-family pairs despite near-identical representations (linear CKA 0.978)—a pattern consistent with identity- or lineage-level coupling rather than representational overlap. We propose two **cross-family rank ensembles** (rank-averaging and conjunctive max-rank) that attain the lowest mean WER across three independent evaluators—1.61% at $N{=}10$ (-12% relative to F5-TTS)—with no measurable degradation under automatic SIM-o/UTMOS metrics; the best single verifier drives WER from 2.06% to 1.72% (-16.5%) under the official F5-TTS evaluator. We recommend cross-evaluator triangulation as default reporting practice.

1. Introduction

Recent flow-matching zero-shot TTS systems—F5-TTS (Chen et al., 2025), E2 TTS (Eskimez et al., 2024), CosyVoice 2 (Du et al., 2024), MaskGCT (Wang et al., 2025), Seed-TTS (Anastassiou et al., 2024), NaturalSpeech 3 (Ju et al., 2024)—produce speech that is indistinguishable from human recordings in naturalness and speaker similarity, yet still produce word-level content errors on a non-trivial fraction of utterances. Best-of-N (BoN) inference is the standard inference-time remedy: synthesize N candidates with different random initializations and select the one whose ASR-decoded transcript matches the target text most closely.

BoN is reported to reduce WER by 10–30$\%$ relative across recent flow-matching systems, but the literature is inconsistent on which *verifier* ASR to use—reported choices include wav2vec 2.0 (Baevski et al., 2020), Whisper (Radford et al., 2023), and its distillations (Gandhi et al., 2023)—and the evaluation ASR is usually a single fixed model. The choice of evaluator is a design decision that has so far received no systematic attention.

In this paper we run a four-way evaluator ablation spanning the Whisper (Radford et al., 2023), wav2vec 2.0 (Baevski et al., 2020), and HuBERT (Hsu et al., 2021) families (§3) over a shared set of BoN candidates. Our contributions are the following.

- We document that BoN verifier rankings are systematically evaluator-dependent: the same generated outputs are ranked in opposing directions by evaluators from different ASR families, so the verifier preferred under one family can lose under another (§4).
- We quantify the magnitude of this confound on LibriSpeech-PC test-clean and show that the (verifier, evaluator) family pairing is a much larger lever on reported WER than the choice between common verifier checkpoints, while a large oracle headroom remains unexploited (§4).
- We rule out audio-encoder representation similarity (linear CKA) as the dominant explanation; the pattern is instead consistent with identity- or lineage-level coupling between verifier and evaluator, a speech analog of LLM-as-a-judge self-bias (§5).
- We propose *cross-family rank ensembles* that select candidates by aggregating verifiers across ASR families, and recommend *cross-evaluator triangulation*—reporting WER under at least two ASR families with disjoint training lineages—as a default reporting practice (§4.4).

2. Related Work

Inference-time remedies for flow-matching TTS—classifier-free guidance (CFG) (Ho & Salimans, 2021), BoN reranking (Du et al., 2024; Yang et al., 2025), reverse inference (Hu

[1]KAIST, Daejeon, South Korea. Correspondence to: Taehyung Yu <taehyung.yu@kaist.ac.kr>.

Accepted at ICML 2026 Workshop on Machine Learning for Audio, Seoul, South Korea.

et al., 2024)—report no cross-evaluator validation; recent strong systems (Anastassiou et al., 2024; Ju et al., 2024) similarly tune verifiers against a single fixed ASR. Speech preference optimization (Zhang et al., 2024) shares the same single-judge risk. BoN scaling theory in LMs (Yu et al., 2025; Wang et al., 2024) motivates verifier ensembling but has no TTS analog. The structural precedent is LLM-as-judge self-bias (Zheng et al., 2023); we document the speech analog. Evaluator-driven shifts in spoken-language conclusions are established for ASR fairness (Koenecke et al., 2020; Feng et al., 2024) and for MOS-predictor cross-domain generalization (Cooper et al., 2022; Mittag et al., 2021; Baba et al., 2024).

