How Far Can Synthetic Speech Take Us? Synthetic Data Recipes for Pretraining Multilingual Speech Language Models

Anonymous authors
Paper under double-blind review

Abstract

Scaling pretraining data has been central to the progress in large language models (LLMs). However, extending scaling recipes to speech-enabled multimodal LLMs is constrained by the scarcity of large, diverse, and permissively licensed speech corpora, a bottleneck even more acute in the speech domain for underrepresented languages. Synthetic speech data can potentially address this gap, but its effectiveness as a training substitute for real data remains unquantified. In this work, we systematically study this question across six languages, varying real-synthetic mixtures, training curricula, and augmentations to the synthetic corpus. Our results show that the quality gap between real and synthetic-only pretraining increases as the training-token budget increases. However, this gap largely disappears when models are pretrained on joint mixtures of synthetic and real speech data: up to 50% of the training data can be synthetic with no detectable degradation across any tested language, while some languages tolerate substantially larger synthetic fractions. Diversifying synthetic data by broadening speaker coverage and performing simple augmentations like reverberation and noise also narrows the synthetic-real gap. Overall, our work provides practical guidance for constructing and using synthetic speech to scale multilingual speech language models under real-world data availability and licensing constraints.

1 Introduction

Recent progress in the capabilities of speech LLMs has relied heavily on the use of large, high-quality speech corpora in pretraining mixtures (Xu et al., 2025; KimiTeam et al., 2025). Qwen3-Omni (Xu et al., 2025), for instance, trains on 20 million hours of supervised audio, while Kimi-Audio uses more than 13 million hours of audio during pretraining (KimiTeam et al., 2025). However, acquiring speech data at this scale remains a challenge for developers outside of frontier labs due to the lack of large and diverse speech corpora that are openly available. Moreover, public availability does not necessarily guarantee that the data are permissively licensed for model development. Finally, the resources that are available and compliant provide highly uneven coverage across languages (Ardila et al., 2020; Pratap et al., 2024).

Synthetic speech provides an avenue for extending limited training data without requiring expensive collection of additional real speech recordings. TTS-generated speech has been successfully used to supplement scarce speech data in ASR and, more recently, in speech language model training (Rossenbach et al., 2020; Laptev et al., 2020; KimiTeam et al., 2025; Maimon et al., 2025). However, existing results do not clarify how closely synthetic speech can match the training benefits of real speech during pretraining, how synthetic data should be combined when real data is available, and what modifications make synthetic data more effective. To study these questions, we conceptualize the *synthetic–real speech performance gap*, which measures the performance difference between pretraining speech-enabled LLMs on real speech data compared to synthetically generated speech, and measure this gap through controlled speech language model pretraining across six languages and two model scales. For each language, we generate synthetic speech using the transcripts of real transcribed speech, yielding real and synthetic datasets with matched linguistic content, but different acoustic realizations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Using this matched framework, we perform different interventions on how synthetic speech is used and constructed: its proportion in the real–synthetic training mixture, the ordering of real and synthetic data during training, and properties of the synthetic corpus, including speaker coverage, acoustic variation, and quality-based filtering. We measure the impact of these interventions on the synthetic–real performance gap. Our experiments yield the following findings:

- **Synthetic speech remains an imperfect substitute for real speech, and the gap grows with training budget.** Synthetic-data-only pretraining consistently underperforms real-data pretraining, although the size of the gap varies substantially across languages. At a 1B-token budget, the increase in loss between the two settings ranges from just 1.0% for Spanish to 16.7% for Korean. Across all tested languages and model scales, this gap widens as the training budget increases.
- **Combining synthetic speech with real speech closes most of this gap.** Replacing up to 50% of the training tokens with synthetic speech does not lead to detectable degradation relative to real-data-only training. Several mixtures even improve performance. The amount of synthetic data that can be used without degradation, though, varies greatly across languages. Training order matters too: training on real data after synthetic data recovers most of the performance gap.
- **Ensuring diversity in synthetic speech data improves its effectiveness, more so than conventional quality filtering.** Increasing speaker coverage and introducing variation in recording and transmission conditions substantially narrows the synthetic-real gap, closing a median 32.9% of the gap across our augmentation experiments. In contrast, filtering synthetic utterances using common quality criteria does not consistently outperform random removal. These results suggest that ensuring synthetic data covers the variability of natural speech is more important for pretraining utility than selecting individual examples according to current quality criteria.

Together, these results show that the value of synthetic speech for model pretraining depends not just on how much data can be generated: how it is generated and how it is used during training greatly influence its utility. Our findings provide practical takeaways for using synthetic speech as a scalable complement when large, diverse, and permissively licensed real-speech corpora are unavailable.

2 Related Work

Synthetic speech for ASR. Prior work has used TTS-generated speech to improve ASR when real transcribed speech is limited (Rosenberg et al., 2019; Rossenbach et al., 2020). More recently, Minixhofer et al. (2024) investigate when synthetic speech can approach real-data ASR performance, whereas Mizumoto et al. (2025) find that synthetic-only training can degrade ASR/AST (automatic speech translation) performance but mixtures with real data help. Multilingual studies explore real–synthetic ratios and synthesis choices in low-resource ASR (DeRenzi et al., 2025; Hao et al., 2026). Most closely, Labrak et al. (2026) study fixed-budget real-synthetic replacement in LLM-based ASR. These works largely study synthetic speech in the context of downstream ASR/AST performance. We instead study how synthetic speech behaves as a pretraining resource for audio and multimodal LLMs, including its effects on both audio-token and audio-text modeling.

Synthetic speech for audio and multimodal pretraining. Synthetic speech is also increasingly used in large-scale pretraining. Zeng et al. (2024) train on 600 billion synthetic interleaved speech–text tokens, while Maimon et al. (2025) study synthetic speech as part of a broader compute-efficient SpeechLM recipe. Synthetic speech is likewise used in large-scale audio foundation models such as Kimi-Audio (KimiTeam et al., 2025). Relatedly, Kulkarni et al. (2026) employ a synthetic-to-real curriculum for task-specific audio post-training. These works establish that synthetic speech can be useful at scale, but do not systematically characterize the principles governing its construction and use relative to real speech. We address this gap.

3 Method

We study how replacing real speech with synthetic speech affects audio-text pretraining. Our goal is to isolate the effect of replacing real recordings with synthetic versions while controlling for both the linguistic content seen during training and the total training budget. To do so, we construct real and synthetic versions of the same underlying transcript data and compare models trained with different proportions of the two under a fixed training-token budget. All comparisons are evaluated

on held-out real speech, so that differences measure how well training on synthetic speech transfers to the real-data target distribution.

Note that the models, discrete tokenizers, and autoregressive objectives for speech are the same as those used in broader audio language modeling. We may therefore use *audio* when referring to these components where appropriate, but our experiments only use speech data.

3.1 Controlled Real-Synthetic Comparison

For an utterance with transcript $\mathbf{t}^i$, let $\mathbf{a}_R^i$ denote its real recording and $\mathbf{a}_S^i$ the synthetic speech obtained by applying a text-to-speech model to the same transcript. The corresponding training examples are therefore $(\mathbf{a}_R^i, \mathbf{t}^i)$ and $(\mathbf{a}_S^i, \mathbf{t}^i)$, which contain the same textual content while differing in their acoustic realization. This construction lets us intervene on the source of the speech while holding the transcript distribution fixed.

Synthetic and real versions of the same utterance can differ in duration and therefore in the number of resulting speech tokens. Thus, our comparisons use the same number of training tokens rather than the same number of hours or examples. Let B denote the total number of speech and text tokens processed during training. For a real-synthetic mixture, we define its synthetic share as

$$\alpha = \frac{B_S}{B_R + B_S}, \tag{1}$$

where B_R and B_S are the number of tokens contributed by audio-text examples containing real and synthetic speech, respectively.

For real-only versus synthetic-only comparisons, we use matched transcript content whenever possible. When real and synthetic speech occur within the same training run, we instead partition the transcript set before constructing the two sources, so that a transcript appears in either real or synthetic form but never both. This split prevents effects unrelated to the speech source from driving the observed results, e.g., changes due to repeated exposure to the same linguistic content, or gains from effectively augmenting a sentence with multiple acoustic realizations.

3.2 Real and Synthetic Speech Corpora

We conduct experiments across six languages: English, German, Spanish, French, Korean, and Chinese. All six use the same TTS system (Hu et al., 2026) and synthesis pipeline. In addition to the monolingual settings, we train a multilingual model with an equal number of training tokens sampled from each language.

Data. For each language, we construct a real-speech training corpus of approximately 10k hours from existing speech datasets. English is drawn from Granary (Koluguri et al., 2025), GigaSpeech (Chen et al., 2021), and People's Speech (Galvez et al., 2021). German combines Granary, EuroSpeech (Pfisterer et al., 2025), Multilingual LibriSpeech (MLS; Pratap et al., 2020), ASR Bundestag (Wirth & Peinl, 2024), and mTEDx (Salesky et al., 2021). Spanish uses Granary, MLS, and mTEDx, while French additionally includes EuroSpeech and Vibravox (Hauret et al., 2025). Korean is assembled from KoreaSpeech, KresSpeech, and GyeongsangSpeech distributed through AI Hub (National Information Society Agency (NIA), 2026). Chinese combines WenetSpeech (Zhang et al., 2022), AISHELL-1 (Bu et al., 2017), AISHELL-4 (Fu et al., 2021), MAGICDATA (Magic Data Technology Co., Ltd., 2019), ST-CMDS (Surfingtech, 2017), MagicData-RAMC (Yang et al., 2022), and AliMeeting (Yu et al., 2022). Table 1 summarizes the resulting training pools.

Synthetic Speech Generation. We generate synthetic speech for the transcript of each real-speech sample using Qwen3-TTS (Hu et al., 2026) because of its strong performance and language coverage. For each transcript, we randomly sample a reference speaker from the corresponding language's Common Voice (Ardila et al., 2020) speaker pool and synthesize the speech in that speaker's voice. Reference speakers are sampled independently of the real training corpus.

Language	Real speech sources	Hours
English	Granary, GigaSpeech, People's Speech	10,146
German	Granary, EuroSpeech, MLS, ASR Bundestag, mTEDx	10,139
Spanish	Granary, MLS, mTEDx	10,023
French	Granary, EuroSpeech, MLS, mTEDx, Vibravox	10,034
Korean	KoreaSpeech, KresSpeech, GyeongsangSpeech	8,949
Chinese	WenetSpeech, AISHELL-1, AISHELL-4, MAGICDATA, ST-CMDS, MagicData-RAMC, AliMeeting	10,000

Table 1: Real-speech training corpora used for each language.

3.3 Audio-Text Pretraining

Multimodal LLMs represent different modalities as token sequences and optimize a shared autoregressive objective. For our study, we consider only the audio-text setting, where speech audio data is represented by discrete tokens $\mathbf{a} = (a_1, \ldots, a_{|\mathbf{a}|})$ followed by the tokens of its transcript $\mathbf{t} = (t_1, \ldots, t_{|\mathbf{t}|})$. Omitting boundary tokens for clarity, a decoder-only language model with parameters θ factorizes their joint likelihood as

$$p_\theta(\mathbf{a}, \mathbf{t}) = \prod_{i=1}^{|\mathbf{a}|} p_\theta(a_i \mid a_{<i}) \prod_{j=1}^{|\mathbf{t}|} p_\theta(t_j \mid \mathbf{a}, t_{<j}). \tag{2}$$

Thus, a next-token prediction objective can be used to train the model to both a) model the speech-token sequence and b) predict the transcript conditioned on the preceding speech. The losses corresponding to these two components are

$$\mathcal{L}_{\text{speech}} = -\frac{1}{|\mathbf{a}|} \sum_{i=1}^{|\mathbf{a}|} \log p_\theta(a_i \mid a_{<i}), \qquad \mathcal{L}_{\text{text}} = -\frac{1}{|\mathbf{t}|} \sum_{j=1}^{|\mathbf{t}|} \log p_\theta(t_j \mid \mathbf{a}, t_{<j}). \tag{3}$$

In words, $\mathcal{L}_{\text{speech}}$ tells us how well the model predicts the distribution of speech tokens, while $\mathcal{L}_{\text{text}}$ measures transcript text prediction abilities conditioned on the speech. The joint loss over the complete sequence is then defined as

$$\mathcal{L}_{\text{joint}} = \frac{|\mathbf{a}|\,\mathcal{L}_{\text{speech}} + |\mathbf{t}|\,\mathcal{L}_{\text{text}}}{|\mathbf{a}| + |\mathbf{t}|}. \tag{4}$$

where $\mathcal{L}_{\text{joint}}$ is the objective used during training. Unless otherwise specified, loss refers to $\mathcal{L}_{\text{joint}}$ throughout the paper.

Model and Training Configuration. We train all models from scratch using the Megatron-LM (Shoeybi et al., 2020) framework. We use WavTokenizer (Ji et al., 2025) at 40 tokens per second as the discrete speech tokenizer and keep it frozen for all experiments. We consider two model scales, 490M and 1.17B parameters; their complete architectural and training configurations are provided in Appendix A.1. Unless otherwise stated, experiments are conducted at both scales.

Models use a sequence length of 4096 tokens and a global batch size of 128. The default training budget is 2B tokens unless otherwise specified. When a training corpus contains fewer unique tokens than its assigned budget, examples are repeated until the budget is exhausted.

3.4 Evaluation Protocol

All models are evaluated on held-out real speech. For each language, we use the test sets of the corresponding real-speech corpus after removing documents whose transcripts overlap the training data. The multilingual model is evaluated on the union of the six language-specific evaluation sets.

Our primary metrics are the percentage changes in speech, text, and joint losses defined in Eqs. 3 and 4, relative to the real-data-only baseline with matched token budget. Positive values denote higher loss and therefore worse performance. We separately evaluate the speech losses separately to distinguish whether differences between real and synthetic speech arise in speech-token modeling

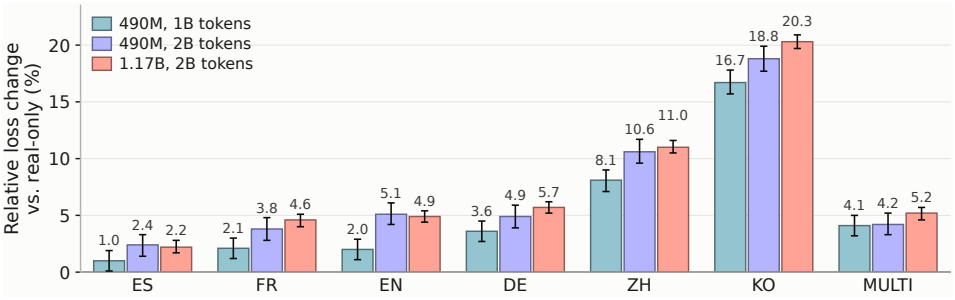


Figure 1: **The synthetic–real pretraining gap varies substantially across languages.** Relative change (%) in $\mathcal{L}_{\text{joint}}$ under synthetic-only training compared with matched real-only training at 1B and 2B tokens; lower is better, and 0% denotes parity with real-only training. Error bars are 95% confidence intervals. Spanish has the smallest gap, whereas Korean exhibits the largest gap.

or in transcript-token prediction conditioned on the speech. We estimate uncertainty from a paired bootstrap over held-out recordings and run-to-run variation from repeated seeds, and combine their variances under an independence assumption. We report 95% confidence intervals. Appendix B.11 reports additional results on downstream tasks.

4 Results

We first characterize the synthetic–real speech performance gap and study how it changes as the training-token budget grows. We then examine how much of this gap remains when synthetic speech is combined with real data or introduced at different stages of training, before testing whether filtering, broader speaker coverage, and acoustic augmentation can narrow it. Finally, Appendix B.5 shows that the same qualitative conclusions extend to interleaved speech-text spans.

4.1 What is the synthetic–real speech performance gap?

We compare synthetic-only and real-only pretraining at matched training-token budgets across two model scales. Figure 1 first measures the relative change in joint loss ($\mathcal{L}_{\text{joint}}$, Eq. 4) across all six languages, trained individually and jointly, at 1B and 2B tokens. Figure 2 then tracks how this gap changes as the training-token budget increases.

The synthetic–real performance gap exists, and the size of the gap varies substantially across languages. For the 490M model trained on 1B tokens, the relative joint-loss increase ranges from only 1.0% for Spanish to 16.7% for Korean. Other European languages, English, French, and German also have relatively smaller loss increases of 2.0%, 2.1%, and 3.6%, respectively, but Chinese shows a substantially larger 8.1% gap. The multilingual runs have smaller gaps (4.1-5.2%) than the macro-average of the corresponding monolingual runs (5.6-8.1%), suggesting that joint multilingual training can partially attenuate the synthetic–real performance gap. This also hints at positive cross-lingual transfer.

The synthetic–real performance gap grows as the training-token budget increases. Figure 2 shows a consistent increase in the gap as training tokens increase across languages and model scales. For the 1.17B model, the relative joint-loss gap grows from 1.1% to 7.7% for English, from 6.6% to 13.6% for Chinese, and from 18.3% to 22.2% for Korean between 1B and 6B training tokens. The same trend holds for Spanish (2.2% $\rightarrow$ 4.4%), French (3.4% $\rightarrow$ 6.9%), German (2.2% $\rightarrow$ 8.4%), and multilingual training (4.2% $\rightarrow$ 6.9%). The 490M models follow the same qualitative trend. Thus, the synthetic–real performance gap is not constant with the training-token budget in our experiments, but the broader cross-language ordering remains stable from the 490M model at 1B tokens to the 1.17B model at 6B tokens: Spanish remains closest to real-only training and Korean the furthest. However, we see some signs of the increase plateauing, particularly in Chinese (13.1% $\rightarrow$ 13.6% when going from 4B to 6B) and Korean (21.1% $\rightarrow$ 22.2% when going from 4B to 6B). At larger budgets, we were constrained by the amount of publicly available speech data

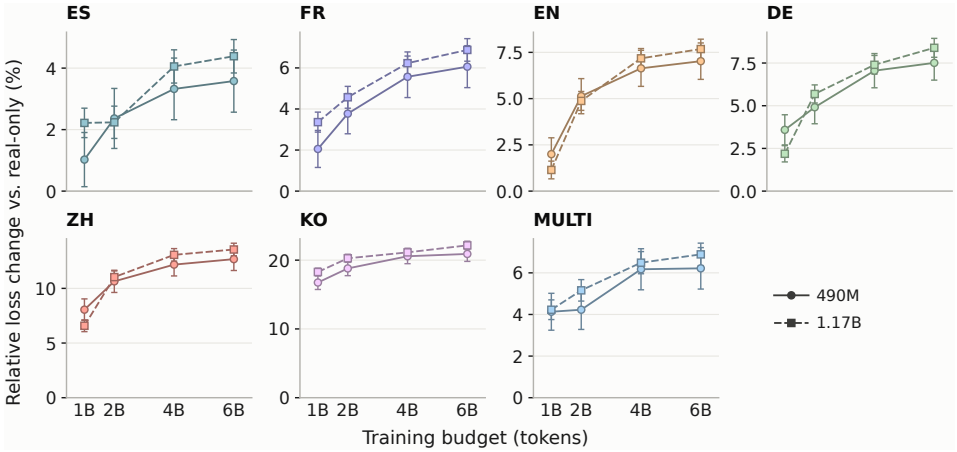


Figure 2: **The synthetic–real performance gap widens as the training-token budget increases.** Relative change (%) in $\mathcal{L}_{\text{joint}}$ compared with matched real-only training as a function of training-token budget for the 490M (solid, circles) and 1.17B (dashed, squares) models; lower is better and 0% denotes parity with real-only training. Error bars are 95% confidence intervals. The gap increases with budget in all seven settings and at both model scales, while the broader cross-language ordering remains stable: Spanish remains closest to real-only training and Korean the furthest. At larger budgets, training entails repeated passes over the underlying corpus.

for each language, and therefore had to repeat training examples. We do not expect increasing the amount of unique data to change the qualitative trend, although it may shift the absolute synthetic–real performance gaps.

The source of the synthetic–real performance gap differs substantially across languages. Appendix Table 10 decomposes the gap into $\mathcal{L}_{\text{speech}}$ and $\mathcal{L}_{\text{text}}$ from Eq. 3. For German, Spanish, and French, synthetic speech preserves or improves transcript-token prediction, with relative text-loss changes of -0.8%, -24.2%, and -4.0%, respectively, despite consistent degradation in speech-token modeling (2.6-5.7%). The pattern differs for English, Chinese, and Korean. English text loss increases by 13.2% compared with a 4.6% increase in speech loss, while the disparity is much larger for Chinese (82.3% versus 9.6%) and Korean (93.9% versus 18.4%). Surprisingly, Appendix B.5 does not show a similar degradation in text-token prediction for interleaved speech-text sequences. One possible explanation is that the predicted text tokens receive more textual grounding from their context, leading to smaller errors. Appendix C.1 examines the synthetic–real performance gap with a smaller TTS model: ZipVoice-Distill.

> **Takeaway 1:** The synthetic–real performance gap is significant in all measured languages and model sizes and grows with training budget, but its relative size across languages is stable across scales. Small-scale runs can therefore identify for which languages synthetic speech data is most likely to be effective.

4.2 How to combine synthetic data if some real data is available?

Having characterized the synthetic–real performance gap, we next ask how synthetic and real data should be combined to best utilize the synthetic data: can we lessen the synthetic–real speech performance gap with a curriculum, or should the two be mixed throughout training? We subsequently ask how large the synthetic proportion can become before degradation relative to real-only training is detected.

Ending on real speech is substantially better than ending on synthetic speech. We fix the training budget at 2B tokens with a 1:1 real-synthetic allocation and compare a mixture of real and

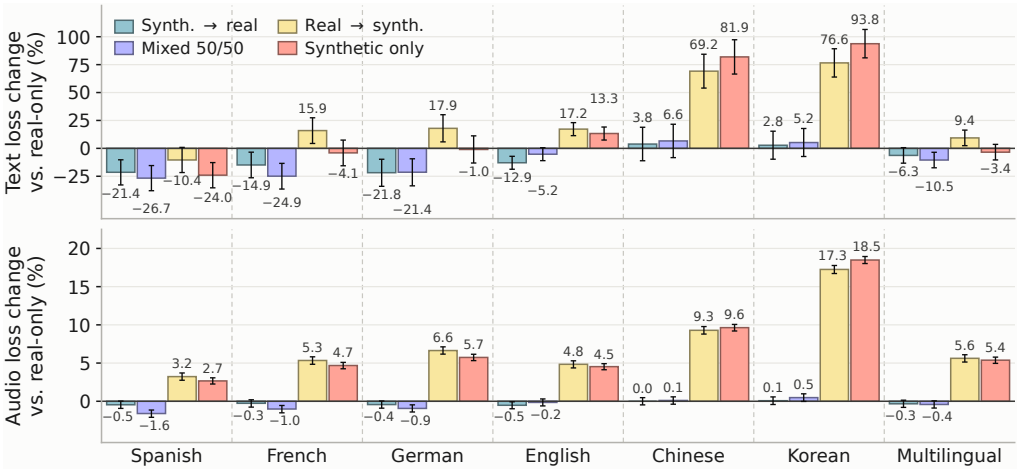


Figure 3: **Mixing or ending training on real speech performs substantially better than ending on synthetic speech.** Relative changes in text loss ($\mathcal{L}_{\text{text}}$, Section 3.3) and speech loss ($\mathcal{L}_{\text{speech}}$, Section 3.3) for synthetic-only training, mixing real-synthetic throughout training, synthetic→real and real→synthetic curricula at the 1.17B scale and a 2B-token budget. Lower is better; zero denotes the matched real-only baseline. Bars show 95% confidence intervals.