3. Setup

Synthesis. We use the publicly released F5-TTS base model (Chen et al., 2025) with its default inference recipe: 32 ODE solver steps, CFG (Ho & Salimans, 2021) scale 2.0, and sway sampling. We generate 10 BoN candidates per utterance; for each $N \in \{3, 5, 10\}$ we take the leading N candidates and select the one with the lowest joint WER+CER score against the reference transcript.

Verifiers. We compare three single-model verifiers and one three-way ensemble. The single verifiers are a wav2vec 2.0 base model (95M parameters; *w2v2-base*), a small Distil-Whisper (166M; *distil-sm*), and a large Distil-Whisper (756M; *distil-v3*). *ens3* denotes the three-way rank-average of these three verifiers. In the scaling study (§4.4) we additionally consider two cross-family rank ensembles that aggregate only the cross-family pair {w2v2-base, distil-v3}.

Evaluators. For the main results (§4) we score every candidate with four ASR evaluators spanning three families: a Whisper-large-v3 evaluator served via the faster-whisper runtime (*fwhisper-lgv3*, the official F5-TTS evaluator (Chen et al., 2025)); a large wav2vec 2.0 model (*w2v2-lv60*); the same Distil-Whisper used as a verifier (*distil-sm*); and a large HuBERT model (*hubert-lg*). For the CKA–WER analysis (§5) we additionally include Whisper-medium (*whisper-med*) as a fourth evaluator to expand the pool of evaluator pairs.

Data & normalization. We evaluate on the LibriSpeech-PC test-clean cross-sentence subset (Meister et al., 2023) (1127 samples, 4–10 s reference) and run the F5-TTS official evaluation pipeline verbatim, applying its default text normalization uniformly to all four ASR outputs. Whisper's English-specific normalizer is not applied separately, to keep the cross-ASR WER comparison on the same footing.

Quality. Speaker similarity SIM-o is the cosine of WavLM (Chen et al., 2022) speaker embeddings; reference and synthesized audio are resampled symmetrically from 24 kHz to 16 kHz before scoring. Naturalness is the UTMOS (Saeki et al., 2022) score from the standard 22-strong checkpoint. Real-time factors are measured isolated on a single RTX 4090 GPU.

Table 1. LibriSpeech-PC test-clean under the official fwhisper-lgv3 evaluator. "rel" is WER reduction vs. F5-TTS; "p" is a two-sided paired permutation test against F5-TTS (10,000 permutations).

Method	WER	CER	RTF	rel	p
F5-TTS (baseline)	2.06	0.71	0.190	—	—
BoN w2v2-base	1.99	0.62	0.383	−3.5%	0.31
BoN distil-sm	2.04	0.59	0.388	−1.0%	0.93
BoN distil-v3	**1.88**	**0.53**	0.379	**−8.7%**	**0.030**
BoN ens3	1.91	0.54	0.429	−7.2%	0.057

Table 2. Cross-evaluator WER (%) on LibriSpeech-PC test-clean. The same generated audio is evaluated by ASR checkpoints from different families (aliases defined in §3); the preferred verifier (bold) changes with the evaluator.

Method	fwhisper-lgv3	w2v2-lv60	hubert-lg
F5-TTS (baseline)	2.06	1.52	1.92
BoN w2v2-base	1.99	**1.41**	**1.74**
BoN distil-sm	2.04	1.50	1.77
BoN distil-v3	**1.88**	1.45	**1.74**
BoN ens3	1.91	1.45	1.76

Reproducibility. All models are used at their default public revisions; code, decoding settings, and evaluation scripts are available at `https://github.com/yu1012/BoN-TTS`.

4. Main Results

4.1. Results under the F5-TTS Evaluator

Our single-shot F5-TTS baseline reaches 2.06% WER, within 0.36 pp of the 2.42% multi-seed mean reported in the F5-TTS paper (Chen et al., 2025)—a single-shot vs. multi-seed-mean gap, not a pipeline discrepancy. BoN with distil-v3 drives WER to 1.88% (−8.7%, p=0.030) at RTF 0.379, the only significant single-verifier reduction; ens3 is a close runner-up (−7.2%, p=0.057). The w2v2-base verifier improves the point estimate (−3.5%) but is not significant (p=0.31), and distil-sm is flat (p=0.93). Speaker similarity (SIM-o) stays within ±0.0006 and UTMOS within ±0.005 of F5-TTS across all configurations (§5.1), so the WER gain comes at no measurable quality cost.