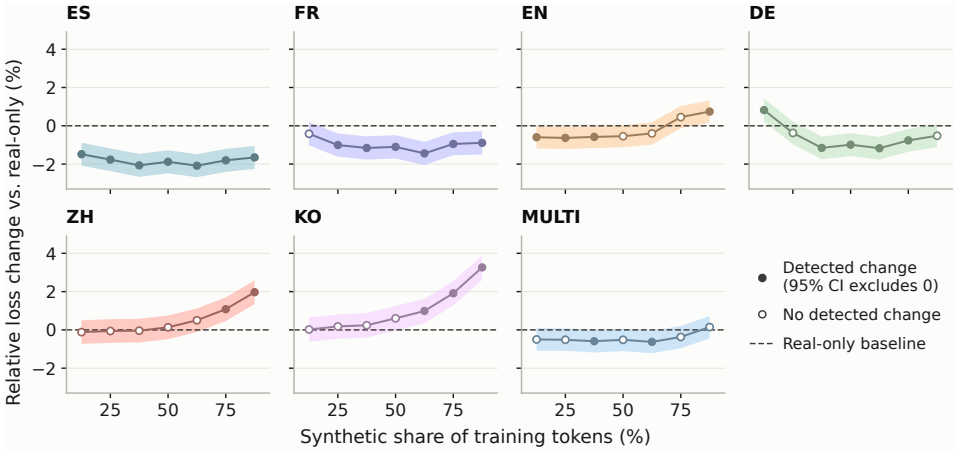


Figure 4: **Up to 50% of the training tokens can be synthetic with no detected degradation.** Relative loss change (%) across real-synthetic mixture ratios for the 1.17B model at a 2B-token budget. Lower is better; zero denotes parity with real-only pretraining. Error bars are 95% confidence intervals. Spanish, French, and German show no detected degradation even with 87.5% synthetic training tokens (German has one isolated detected degradation at 12.5% synthetic, which does not persist at any larger tested proportion). Korean first shows degradation at 62.5% synthetic speech.

synthetic speech with synthetic→real and real→synthetic curricula. The real and synthetic datasets use disjoint transcript sets. Figure 3 reports results for the 1.17B model.

The ordering effect is visible across both speech-token and text-token prediction. Synthetic→real finishes with speech-token loss close to the real-only baseline across all languages, with relative speech-loss changes ranging from -0.5% to 0.1%. Mixed training performs similarly, with speech-loss changes ranging from -1.6% to 0.5%. In contrast, real→synthetic increases speech loss by 4.8% for English, 9.3% for Chinese, 17.3% for Korean, and 5.6% for multilingual training. These values closely approach the corresponding synthetic-only values, and Spanish, French, and German show similar trends. Similarly, combining synthetic speech with real data largely mitigates the sharp degradation in transcript prediction for Chinese and Korean. Again, ending on synthetic speech is much less effective for these two languages (69.2% for Chinese and 76.6% for Korean). Interestingly, for French (15.9% versus -4.1%) and German (17.9% versus -1.0%), real→synthetic train-

ing performs substantially worse than synthetic-only training in terms of relative text-loss change. Overall, when the real-synthetic composition is held fixed, either mixing the sources or ending on real speech keeps speech-token performance close to real-only training. Appendix B.4 further shows that restricting real speech to the cooldown phase, only 20% of the training tokens, still brings loss close to the real-only baseline.

A substantial proportion of pretraining data can be synthetic before its effectiveness declines. We vary the synthetic share while holding the total training budget fixed at 2B tokens. Figure 4 shows the results. For the 1.17B model, synthetic speech can constitute up to 50% of the training tokens with no detected degradation across any tested language. Several intermediate mixtures additionally improve over real-only training, particularly for Spanish and French, where nearly every tested synthetic proportion improves over real-only training. This suggests that the degradation observed under synthetic-only training can be largely mitigated when some real data is available. Korean departs from the real-only regime earlier than the other settings, consistent with its much higher synthetic-only loss change. For the 1.17B model, degradation is first detected at 62.5% synthetic speech for Korean, compared with 87.5% for English and 75% for Chinese. Again, surprisingly, multilingual training shows no detected degradation across the tested mixture ratios.

> **Takeaway 2:** Even small amounts of real data can close most of the synthetic–real performance gap when mixed into a synthetic speech dataset or presented at the end of training. For some languages, using just 12.5% real speech tokens in the training data mixture is sufficient to avoid detectable degradation relative to real-only training. Other languages, though, such as Korean, require a more balanced mixture.

4.3 Can We Make Synthetic Speech More Effective?

We study whether three common choices for improving the quality and realism of synthetic data can To further narrow the synthetic–real performance gap: filtering low-quality utterances, increasing speaker coverage, and introducing additional acoustic variation.

Standard quality scores are poor predictors of pretraining value. We test removing the lowest-scoring 20% of clips according to ASR error rate, transcript-length expansion, and a composite criterion additionally incorporating voice-activity ratio. We upsample the remaining clips to reach 2B tokens. We see that filtering ($+0.51\%$ relative loss change on average across filtering methods and languages) does not consistently outperform removing the same number of clips at random ($+0.17\%$ relative loss change on average across languages). A synthetic utterance can therefore score well under these quality criteria while still contributing little towards covering the distribution encountered in real speech. For more details and results, refer to Appendix B.6.

Variation in recording and transmission conditions substantially narrows the synthetic–real performance gap. We augment synthetic speech with either delivery variation (speed, pitch, and compression), variation in recording and transmission conditions (reverberation, noise, and codec effects), or both. Appendix A.3 contains the augmentation recipes and Appendix B.7 reports the corresponding loss results. Figure 5 (left) shows that introducing recording and transmission variation (also referred to as *channel augmentation* in this work) consistently improves performance across languages. At the 1.17B scale, channel augmentation reduces joint loss relative to unaugmented synthetic speech by 2.2% on average across the six monolingual settings, with larger gains for Chinese (3.1%) and Korean (3.3%). Delivery perturbations are substantially less effective. At the 1.17B scale, they change the loss by at most 0.4% and can sometimes hurt (e.g., multilingual: $+0.3\%$). Moreover, combining delivery and channel augmentation does not improve upon channel augmentation alone, and channel augmentation has a lower joint loss than the combined recipe in every 2B setting. Interestingly, channel augmentation improves the text-loss point estimate for all six monolingual settings, ranging from 2.1% for Spanish to 11.4% for French, but worsens the combined multilingual text loss by 6.3%.

Broader speaker coverage substantially improves transfer. We test the effect of adding more reference speakers by varying the number of speakers sampled from Common Voice across four lan-

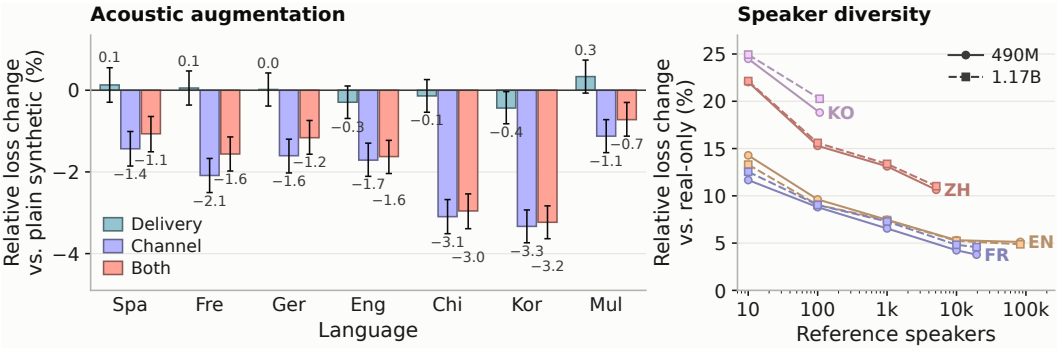


Figure 5: **Channel augmentation and broader speaker coverage narrow the synthetic–real performance gap.** (Left) Relative loss changes (%) compared with plain synthetic speech for augmentation recipes on the 1.17B model at 2B tokens. Lower is better; 0% denotes parity with synthetic-only training without augmentation. Channel augmentations such as noise, reverberation, and codec effects provide larger gains than delivery augmentations such as speed, pitch, and compression. (Right) Relative loss changes (%) compared with real-only training at 2B tokens as the reference-speaker pool grows. Lower is better; English shows little additional improvement beyond 10k speakers, while the other languages do not plateau within the available Common Voice pools.

guages: English, French, Chinese, and Korean. Figure 5 (right) shows that holding the transcripts fixed while increasing the reference-speaker pool closes roughly two-thirds of the initial English loss gap to real speech (14.3% → 5.3%). Joint losses improve rapidly and begin to saturate at around 10,000 speakers, while speech loss continues to improve modestly as the pool expands further (Appendix Table 11). Korean exhibits a similar effect when increasing the pool from 10 to all 107 available speakers (24.5% → 18.8%). However, Figure 5 also suggests that cross-language differences cannot be explained simply by reference-speaker count. Appendix B.2 likewise finds no clear explanation from dataset-quality metrics such as SNR and MOS. Differences in dialectal coverage may be another factor, but we cannot test this with the available metadata.

Collectively, the results from this section suggest that using methods that improve coverage of the variability of natural speech are better strategies for narrowing the synthetic–real performance gap than selecting individual synthetic utterances on the basis of conventional quality scores.

Takeaway 3: Broader speaker coverage and augmenting the generated speech with variation in recording and transmission conditions can significantly improve the pretraining value of a synthetic corpus; this strategy should be prioritized over applying conventional quality filtering techniques to synthetic data, which do not lead to consistent performance benefits.

5 Conclusion

We systematically study how useful synthetic speech is for multilingual speech language model pretraining and present guidelines for increasing its effectiveness. At a 1B-token budget, the synthetic-real loss gap varies substantially across languages, from 1% for Spanish to 16.7% for Korean, and grows as the training-token budget increases. However, this synthetic-only gap considerably overstates the cost of using synthetic speech alongside real data: our experiments show that synthetic speech can constitute up to 50% of the training tokens without detected degradation across any tested language. In a curriculum, ending training on real speech largely recovers the deficits observed under synthetic-only training, even when real speech is confined to the cooldown phase of pretraining. Our analyses further show that broader speaker coverage and augmenting generated speech with variation in recording and transmission conditions, such as noise and reverberation, substantially improve the effectiveness of synthetic speech. Overall, our results suggest that synthetic speech can help democratize speech language model pretraining, but its effectiveness depends on the language and how it is combined with real speech, with real data being particularly valuable towards the end of training.