4.2. Cross-Evaluator Triangulation

Same-family preference. Under fwhisper-lgv3 (Table 2), the Distil-Whisper verifier (distil-v3) wins; under w2v2-lv60, the wav2vec 2.0 verifier (w2v2-base) wins; under hubert-lg, w2v2-base again ranks at the top alongside distil-v3.

Table 3. Oracle headroom on LibriSpeech-PC test-clean with N=3 candidates per sample. *Single-shot* = first candidate; *Oracle* = per-sample best of 3. The bottom block reports recovery (Single−BoN)/(Single−Oracle) in %; same-family BoN recovers 2–3× more than cross-family BoN on the same audio. The 0.02 pp gap between Single-shot = 2.04 here and F5-TTS = 2.06 in Tables 1/4 is Whisper batching non-determinism (decoded in an N=3 batch vs. a single utterance); both rows describe identical synthesized audio.

	fwhisper-lgv3	w2v2-lv60	hubert-lg
Single-shot	2.04	1.52	1.94
Oracle (N=3)	**1.42**	**1.09**	**1.18**
Headroom (pp)	0.63	0.43	0.76
Recovery (%) of oracle headroom:			
BoN w2v2-base	7.9	**26.1**	27.1
BoN distil-sm	0.0	6.8	22.6
BoN distil-v3	**26.0**	18.2	27.1
BoN ens3	20.5	17.0	24.5

Magnitude of the swing. The distil-v3 BoN reduction is 0.18 pp under fwhisper-lgv3 but only 0.07 pp under w2v2-lv60, reversing the verifier ranking between Whisper and wav2vec 2.0 evaluators. A reader given only the F5-TTS official evaluator would prefer distil-v3 BoN; one given only w2v2-lv60 would prefer w2v2-base BoN. Both are correct conditional on their evaluator, and incompatible.

4.3. Oracle Headroom and Recovery

To bound how far BoN can go at N=3, we synthesize all three candidates per sample and transcribe each with every evaluator (Table 3, top block). On the official evaluator, an oracle verifier would drive WER to 1.42%, below F5-TTS's reported ∼1.5–1.7% (Chen et al., 2025): headroom for verifier improvement is substantial. The bottom block makes the 2–3× family swing precise: under fwhisper-lgv3, distil-v3 recovers **26.0%** of headroom vs. only 7.9% for cross-family w2v2-base (3.3×); under w2v2-lv60, w2v2-base recovers **26.1%** vs. 18.2% for cross-family distil-v3 (1.4×). Recovery is a property of the (verifier, evaluator) pair, not the verifier alone.

4.4. Scaling and Bias-Corrected Ensembles

To test whether the confound persists at higher N and whether cross-family aggregation mitigates it, we synthesize N=10 candidates per sample and add six BoN configurations: single wav2vec 2.0 and Distil-Whisper verifiers at $N \in \{5, 10\}$, plus two cross-family rank ensembles over the (w2v2-base, distil-v3) pair.

- **rank-avg.** Each verifier independently ranks the N candidates by WER+CER score. We then pick the candidate with the lowest average rank across the two families.

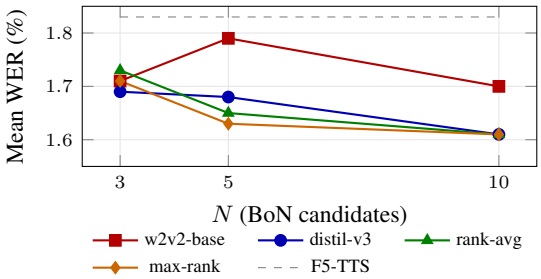

Figure 1. Mean WER (averaged over fwhisper-lgv3, w2v2-lv60, hubert-lg) vs. N, from Table 4. Both cross-family rank ensembles (rank-avg, max-rank) at N=5 reduce WER under all three evaluators simultaneously, while the single w2v2-base verifier regresses on fwhisper-lgv3 at N=5. Cross-family ensembles scale monotonically with N.

Table 4. Scaling N on LibriSpeech-PC test-clean. Per-evaluator WER (%) under three evaluators (aliases as defined in §3). The p-value column reports a two-sided paired permutation test against F5-TTS under *fwhisper-lgv3* (10,000 permutations). "↑" marks a regression vs. the same row at the preceding N.

Method (N)	fwhisper-lgv3	w2v2-lv60	hubert-lg	mean	p
F5-TTS (baseline)	2.06	1.52	1.92	1.83	—
w2v2-base (3)	1.99	1.41	1.74	1.71	0.05
w2v2-base (5)	2.20↑	1.41	1.75	1.79	0.97
w2v2-base (10)	1.98	1.40	1.72	1.70	0.08
distil-v3 (3)	1.88	1.45	1.74	1.69	<.001
distil-v3 (5)	1.80	1.48	1.75	1.68	<.001
distil-v3 (10)	**1.72**	1.44	1.68	**1.61**	<.001
rank-avg (3)	2.01	1.43	1.74	1.73	0.17
rank-avg (5)	1.90	**1.40**	1.66	1.65	**.001**
rank-avg (10)	1.81	1.41	**1.60**	1.61	<.001
max-rank (3)	1.99	1.42	1.73	1.71	0.12
max-rank (5)	1.80	1.43	1.67	1.63	<.001
max-rank (10)	1.80	**1.40**	1.62	1.61	<.001

- **max-rank.** We pick the candidate with the lowest worst-case rank, requiring it to rank well in both families. This conjunctive form explicitly penalizes single-family inflation.

Findings (Figure 1). (i) The single cross-family verifier (w2v2-base) under fwhisper-lgv3 regresses from 1.99% at N=3 to 2.20% at N=5 (p=0.97): more candidates expose a same-family-evaluator penalty. (ii) The same-family verifier (distil-v3) scales monotonically on its own family (1.88% → 1.72% for N=3→10, p<.001). (iii) Cross-family rank ensembles match the single best on mean WER and are the most cross-evaluator-robust. The rank-avg configuration at N=5 is the only one reaching p<0.05 under all three evaluators at once (.001/.022/.0002); at N=10, rank-avg ties distil-v3 at mean 1.61% while winning the hubert-lg column (1.60%).

Practical rule. With a known same-family evaluator, use

Table 5. Linear CKA (mean-pooled last hidden state) between audio encoders of six ASR checkpoints on 500 F5-TTS-synthesized waveforms (a random subset of LibriSpeech test-clean). Aliases as in §3: wav2vec 2.0 family (*w2v2-base*, *w2v2-lv60*), Whisper family (*distil-sm*, *distil-v3*, *whisper-med*; the latter three are Distil-Whisper or Whisper checkpoints), HuBERT family (*hubert-lg*).

	w2v2-base	w2v2-lv60	distil-sm	distil-v3	whisper-med	hubert-lg
w2v2-base	1.00	0.65	0.14	0.14	0.15	0.67
w2v2-lv60	0.65	1.00	0.06	0.05	0.06	0.55
distil-sm	0.14	0.06	1.00	**0.98**	**0.98**	0.39
distil-v3	0.14	0.05	**0.98**	1.00	**0.98**	0.40
whisper-med	0.15	0.06	**0.98**	**0.98**	1.00	0.40
hubert-lg	0.67	0.55	0.39	0.40	0.40	1.00

distil-v3 at $N{=}10$. Otherwise, use rank-avg or max-rank over the cross-family pair at $N{=}5$ for the best compute–robustness frontier.

5. Analysis: Toward a Mechanism for the Confound

To test whether the confound reduces to representation similarity, we measure linear CKA (Kornblith et al., 2019) between audio encoders of six ASR checkpoints on 500 F5-TTS-synthesized waveforms (a random subset of LibriSpeech test-clean). Table 5 confirms three intuitive clusters: Whisper–Distil-Whisper (CKA 0.97–0.98 internally), wav2vec 2.0 / HuBERT (0.55–0.67), and near-zero cross-cluster.