AI USE STATEMENT

The authors used LLMs to assist with implementing the proposed multimodal changes to Megatron-LM and cleaning the real speech datasets. All AI-assisted code was reviewed and verified by the authors before use, and all reported results and interpretations were independently checked by the authors. LLMs were also used to refine author-written text for clarity and presentation. They were not used to generate the research idea, identify the research gap, propose or refine the central hypotheses of this work, design or provide feedback on the research methodology or experiments, or interpret the results. Finally, generative AI was used to generate the synthetic speech studied in this work using Qwen3-TTS, though the underlying transcripts are from real non-synthetic datasets. The other required disclosure categories, such as developing theoretical models, formulating mathematical claims, etc do not apply to this work. We have reviewed all AI-assisted material and take responsibility for the final content of the paper, including any text, code, claims, or other artifacts produced with the aid of generative AI.

REFERENCES

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520/.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pp. 1–5, 2017. doi: 10.1109/ICSDA.2017.8384449.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech 2021*, 2021.

Brian DeRenzi, Anna Dixon, Mohamed Aymane Farhi, and Christian Resch. Synthetic voice data for automatic speech recognition in African languages. In Ernesto Luis Estevanell-Valladares, Alicia Picazo-Izquierdo, Tharindu Ranasinghe, Besik Mikaberidze, Simon Ostermann, Daniil Gurgurov, Philipp Mueller, Claudia Borg, and Marián Šimko (eds.), *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pp. 152–186, Varna, Bulgaria, September 2025. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2025.lowresnlp-1.16/.

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, and Jingdong Chen. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario, 2021. URL https://arxiv.org/abs/2104.03603.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage, 2021. URL https://arxiv.org/abs/2111.09344.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux,

Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy

Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yun Hao, Reihaneh Amooie, Wietse de Vries, Rik van Noord, and Martijn Wieling. Investigating the role of synthetic data augmentation and training strategies on improving low-resource language ASR. In Stelios Piperidis, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral (eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 5633–5639, Palma de Mallorca, Spain, May 2026. ELRA Language Resource Association. doi: 10.63317/4c4ad7t5i967. URL https://aclanthology.org/2026.lrec-1.442/.

Julien Hauret, Malo Olivier, Thomas Joubaud, Christophe Langrenne, Sarah Poirée, Véronique Zimpfer, and Éric Bavu. Vibravox: A dataset of french speech captured with body-conduction audio sensors, 2025. URL https://arxiv.org/abs/2407.11828.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report, 2026. URL https://arxiv.org/abs/2601.15621.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling, 2025. URL https://arxiv.org/abs/2408.16532.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025. URL https://arxiv.org/abs/2504.18425.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, 2017. doi: 10.1109/ICASSP.2017.7953152.

Nithin Rao Koluguri, Monica Sekoyan, George Zelenfroynd, Sasha Meister, Shuoyang Ding, Sofia Kostandian, He Huang, Nikolay Karpov, Jagadeesh Balam, Vitaly Lavrukhin, Yifan Peng, Sara Papi, Marco Gaido, Alessio Brutti, and Boris Ginsburg. Granary: Speech recognition and translation dataset in 25 european languages, 2025. URL https://arxiv.org/abs/2505.13404.

Apoorva Kulkarni, Kaousheik Jayakumar, Sreyan Ghosh, Utathya Aich, Ramani Duraiswami, and Dinesh Manocha. Tempo: Temporally-grounded multi-task post-training for large audio-language models, 2026. URL https://arxiv.org/abs/2608.29999.

Yanis Labrak, Dairazalia Sanchez-Cortes, Sergio Burdisso, Séverin Baroudi, Shashi Kumar, Esaú Villatoro-Tello, Srikanth Madikeri, Manjunath K E, Oldřich Plchot, Kadri Hacioğlu, Petr Motlicek, and Andreas Stolcke. How to leverage synthetic speech for llm-based asr systems?, 2026. URL https://arxiv.org/abs/2606.29031.

Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin. You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 439–444. IEEE, October 2020. doi: 10.1109/cisp-bmei51763.2020.9263564. URL http://dx.doi.org/10.1109/CISP-BMEI51763.2020.9263564.

Magic Data Technology Co., Ltd. MAGICDATA mandarin chinese read speech corpus. Open Speech and Language Resources (OpenSLR), SLR68, May 2019. URL https://www.openslr.org/68/. Accessed: 2026-09-24.

Gallil Maimon, Avishai Elmakies, and Yossi Adi. Slamming: Training a speech language model on one GPU in a day. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12201–12216, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.631. URL https://aclanthology.org/2025.findings-acl.631/.

Christoph Minixhofer, Ondrej Klejch, and Peter Bell. Beyond oversmoothing: Evaluating ddpm and mse for scalable speech synthesis in asr, 2024. URL https://arxiv.org/abs/2410.12279.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, pp. 2127–2131. ISCA, August 2021. doi: 10.21437/interspeech.2021-299. URL http://dx.doi.org/10.21437/Interspeech.2021-299.

Tomoya Mizumoto, Atsushi Kojima, Yusuke Fujita, Lianbo Liu, and Yui Sudo. Is Synthetic Data Truly Effective for Training Speech Language Models? In *Interspeech 2025*, pp. 1808–1812, 2025. doi: 10.21437/Interspeech.2025-2693.

National Information Society Agency (NIA). AI Hub. https://www.aihub.or.kr/, 2026. Source of the KoreaSpeech, KrespSpeech, and GyeongsangSpeech corpora. Accessed 24 September 2026.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. SpiRit-LM: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025. doi: 10.1162/tacl_a_00728. URL https://aclanthology.org/2025.tacl-1.2/.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

Samuel Pfisterer, Florian Grötschla, Luca Lanzendörfer, Florian Yan, and Roger Wattenhofer. Eurospeech: A multilingual speech corpus. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-2061. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/58ea63de01321ee52d06b48026981c40-Paper-Datasets_and_Benchmarks_Track.pdf.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pp. 2757–2761. ISCA, October 2020. doi: 10.21437/interspeech.2020-2826. URL http://dx.doi.org/10.21437/Interspeech.2020-2826.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *J. Mach. Learn. Res.*, 25(1), January 2024. ISSN 1532-4435.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Chandan K A Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors, 2021. URL https://arxiv.org/abs/2010.15258.

Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. Speech recognition with augmented synthesized speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 996–1002, 2019. doi: 10.1109/ASRU46091.2019.9003990.

Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7069–7073, 2020. doi: 10.1109/ICASSP40776.2020.9053008.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022, 2022. URL https://arxiv.org/abs/2204.02152.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. The multilingual tedx corpus for speech recognition and translation, 2021. URL https://arxiv.org/abs/2102.01757.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020. URL https://arxiv.org/abs/1909.08053.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus, 2015. URL https://arxiv.org/abs/1510.08484.

Surfingtech. ST-CMDS-20170001_1: Free ST chinese mandarin corpus. https://openslr.org/38/, 2017. OpenSLR 38.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective, 2024. URL https://arxiv.org/abs/2410.05192.

Johannes Wirth and René Peinl. Asr bundestag: A large-scale political debate dataset in german. In Kohei Arai (ed.), *Intelligent Systems and Applications*, pp. 190–202, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-47724-9.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, et al. Open source magicdata-ramc: A rich annotated mandarin conversational (ramc) speech dataset. *arXiv preprint arXiv:2203.16844*, 2022.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6167–6171, 2022. doi: 10.1109/ICASSP43922.2022.9746465.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. Scaling speech-text pre-training with synthetic interleaved data, 2024. URL https://arxiv.org/abs/2411.17607.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6182–6186, 2022. doi: 10.1109/ICASSP43922.2022.9746682.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. *arXiv preprint arXiv:2506.13053*, 2025.

Appendix

A Experimental Details **17**

A.1 Model configurations and training 17

A.2 Validation and Test Sets 17

A.3 Acoustic augmentation recipes 17

A.4 Common Voice Reference-Speaker Pools 18

B Additional Results **19**

B.1 Real and Synthetic Training Trajectories 19

B.2 Corpus quality scores 20

B.3 Mixture-sweep summary 20

B.4 Real Speech During Learning-Rate Decay 20

B.5 Interleaved audio-text training 21

B.6 Quality-based filtering 23

B.7 Acoustic augmentation on each loss axis 23

B.8 Where the synthetic-real gap comes from 25

B.9 Reference-speaker diversity 26

B.10 Multilingual training, language by language 26

B.11 Downstream transcription 26

C Controls **29**

C.1 Replication with a Smaller TTS System 29

A Experimental Details

A.1 Model configurations and training

We train two model sizes in Megatron-LM (Shoeybi et al., 2020), both decoder-only transformers with grouped-query attention and a frozen WavTokenizer (Ji et al., 2025) front end. Table 2 gives their configurations. We report embedding and non-embedding counts separately; the latter is the more meaningful measure of capacity when comparing the two scales.

Configuration	Small	Large
Layers	18	28
Hidden dimension	1024	1536
Attention heads	16	16
Key-value heads	4	4
FFN dimension	3072	4608
Embedding parameters	273M	409M
Non-embedding parameters	217M	760M
Total parameters	490M	1.17B

Table 2: Model configurations. The two scales differ by $3.5\times$ in non-embedding capacity. The speech tokenizer is frozen throughout training and is not counted in these parameter totals.

We use a warmup-stable-decay (WSD) learning-rate schedule (Wen et al., 2024) with 500 warmup steps, followed by a stable phase at the peak learning rate and a linear decay to zero over the final 20% of training. The peak learning rate is 3×10^{-4} for the 490M model and 2×10^{-4} for the 1.17B model.

A.2 Validation and Test Sets

Endpoint results are evaluated on held-out real-speech test sets, while the in-training trajectories use the corresponding validation sets. We use the validation and test splits distributed with each corpus rather than constructing new splits. Documents whose transcripts overlap the training data are removed before endpoint evaluation, following Section 3.4. Table 3 gives the hours and token counts of the resulting real-speech evaluation sets.

Lang.	Split	Hours	Text tok.	Speech tok.	Total tok.
English	Val	44.16	720,814	6,252,424	7,043,564
	Test	72.24	1,436,937	10,219,748	11,776,271
German	Val	14.28	184,454	2,041,098	2,235,956
	Test	44.59	610,006	6,354,909	7,010,266
Spanish	Val	9.95	118,923	1,422,202	1,548,316
	Test	9.95	119,546	1,422,610	1,549,272
French	Val	10.05	118,422	1,436,332	1,561,984
	Test	10.06	127,560	1,438,774	1,573,609
Korean	Val	100.00	1,487,831	14,084,906	15,808,567
	Test	100.00	1,487,770	14,085,631	15,809,435
Chinese	Val	65.85	915,173	9,236,066	10,317,208
	Test	100.00	1,519,984	13,980,139	15,783,188

Table 3: Hours and token counts of the real validation and test sets.

A.3 Acoustic augmentation recipes

Synthetic speech does not capture the full diversity present in real recordings, including differences in speaking style, recording environments, background noise, and transmission conditions. We therefore test whether introducing this variability to synthetic data can narrow the synthetic-real pretraining gap. We augment each synthetic utterance once while keeping the underlying clip set and total training-token budget fixed.

Table 4 lists the recipes. We consider two types of augmentation. Delivery augmentation changes how the speech is produced using speed perturbation, pitch shifting, or dynamic-range compression. Channel augmentation changes the recording or transmission conditions using reverberation, additive noise, or codec effects. For each synthetic clip, one augmentation is sampled from the corresponding family. We additionally evaluate a both condition that applies one delivery augmentation followed by one channel augmentation. The perturbations are intentionally moderate so that they increase variation without substantially changing intelligibility. All augmentations are applied to the waveform before speech tokenization.