The naive prediction (high CKA implies similar WER rankings) is not supported (Figure 2). The Whisper-family pair (distil-sm ↔ whisper-med, CKA 0.978) is anti-correlated ($r{=}-0.52$), while the wav2vec 2.0/HuBERT pair (CKA 0.55) agrees nearly perfectly ($r{=}+0.94$). Pearson(CKA, r) is -0.36 over the six pairs but reverses to $+0.32$ once the same-family outlier is removed: the trend is single-point-driven.

The pattern is consistent not with representation similarity but with an **identity- or lineage-level coupling effect**: a verifier's selections are disproportionately favored by evaluators sharing its ASR lineage, while an independent same-family member returns an uninflated score—structurally analogous to LLM-as-a-judge self-bias (Zheng et al., 2023). Disentangling identity- from family-level coupling is left to future work.

5.1. Quality and cross-evaluator robustness

SIM-o and UTMOS at $N{=}3$ stay within ±0.001 and ±0.005 of F5-TTS (0.9426/3.879), with no measurable degradation under either predictor. Both may be near their resolution limit (UTMOS saturates around 4.0), so human-MOS / NISQA / UTMOSv2 triangulation (Mittag et al., 2021; Baba et al., 2024) is deferred. For cross-evaluator robustness, rank-avg and max-rank at $N{=}5$ reduce WER under all three evaluators simultaneously, while distil-v3 at $N{=}10$ stays flat on w2v2-lv60. The (w2v2-base, distil-v3) pair was fixed in advance as the largest publicly available verifier in each family, prior to any cross-evaluator analysis.

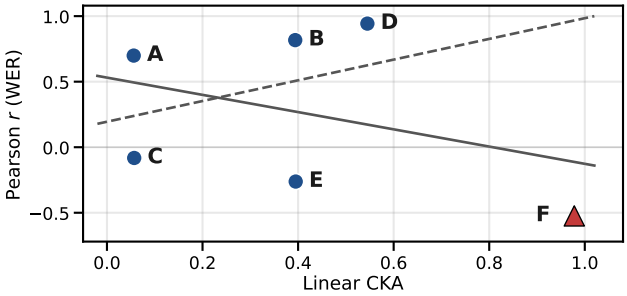

Figure 2. CKA does not predict WER agreement across the $\binom{4}{2}{=}6$ evaluator pairs from {distil-sm, w2v2-lv60, whisper-med, hubert-lg}. Pearson r is computed over 13 BoN configurations on a 500-sample pilot. Pairs: **A**, distil-sm ↔ w2v2-lv60; **B**, distil-sm ↔ hubert-lg; **C**, w2v2-lv60 ↔ whisper-med; **D**, w2v2-lv60 ↔ hubert-lg; **E**, whisper-med ↔ hubert-lg; **F**, distil-sm ↔ whisper-med (same-family, red triangle). The all-6 trend ($r{=}-0.36$, solid) is driven by the same-family outlier F ($r_{\mathrm{WER}}{=}-0.52$ at CKA 0.978). Excluding F, the trend reverses to $r{=}+0.32$ (dashed), confirming that CKA does not monotonically predict WER agreement.

6. Discussion and Conclusion

Any BoN comparison reported under a single ASR evaluator is potentially confounded; we recommend reporting WER under at least two ASR families with disjoint training lineages. Our **cross-family rank ensembles** realize this principle on the selection side: they yield the most cross-evaluator-robust BoN configurations we test, reaching 1.61% mean WER at $N{=}10$ (-12% vs. F5-TTS); the largest reduction under the official F5-TTS evaluator alone, $2.06\% \rightarrow 1.72\%$ (-16.5%), comes from the best same-family verifier. The $N{=}3$ oracle WER of 1.42% leaves substantial headroom for verifier-design work (e.g., adversarial reweighting against in-family inflation).

Limitations. Our study uses one TTS backbone (F5-TTS) on LibriSpeech-PC test-clean; generalization to other backbones (CosyVoice 2, MaskGCT) and a fourth ASR family (Parakeet, Canary), plus human-MOS triangulation, is left to future work.