Family	Member	Augmentation settings
Delivery	Speed	Tempo is increased or decreased by 0.1-0.2, with the direction selected with equal probability.
	Pitch	Pitch is shifted by 1.0-2.0 semitones, with upward and downward shifts selected with equal probability.
	Compression	Dynamic-range compression is applied using a threshold of -30 to -18 dB, a ratio of 3-6, an attack time of 2-10 ms, and a release time of 50-200 ms. Make-up gain is applied to restore the input RMS level.
Channel	Reverberation	A room impulse response is sampled uniformly from the small- and medium-room subsets of RIRS_NOISES (Ko et al., 2017).
	Noise	MUSAN noise (Snyder et al., 2015) is added at an SNR of 15-25 dB using a randomly selected offset; the noise is looped when shorter than the input clip.
	Codec	One of three channel transformations is applied: MP3 compression (VBR quality 7-9), GSM full-rate coding at 8 kHz, or band-limiting via 24 kHz $\rightarrow$ 8 kHz $\rightarrow$ 24 kHz resampling.

Table 4: Augmentation settings used in our experiments. For each clip, we sample one augmentation uniformly from the selected family. In the "both" condition, one delivery augmentation and one channel augmentation are sampled independently.

A.4 Common Voice Reference-Speaker Pools

We draw one reference clip per speaker from Common Voice (Ardila et al., 2020), keeping the reference-speaker pool separate from the real training corpus. This mirrors the setting in which large-scale real speech for a language is unavailable or cannot be used for pretraining, but a smaller permissively licensed set of reference recordings is available for synthesis. Drawing references from a separate source also prevents the synthetic corpus from directly reusing speakers from the real training data.

Sourcing references this way additionally gives direct control over speaker coverage, since the pool can be subsampled to any size while holding the transcripts fixed. We use this in Section 4.3 to vary the number of distinct reference speakers from 10 up to the full pool. Table 5 reports the pool sizes available for each language; they differ by nearly three orders of magnitude, from 83,057 speakers for English to 107 for Korean, which bounds how far the sweep can be extended in each language.

Language	Reference speakers
English	83,057
German	19,455
Spanish	16,706
French	19,467
Korean	107
Chinese	5,104

Table 5: Number of Common Voice reference speakers used for synthesis.

B Additional Results

This section reports the main experiments in more detail than the main text has room for, including the separate loss axes, additional model scales, and downstream transcription metrics. Unless otherwise stated, reported percentage loss changes use the full real-only model trained at the same token budget as the reference, following Section 3.4. Experiments designed to isolate a narrower intervention use the corresponding controlled baseline: acoustic augmentation is compared with plain synthetic speech, quality filtering with a random drop of the same size, and reference-speaker source with Common Voice references. Each exception is stated explicitly in the corresponding figure or table caption. We use the same 95% confidence-interval criterion as in the main text: an interval containing zero indicates no detected separation, not demonstrated equivalence.

B.1 Real and Synthetic Training Trajectories

Figure 6 shows the descriptive in-training validation trajectories underlying the endpoint comparison in Section 4.1. Each panel compares large-model real-only and Qwen3-TTS synthetic-only training over the same 2B-token schedule. Loss decreases over synthetic-only training in every language and for both components, but the final separation is smallest in several European-language settings and largest for Chinese and Korean. Text and speech exhibit different patterns across languages, consistent with the endpoint decomposition in Appendix B.8. These single-run trajectories characterize optimization behavior; the endpoint comparisons and their confidence intervals remain the basis for our quantitative claims.

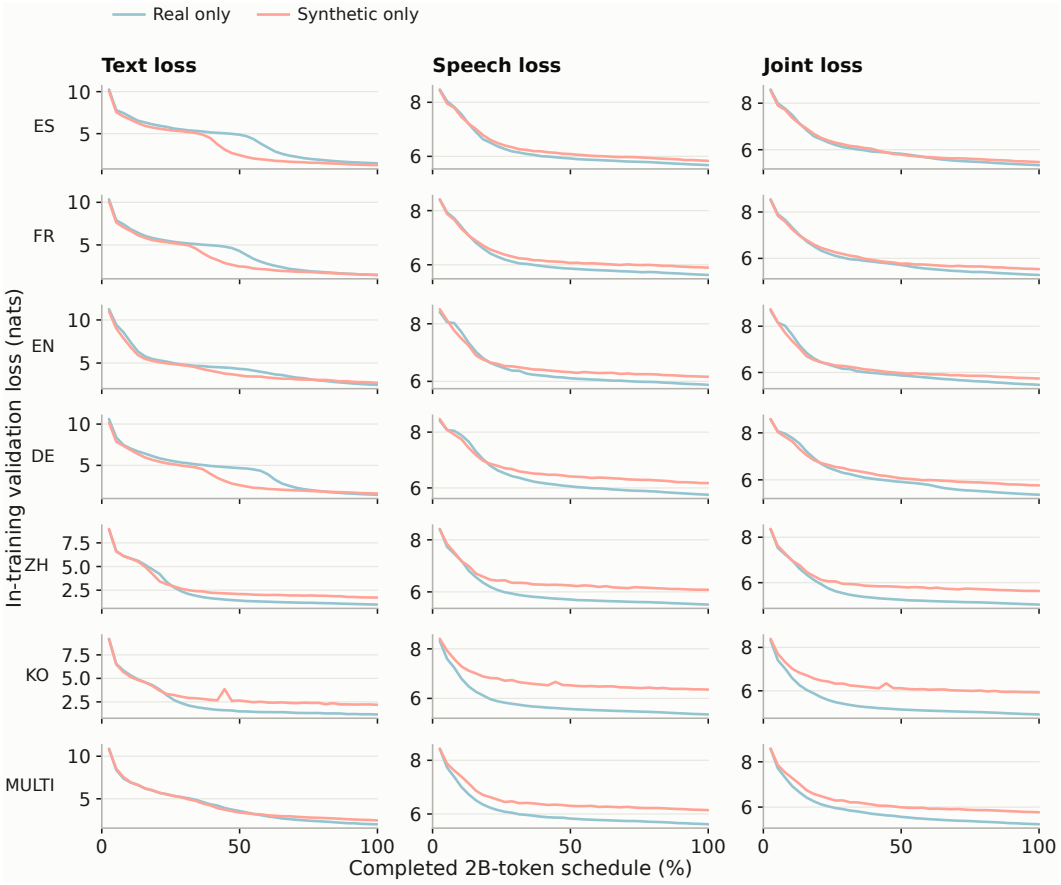


Figure 6: Large-model in-training validation trajectories for real-only and Qwen3-TTS synthetic-only training across all language settings. Each run follows the same 2B-token training schedule. Lower is better.

B.2 CORPUS QUALITY SCORES

We next ask whether conventional corpus-quality measures can explain the synthetic-real gap or its variation across languages. We score every real and synthetic training clip using the same pipeline. Whisper large-v3 is used to measure transcription error against the reference transcript, while UTMOS (Saeki et al., 2022), DNSMOS (Reddy et al., 2021), and NISQA (Mittag et al., 2021) estimate perceptual quality. We additionally report an SNR proxy computed from each clip's frame energies as the decibel ratio between its loudest and quietest frames. Table 6 gives the corpus-level mean for each measure.

		Whisper ASR		Predicted quality				
Lang.	Corpus	WER	CER	UTMOS	NISQA-TTS	NISQA-MOS	DNSMOS	SNR
ES	Real	19.1	17.4	1.97	2.77	3.12	3.10	33.7
	Synthetic	3.8	2.5	2.76	3.12	3.92	3.29	48.3
FR	Real	11.1	9.5	2.06	3.00	3.27	3.13	32.2
	Synthetic	4.9	3.2	2.78	3.29	4.01	3.36	50.0
EN	Real	12.0	9.2	2.55	2.81	3.46	2.92	33.6
	Synthetic	9.9	6.5	3.55	3.13	3.92	3.17	49.1
DE	Real	14.1	12.4	2.05	2.83	3.15	3.02	30.6
	Synthetic	3.7	2.2	3.10	3.21	4.10	3.32	48.6
ZH	Real	14.7	14.7	1.94	2.54	2.69	2.14	23.7
	Synthetic	29.8	29.8	2.84	3.08	3.50	2.94	54.6
KO	Real	37.0	26.0	2.47	3.14	3.47	2.96	44.0
	Synthetic	15.2	4.6	2.93	3.30	3.91	3.11	56.4

Table 6: **Conventional corpus-quality scores do not explain the synthetic-real pretraining gap.** Corpus-level means over all training clips. WER and CER are obtained using Whisper large-v3 against the reference transcript (lower is better); UTMOS, NISQA-TTS, NISQA-MOS, and DNSMOS are non-intrusive quality predictors on a 1–5 scale (higher is better); SNR is reported in dB. Chinese is scored at the character level, so WER and CER coincide. Synthetic speech receives higher predicted-quality and SNR scores in every language and lower transcription error in every language except Chinese, despite synthetic-only pretraining producing higher loss. The cross-language ordering is also inconsistent with these scores: languages with the largest synthetic-real gaps, particularly Chinese and Korean, do not consistently have lower synthetic quality scores.

B.3 MIXTURE-SWEEP SUMMARY

Figure 4 in the main text shows the complete 1.17B joint-loss curves. Table 7 complements that figure by summarizing the best tested synthetic share, the shares for which joint loss differs detectably from real-only training, and the synthetic-only endpoint for both model scales.

B.4 REAL SPEECH DURING LEARNING-RATE DECAY

Section 4.2 shows that placing real speech after synthetic speech is substantially more effective than reversing the order. However, the synthetic→real curriculum allocates half of the total training budget to real speech. We therefore test whether this benefit persists when real speech is reserved only for the final part of training, coinciding with the learning-rate decay phase. This setup is motivated by LLM pretraining recipes that concentrate higher-quality data toward the end of training, when the learning rate is reduced (OLMo et al., 2025; Grattafiori et al., 2024).

All models are trained for 2B tokens using the same learning-rate schedule as in the main experiments. The late-real schedule uses synthetic speech for most of training and switches to real speech only for the final learning-rate decay phase, so that 20% of the training tokens, or 0.4B tokens, are real. As in Section 4.2, the synthetic and real stages use disjoint transcript subsets.

A short real-speech stage at the end of training recovers most of the synthetic-real gap. Figure 7 shows a clear separation based on which source is seen last. Synthetic-only and real→synthetic training remain approximately 2%-20% above the real-only joint-loss baseline, whereas schedules

		Best share						Synthetic only		
Lang.	Model	Share	ΔJoint	ΔAcc@1	ΔWER	Better	Worse	ΔJoint	ΔAcc@1	ΔWER
EN	1.17B/2B	25	-0.04*	+2.5	-20.7*	12.5-37.5	87.5	+0.27*	-3.5*	-19.0*
EN	490M/1B	50	-0.06	+2.0	-47.0	-	-	+0.12*	+5.5	-94.1*
DE	1.17B/2B	62.5	-0.06*	+5.9*	-17.8*	37.5-75	12.5	+0.31*	+1.1	-5.4
DE	490M/1B	75	-0.07	+6.0	-15.0	-	-	+0.21*	+10.1*	-35.3
ES	1.17B/2B	62.5	-0.11*	+6.8*	-40.4*	12.5-87.5	-	+0.12*	+6.1*	-35.4*
ES	490M/1B	75	-0.14*	+2.4	-87.2*	12.5-87.5	-	+0.06*	+2.5	-78.4*
FR	1.17B/2B	62.5	-0.08*	+5.8*	-33.0*	25-87.5	-	+0.24*	+1.1	-22.0*
FR	490M/1B	12.5	-0.07	+0.8	+3.8	-	-	+0.12*	+6.9	-27.2
KO	1.17B/2B	12.5	0.00	-0.3	+1.4	-	62.5-87.5	+1.01*	-20.2*	+42.3*
KO	490M/1B	12.5	-0.01	+0.8	-0.9	-	75-87.5	+0.88*	-14.4*	+20.5
ZH	1.17B/2B	12.5	-0.01	-0.3	+1.7	-	75-87.5	+0.57*	-12.4*	+10.6
ZH	490M/1B	25	-0.05	+2.0	-2.8	-	-	+0.44*	-4.8	-5.0
Multi	1.17B/2B	62.5	-0.03*	+4.4*	-18.3*	37.5, 62.5	-	+0.28*	+0.7	-13.9
Multi	490M/2B	37.5	-0.09*	+5.2	-8.7	25-62.5	-	+0.23*	+3.7	-19.2

Table 7: Mixture-sweep summary. Shares are percentages of synthetic training tokens. "Best" is the tested share with the lowest joint loss, and the adjacent columns give its difference from the matched real-only baseline. "Better" and "Worse" list shares whose 95% confidence interval for the joint-loss difference from real-only training excludes zero in the corresponding direction; a single hyphen indicates none. The final three columns report the synthetic-only endpoint. Figure 4 shows the complete 1.17B joint-loss curves, while this table additionally summarizes the 490M sweep and downstream endpoint metrics.

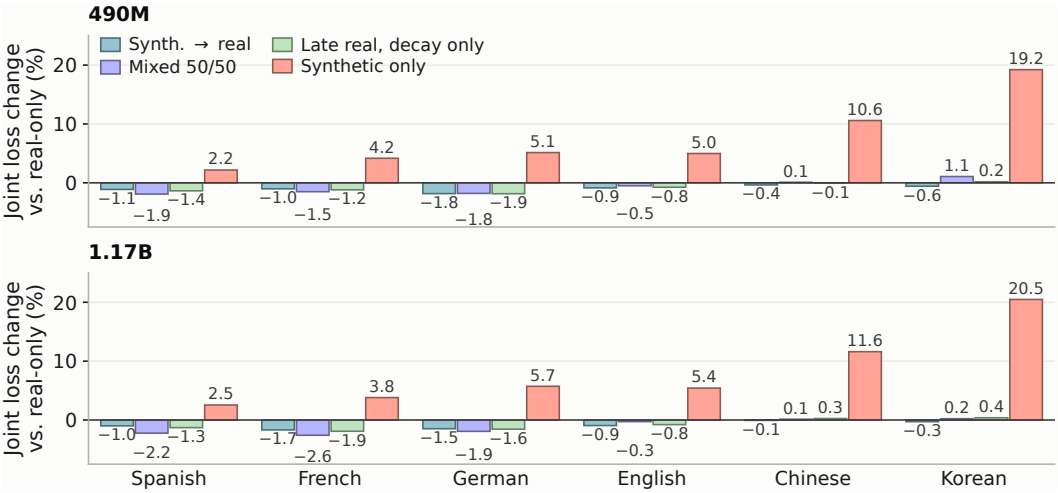


Figure 7: Relative joint-loss change compared with real-only training at a 2B-token budget for different placements of real speech. Lower is better; 0% denotes the real-only baseline. The late-real schedule uses real speech only during the final 20% of training, or 0.4B tokens, compared with 1.0B real tokens for the standard synthetic→real curriculum. Reserving real speech for the end of training recovers most of the benefit of the longer real-speech stage.

that finish on real speech approach or improve upon it. Reducing the real-speech stage from half of training to only the final 20% changes joint loss by at most 0.8 percentage points across the evaluated configurations. Thus, most of the benefit of synthetic→real training does not require allocating half of the training budget to real speech. In the evaluated settings, using real speech only at the end of training reaches nearly the same endpoint while requiring 2.5× fewer real tokens than the 50/50 curriculum and 5× fewer than real-only training.

B.5 Interleaved audio-text training

Since Nguyen et al. (2025) showed that interleaving speech and text during training improves cross-modal transfer, we test whether synthetic speech can serve in this setting as well. Elsewhere in this work, each training example presents an entire recording's audio followed by its full transcript.

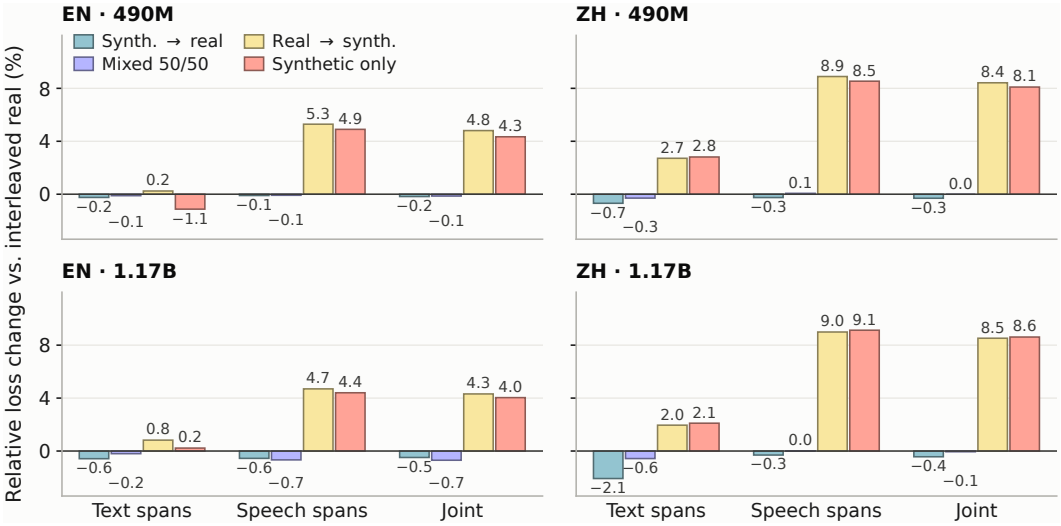


Figure 8: Interleaved training with synthetic speech, evaluated in the interleaved format at a 2B-token budget. Relative loss change against training on real audio throughout; lower is better and 0% denotes parity with the real-audio baseline. Synthetic speech incurs little cost on words presented as text but a substantially larger cost on words presented as audio. Mixing sources across recordings and ending the curriculum on real speech largely close this gap, whereas ending on synthetic speech remains near the synthetic-only result.

Here we instead divide each recording into short alternating spans, so the model sees a few words of audio, then a few words of text, then audio again, with each word appearing in exactly one modality: the model either hears it or reads it, never both.

Segment lengths are drawn from a geometric distribution with mean four words. Using forced alignment (Pratap et al., 2024), we place each cut between consecutive words at the midpoint of the intervening silence, so audio spans end on word boundaries rather than mid-word. Following Nguyen et al. (2025), roughly 60% of recordings are interleaved this way; the remainder are emitted as audio-only or text-only so that both pure modalities remain in training.

Five training corpora share the same segmentation, re-seeded identically across conditions, and differ only in the source of the audio. Real uses the original recording throughout and serves as the baseline. Synthetic only uses the corresponding synthesized clip throughout. Mixed chooses one source per recording by equal chance, so each training sequence is entirely real or entirely synthetic and the two sources appear in roughly equal proportion across the corpus. The two curricula repeat the comparison of Section 4.2 in this format: transcripts are split into two non-overlapping halves, one stage trains on half A and the other on half B, and only the order of the two sources changes, so no sentence is seen with both renderings. Every condition uses the same 2B-token budget, and the Mixed condition and both curricula consume approximately equal amounts of real and synthetic speech. Figure 8 evaluates them on held-out recordings in the same interleaved format.

Synthetic speech is adequate for the text side of interleaved training but not for the audio side. Training on synthetic audio alone leaves text-span prediction within 2.9% of real, and improves upon real for English at 490M, but increases audio-span loss by 4.4%-9.1% in every configuration. The split mirrors the audio-then-transcript setting: predictions over text spans are much less sensitive to the source of the audio than predictions over speech codes.

Mixing sources, or ending on real speech, closes the gap. When half of the recordings use real speech and half use synthetic speech, text, speech, and joint losses all remain within 0.7% of the real-only baseline across both languages and model sizes. The curricula reproduce the ordering effect from Section 4.2: ending on real speech places joint loss 0.2%-0.5% below the real-audio baseline in all four configurations, whereas ending on synthetic speech leaves it 4.3%-8.5% above, close to synthetic-only training. Because the two curricula use identical amounts of each source and differ

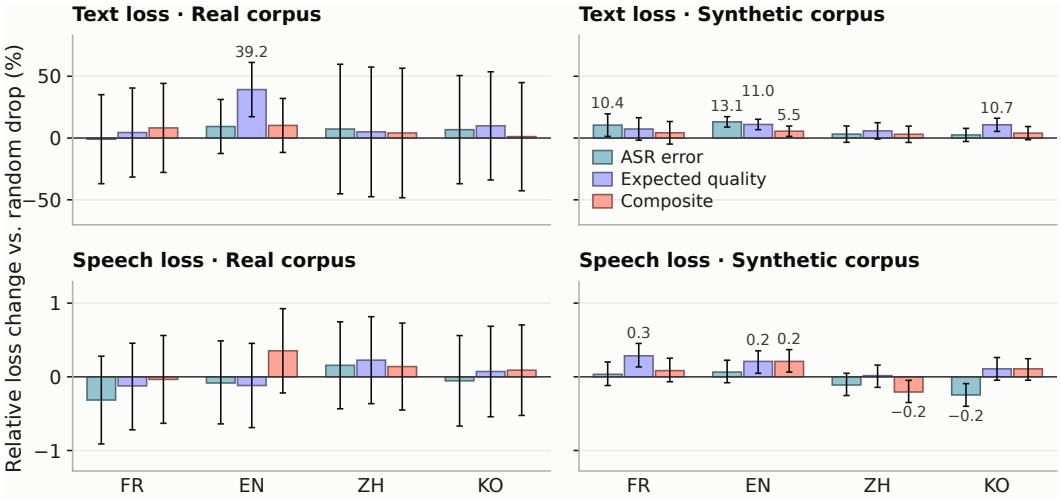


Figure 9: Quality filtering compared with a random drop of the same size, for the 490M model at a 2B-token budget. Positive values indicate that the criterion performs worse than random removal. Error bars are 95% confidence intervals, and values are printed only when the interval excludes zero. No criterion detectably improves over random removal in any setting. On real speech, only expected-quality filtering in English separates from random, worsening loss by 1.7%; on synthetic speech, the criteria separate in five cells and are worse in all five. Under these criteria and this removal rate, quality-based filtering does not improve over random removal.

only in their order, source order is the controlled difference between them. The order-free mixture performs comparably to the better curriculum rather than falling between the two orderings. Thus, both strategies carry over to the interleaved sequence format without requiring more real speech than the other.

B.6 Quality-based filtering

Each criterion removes the worst 20% of clips from a corpus and upsamples the remainder to the same token budget. Because this also removes one fifth of the unique data, the comparison that isolates the effect of which clips are removed is against a random 20% drop rather than the unfiltered corpus.