References

Anastassiou, P., Chen, J., Chen, J., Chen, Y., Chen, Z., et al. Seed-TTS: A family of high-quality versatile speech

generation models. *arXiv:2406.02430*, 2024.

Baba, K., Nakata, W., Saito, Y., and Saruwatari, H. The T05 system for the VoiceMOS challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop*, pp. 818–824, 2024. doi: 10.1109/SLT61566.2024.10832315.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 2020.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K., and Chen, X. 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)*, pp. 6255–6271, 2025.

Cooper, E., Huang, W.-C., Toda, T., and Yamagishi, J. Generalization ability of MOS prediction networks. In *ICASSP*, pp. 8442–8446, 2022.

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

Eskimez, S. E., Wang, X., Thakker, M., Li, C., Tsai, C.-H., Xiao, Z., Yang, H., Zhu, Z., Tang, M., Tan, X., Liu, Y., Zhao, S., and Kanda, N. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot TTS. *arXiv:2406.18009*, 2024.

Feng, S., Halpern, B. M., Kudina, O., and Scharenborg, O. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567, 2024.

Gandhi, S., von Platen, P., and Rush, A. M. Distil-Whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv:2311.00430*, 2023.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *NeurIPS Workshop on Deep Generative Models and Downstream Applications (arXiv:2207.12598)*, 2021.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM TASLP*, 2021.

Hu, Y., Chen, C., Wang, S., Chng, E. S., and Zhang, C. Robust zero-shot text-to-speech synthesis with reverse inference optimization. *arXiv:2407.02243*, 2024.

Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., et al. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *ICML*, 2024.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 2020.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *ICML*, 2019.

Meister, A., Novikov, M., Karpov, N., Bakhturina, E., Lavrukhin, V., and Ginsburg, B. LibriSpeech-PC: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end ASR models. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2023.

Mittag, G., Naderi, B., Chehadi, A., and Möller, S. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech*, pp. 2127–2131, 2021.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023.

Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., and Saruwatari, H. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Interspeech*, pp. 4521–4525, 2022.

Wang, T., Chen, J., Jia, Q., Wang, S., Fang, R., Wang, H., Gao, Z., Xie, C., Xu, C., Dai, J., Liu, Y., Wu, J., Ding, S., Li, L., Huang, Z., Deng, X., Yu, T., Ma, G., Xiao, H., Chen, Z., Xiang, D., Wang, Y., Zhu, Y., Xiao, Y., Wang, J., Wang, Y., Ding, S., Huang, J., Xu, J., Tayier, Y., Hu, Z., Gao, Y., Zheng, C., Ye, Y., Li, Y., Wan, L., Jiang, X., Wang, Y., Cheng, S., Song, Z., Tang, X., Xu, X., Zhang, N., Chen, H., Jiang, Y. E., and Zhou, W. Weaver: Foundation models for creative writing. *arXiv:2401.17268*, 2024.

Wang, Y., Zhan, H., Liu, L., Zeng, R., Guo, H., Zheng, J., Zhang, Q., Zhang, X., Zhang, S., and Wu, Z. MaskGCT: Zero-shot text-to-speech with masked generative codec transformer. In *International Conference on Learning Representations*, 2025.

Yang, Y., Liu, S., Li, J., Hu, Y., Wu, H., Wang, H., Yu, J., Meng, L., Sun, H., Liu, Y., Lu, Y., Yu, K., and Chen, X. Pseudo-autoregressive neural codec language models for efficient zero-shot text-to-speech synthesis. *arXiv:2504.10352*, 2025.

Yu, F., Li, Y., and Wang, B. Scaling flaws of verifier-guided search in mathematical reasoning. *arXiv:2502.00271*, 2025.

Zhang, D., Li, Z., Li, S., Zhang, X., Wang, P., Zhou, Y., and Qiu, X. SpeechAlign: Aligning speech generation to human preferences. *NeurIPS (arXiv:2404.05600)*, 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *NeurIPS*, 2023.