We transcribe every clip with Whisper (Radford et al., 2022) and rank clips by three criteria. The first is per-clip word error rate against the reference transcript, using character error rate for Chinese where word segmentation is unreliable. The second is expected quality, which penalizes both truncation and over-generation while ignoring substitutions that preserve length. The third is a composite rank-average of the first two together with the deviation of each clip's voice-activity speech ratio from the corpus median.

Table 8 reports the corresponding absolute metrics and changes from the unfiltered corpus. The figure is retained because it directly compares each filtering criterion with random removal using paired confidence intervals.

B.7 Acoustic augmentation on each loss axis

Figure 10 separates the two loss axes combined in Figure 5 (left), since they respond differently to augmentation.

Variation in recording and transmission conditions consistently improves synthetic speech. Figure 10 shows that channel augmentation reduces speech and joint loss across every language, model size, and training budget we evaluate. Across settings, it recovers a median of 32.9% of the synthetic-real joint-loss gap and 35% of the synthetic-real speech-loss gap, although it does not fully

		Real corpus					Synthetic corpus				
Lang.	Criterion	Text	Audio	Joint	Acc@1	WER	Text	Audio	Joint	Acc@1	WER
EN	Full	2.62	5.95	5.56	51.1	73.9	3.01	6.23	5.84	45.8	85.3
	WER filter	+0.26	0.00	+0.03	-3.0	+9.9	+0.37*	+0.01	+0.05*	-2.6*	-9.4
	Expansion	+1.05*	-0.01	+0.10*	-10.7*	+17.4	+0.31*	+0.02*	+0.05*	-2.7*	-4.8
	Composite	+0.29	+0.02	+0.05	-3.5	+14.8	+0.15*	+0.02*	+0.03*	-1.4	-2.2
	Random	+0.02	0.00	0.00	-0.2	+9.1	-0.02	0.00	0.00	+0.5	-1.8
FR	Full	1.62	5.70	5.37	64.8	79.8	1.30	5.95	5.58	70.7	64.7
	WER filter	-0.04	-0.01	-0.02	+0.8	+7.4	+0.21*	+0.01*	+0.03*	-2.9*	-5.0
	Expansion	+0.05	0.00	0.00	-0.9	+0.5	+0.17*	+0.03*	+0.04*	-2.7*	-1.8
	Composite	+0.11	0.00	+0.01	-1.6	+5.5	+0.12	+0.02*	+0.02*	-1.9	-5.7
	Random	-0.02	+0.01	0.00	+0.2	+1.3	+0.06	+0.01*	+0.01	-1.0	+4.5
KO	Full	1.29	5.54	5.07	75.4	45.8	2.34	6.49	6.02	55.9	78.9
	WER filter	+0.12	0.00	+0.01	-1.3	-1.8	+0.08	-0.01	0.00	-0.5	0.0
	Expansion	+0.16	0.00	+0.02	-0.8	-3.3	+0.28*	+0.01*	+0.04*	-1.6	-1.1
	Composite	+0.04	0.00	+0.01	-0.8	-3.1	+0.12	+0.01*	+0.03*	-0.5	-1.8
	Random	+0.03	0.00	0.00	-1.5	+6.2	+0.02	+0.01	+0.01	-0.1	+5.2
ZH	Full	1.12	5.76	5.25	73.6	92.2	1.92	6.29	5.80	62.0	101.5†
	WER filter	+0.05	0.00	+0.01	+0.4	-3.2	+0.04	+0.01	+0.01	-0.1	+1.5
	Expansion	+0.03	+0.01	+0.01	+0.3	-1.7	+0.09	+0.01*	+0.03*	-1.1	-1.3
	Composite	+0.02	0.00	0.00	+0.3	-2.6	+0.04	0.00	0.00	-0.3	-2.1
	Random	-0.03	0.00	0.00	+0.7	+1.3	-0.02	+0.01*	+0.01	0.0	-1.3

Table 8: Quality-based filtering for the 490M model at 2B tokens. Each criterion removes 20% of clips from the real or synthetic corpus. The Full row reports absolute values, and the remaining rows report changes from the unfiltered corpus; * therefore marks full-corpus contrasts, and † marks WER above 100%, which can occur because insertion errors can make WER exceed 100%. Figure 9 gives the paired comparison against a random 20% drop, which isolates whether the filtering criterion selects more useful examples than random removal.

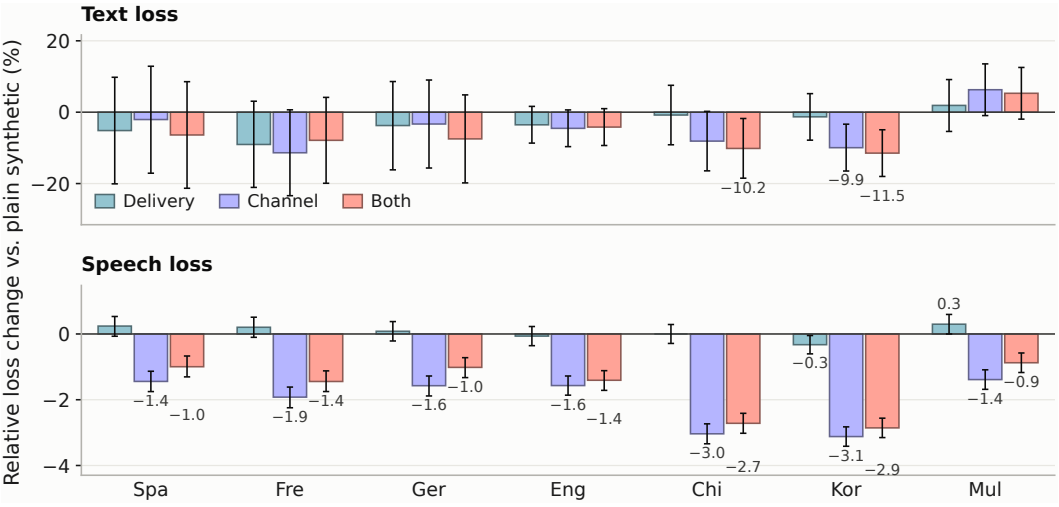


Figure 10: Acoustic augmentation evaluated separately on each loss axis, for the 1.17B model at a 2B-token budget, relative to plain synthetic speech. Error bars are 95% confidence intervals, and values are printed only where the interval excludes zero. Channel augmentation improves speech-token prediction in every language by 1.4%-3.1%, whereas delivery augmentation produces no resolved change. The text axis shows a different pattern: no effect is resolved except in Chinese and Korean, the two languages with the largest synthetic text gaps to begin with. On the speech axis, combining both augmentation families is never better than channel augmentation alone.

close the gap in any setting. The largest recoveries relative to the original gap occur for Spanish and French, where channel augmentation closes approximately 43%-56% of the speech-loss difference.

Changing delivery features provides substantially less benefit. Delivery augmentation does not consistently improve pretraining performance and slightly increases speech loss in several settings. Combining delivery and channel augmentation also provides no additional benefit over channel augmentation alone: channel augmentation has equal or lower speech loss than the combined recipe

			490M					1.17B				
Lang.	Budget	Corpus	Text	Audio	Joint	Acc@1	WER	Text	Audio	Joint	Acc@1	WER
EN	2B	Plain	3.01	6.23	5.84	45.8	85.3	2.83	6.18	5.79	49.1	75.6
		Delivery	-0.11	0.00	-0.01	+2.0	-7.1	-0.10	0.00	-0.02	+1.5	-6.6
		Channel	-0.17*	-0.11*	-0.11*	+2.9*	-9.2	-0.13	-0.10*	-0.10*	+1.7	+2.7
		Both	-0.12	-0.09*	-0.10*	+2.2	-7.0	-0.12	-0.09*	-0.09*	+1.8	-6.2
DE	2B	Plain	1.36	6.17	5.76	71.3	58.4	1.18	6.10	5.68	75.2	51.4
		Delivery	-0.02	+0.01	0.00	+0.2	-3.7	-0.04	0.00	0.00	+0.6	-4.0
		Channel	-0.06	-0.09*	-0.09*	+1.0	-4.3	-0.04	-0.10*	-0.09*	+0.5	+3.5
		Both	-0.04	-0.06*	-0.06*	+0.4	-0.6	-0.09	-0.06*	-0.07*	+1.8	-8.4
ES	2B	Plain	1.17	5.90	5.53	73.6	54.6	0.97	5.82	5.44	78.1	48.9
		Delivery	+0.03	+0.01*	+0.01	-0.3	+1.7	-0.05	+0.01	+0.01	+0.9	-8.5
		Channel	+0.06	-0.08*	-0.07*	-0.7	+1.5	-0.02	-0.08*	-0.08*	+0.3	-5.8
		Both	+0.07	-0.04*	-0.04*	-1.1	+2.1	-0.06	-0.06*	-0.06*	+1.2	-9.7
FR	2B	Plain	1.30	5.95	5.58	70.7	64.7	1.21	5.88	5.51	72.8	52.3
		Delivery	+0.06	+0.02*	+0.02*	-1.1	-3.6	-0.11	+0.01	0.00	+2.0	-2.4
		Channel	+0.08	-0.11*	-0.10*	-1.4	-4.0	-0.14	-0.11*	-0.12*	+2.4	-4.6
		Both	+0.07	-0.08*	-0.07*	-1.5	-3.4	-0.09	-0.08*	-0.09*	+1.7	-2.1
KO	2B	Plain	2.34	6.49	6.02	55.9	78.9	2.23	6.44	5.97	58.1	81.7
		Delivery	-0.05	0.00	-0.01	+1.1	-2.7	-0.03	-0.02*	-0.02*	+0.6	-9.2
		Channel	-0.20*	-0.20*	-0.20*	+3.8*	-9.4	-0.22*	-0.20*	-0.20*	+3.3*	-4.9
		Both	-0.24*	-0.17*	-0.18*	+4.6*	-9.3	-0.26*	-0.18*	-0.19*	+5.0*	-15.1
ZH	2B	Plain	1.92	6.29	5.80	62.0	101.5†	1.75	6.26	5.75	64.8	98.7
		Delivery	-0.06	-0.01*	-0.03*	+1.0	+0.2	-0.01	0.00	-0.01	+0.2	+2.8
		Channel	-0.19*	-0.19*	-0.19*	+2.8*	-6.3	-0.14	-0.19*	-0.18*	+2.2	-2.6
		Both	-0.22*	-0.17*	-0.18*	+3.1*	-3.0	-0.18*	-0.17*	-0.17*	+2.7*	+1.9
Multi	2B	Plain	2.32	6.14	5.80	54.8	77.7	2.00	6.06	5.69	60.2	68.4
		Delivery	+0.07	0.00	0.00	-1.2	+6.5	+0.04	+0.02*	+0.02	-0.8	+5.2
		Channel	+0.17*	-0.10*	-0.08*	-2.6*	+17.3*	+0.13	-0.08*	-0.06*	-2.1	+16.4*
		Both	+0.02	-0.07*	-0.07*	-0.1	+14.1	+0.11	-0.05*	-0.04*	-1.8	+10.2
Multi	6B	Plain	1.58	6.01	5.62	67.6	55.7	1.36	5.94	5.54	72.0	47.8
		Delivery	+0.01	+0.02*	+0.01	-0.2	+1.7	-0.01	+0.01	+0.01	0.0	+0.2
		Channel	-0.03	-0.09*	-0.09*	+0.3	+2.3	-0.06	-0.10*	-0.10*	+0.9	-0.7
		Both	-0.01	-0.07*	-0.07*	-0.1	+1.4	-0.08	-0.07*	-0.07*	+1.1	0.0

Table 9: Acoustic augmentation results. Plain synthetic speech is the baseline of each configuration and is reported in absolute values; the augmented rows report changes from that baseline. * marks contrasts whose 95% confidence interval excludes zero, and † marks WER above 100%, which can occur because insertion errors can make WER exceed 100%. This intervention-specific baseline isolates the effect of each augmentation while keeping the synthetic corpus and training-token budget fixed.

across the evaluated settings. This suggests that the missing variability in synthetic speech is better addressed by modeling differences in recording environments and transmission conditions than by perturbing properties such as speaking rate or pitch.

Multilingual training shows a slightly different pattern. Channel augmentation still improves speech-token prediction, but its text-loss performance degrades, reducing the improvement in joint loss. Overall, these results indicate that broadening the recording and transmission conditions of synthetic speech is a simple and consistent way to increase its pretraining utility, whereas delivery perturbations provide little additional benefit.

Table 9 gives the corresponding absolute metrics and intervention deltas for both model scales, including the multilingual 6B setting. Figure 10 is retained because it makes the different responses of the text and speech loss axes directly visible.

B.8 Where the synthetic-real gap comes from

Table 10 decomposes the synthetic-real comparison across the two loss axes and downstream transcription metrics. The source of the joint-loss gap differs substantially across languages. For German, Spanish, and French, synthetic speech preserves or improves text-token prediction while increasing speech-token loss. English differs from these settings: both losses increase, and its relative text-loss increase is larger than its speech-loss increase. Chinese and Korean show the strongest version of this pattern, with large increases on both axes and particularly large relative changes in text loss. Multilingual training again differs, with a small improvement in text loss but a higher speech-token loss.

The downstream metrics do not reproduce the loss ordering exactly. Korean and Chinese show the largest reductions in transcript-token accuracy, whereas several European-language settings improve

		490M					1.17B				
Lang.	Training data	Text	Audio	Joint	Acc@1	WER	Text	Audio	Joint	Acc@1	WER
EN	Real	2.62	5.95	5.56	51.1	73.9	2.50	5.91	5.52	52.6	94.6
	Synth.	+0.39	+0.28*	+0.29*	-5.3	+11.4	+0.33*	+0.27*	+0.27*	-3.5*	-19.0*
DE	Real	1.55	5.86	5.49	67.1	71.5	1.19	5.77	5.37	74.0	56.8
	Synth.	-0.19	+0.31*	+0.27*	+4.2	-13.1	-0.01	+0.33*	+0.31*	+1.1	-5.4
ES	Real	1.45	5.74	5.40	68.8	58.5	1.28	5.67	5.32	71.9	84.3
	Synth.	-0.28	+0.16*	+0.13*	+4.8	-3.9	-0.31*	+0.15*	+0.12*	+6.1*	-35.4*
FR	Real	1.62	5.70	5.37	64.8	79.8	1.26	5.62	5.27	71.7	74.3
	Synth.	-0.32	+0.25*	+0.20*	+5.9	-15.1	-0.05	+0.26*	+0.24*	+1.1	-22.0*
KO	Real	1.29	5.54	5.07	75.4	45.8	1.15	5.44	4.96	78.3	39.4
	Synth.	+1.05*	+0.95*	+0.95*	-19.5*	+33.1	+1.08*	+1.00*	+1.01*	-20.2*	+42.3*
ZH	Real	1.12	5.76	5.25	73.6	92.2	0.96	5.71	5.18	77.1	88.1
	Synth.	+0.80*	+0.53*	+0.56*	-11.5*	+9.3	+0.79*	+0.55*	+0.57*	-12.4*	+10.6
Multi	Real	2.62	5.86	5.56	51.1	96.9	2.07	5.75	5.42	59.5	82.2
	Synth.	-0.30	+0.28*	+0.24*	+3.7	-19.2	-0.07	+0.31*	+0.28*	+0.7	-13.8

Table 10: Real versus synthetic speech after 2B training tokens. For each language, the real-only row reports absolute values and the synthetic row reports the difference from that baseline. Lower is better for text, audio, and joint loss and for WER; higher is better for Acc@1. Green indicates improvement over real-only training and red indicates degradation, with tint scaled within each metric. * marks contrasts whose 95% confidence interval excludes zero. The contribution of text and audio loss to the synthetic-real gap differs substantially across languages.

despite higher joint loss. Thus, there is no single loss component that explains the synthetic-real gap across languages: speech-token modeling accounts for most of the gap in German, Spanish, and French, while English, Chinese, and Korean also show degradation in transcript-token prediction. The interleaved experiment in Appendix B.5 further shows that this text-side degradation is not reproduced when text tokens receive more textual context within the sequence.

B.9 Reference-speaker diversity

The main text reports the relative gap to real-only training as the reference-speaker pool grows. Table 11 gives the corresponding absolute losses and downstream metrics. Pools are nested subsets over identical transcripts, so increasing the pool changes speaker coverage without changing the underlying linguistic content. Where the sweep reaches 10,000 speakers, the largest gains occur before that point, while speech loss can continue to improve modestly as the pool expands further.

B.10 Multilingual training, language by language

The multilingual results in the main text are reported as a single aggregate over the union of the six evaluation sets, which can obscure language-specific behavior. Figure 11 evaluates the same models separately on each language.

Table 12 reports the same multilingual slices numerically, including the 490M 1B setting. The figure is retained because the reduction in the synthetic-real gap under multilingual training is easier to compare across languages visually.

B.11 Downstream transcription

Beyond the losses reported in Section 3.4, we evaluate the transcripts generated by the model itself. We use two complementary views: under teacher forcing, we compute top-1 accuracy over transcript tokens; under greedy decoding through the end of each utterance, we compute word and character error rate against the gold transcript. Neither evaluation relies on an external recognizer: the corpus transcript provides the reference, and the model's own output provides the hypothesis.

Lang.	Model	Speakers	Text	Audio	Joint	Acc@1	WER
EN	490M/2B	10	3.90	6.70	6.35	35.5	97.0
		100	-0.45	-0.24	-0.26	+4.1	-7.0
		1,000	-0.75	-0.34	-0.38	+8.6	-17.2
		10,000	-0.91	-0.46	-0.50	+11.2	-22.4
		83,057	-0.90	-0.47	-0.51	+10.2	-11.7
FR	490M/2B	10	2.12	6.34	6.00	56.5	83.3
		100	-0.12	-0.16	-0.15	+2.9	-13.0
		1,000	-0.59	-0.25	-0.28	+10.2	-18.0
		10,000	-0.71	-0.38	-0.40	+12.2	-21.0
		19,467	-0.82	-0.39	-0.42	+14.2	-18.6
FR	1.17B/2B	10	1.85	6.29	5.93	61.3	85.3
		100	-0.14	-0.19	-0.18	+3.0	-19.0
		1,000	-0.50	-0.26	-0.28	+8.8	-26.1
		10,000	-0.70	-0.39	-0.41	+12.4	-33.7
		19,467	-0.64	-0.40	-0.42	+11.6	-33.0
KO	490M/2B	10	2.94	6.75	6.31	47.8	87.2
		107	-0.60	-0.26	-0.29	+8.2	-8.3
ZH	490M/2B	10	2.94	6.87	6.40	46.1	111.0†
		5,104	-1.02	-0.57	-0.60	+15.9	-9.5
ZH	1.17B/2B	10	2.70	6.81	6.33	49.2	113.6†
		5,104	-0.95	-0.55	-0.58	+15.5	-14.9

Table 11: Reference-speaker diversity at 2B tokens. The smallest pool of each configuration gives absolute values and larger pools give the difference from that row; pools are nested subsets over identical transcripts. The main-text figure reports the same conditions relative to the matched full real-only baseline. † marks WER above 100%, which can occur because insertion errors can make WER exceed 100%.

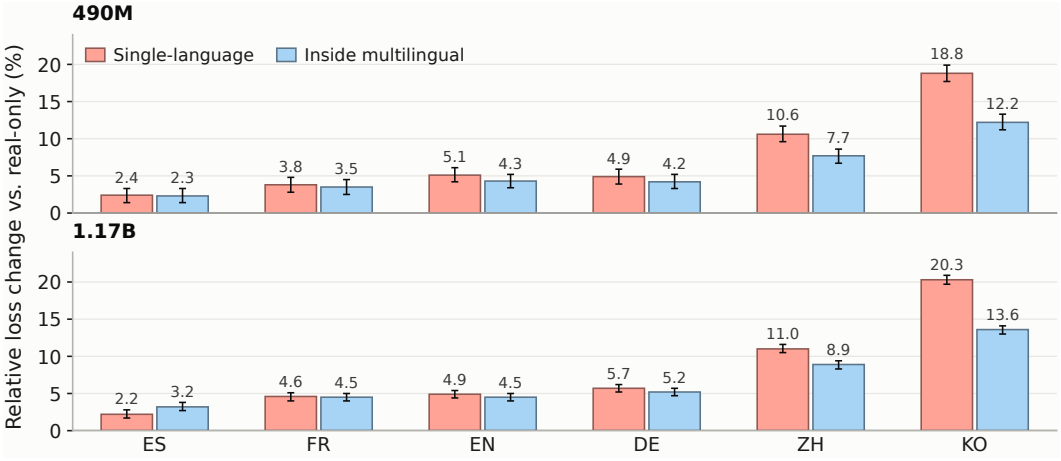


Figure 11: Synthetic-real penalty for each language when trained independently and when trained as one sixth of a multilingual model, each relative to its own real-only baseline at a 2B-token budget. Error bars are 95% confidence intervals. The penalty is smaller under multilingual training in most language-scale settings, with the largest reductions for Korean and Chinese, which also have the largest monolingual penalties: at 490M, Korean falls from 18.8% to 12.2% and Chinese from 10.6% to 7.7%. Because the multilingual model allocates only one sixth of its training tokens to each language, this comparison characterizes how the synthetic penalty changes across training regimes rather than matching the per-language training budget.

We use top-1 accuracy as the primary downstream measure. After 2B pretraining tokens, greedy decoding remains close to degenerate. WER is typically between 40% and 100% and occasionally exceeds 100%. As the intervals stay roughly five times wider than those for token accuracy, it resolves only about half as many contrasts. Figure 12 shows synthetic-only versus real-only pretraining in the first panel and then reports the analogous headline contrast for the remaining experiments; the corresponding error-rate results for every panel are included in the figure data.

Slice	490M / 1B	490M / 2B	1.17B / 2B	vs. single-language
English	$+4.1$ [$+3.2$, $+5.0$]	$+4.3$ [$+3.4$, $+5.2$]	$+4.5$ [$+4.0$, $+5.0$]	-0.8
German	$+4.0$ [$+3.1$, $+4.9$]	$+4.2$ [$+3.3$, $+5.2$]	$+5.2$ [$+4.7$, $+5.7$]	-0.7
Spanish	$+3.1$ [$+2.2$, $+4.0$]	$+2.3$ [$+1.4$, $+3.3$]	$+3.2$ [$+2.7$, $+3.8$]	-0.0
French	$+3.4$ [$+2.5$, $+4.3$]	$+3.5$ [$+2.5$, $+4.5$]	$+4.5$ [$+4.0$, $+5.0$]	-0.3
Korean	$+9.7$ [$+8.8$, $+10.7$]	$+12.2$ [$+11.2$, $+13.3$]	$+13.6$ [$+13.0$, $+14.1$]	-6.6
Chinese	$+7.0$ [$+6.1$, $+7.9$]	$+7.7$ [$+6.7$, $+8.6$]	$+8.9$ [$+8.3$, $+9.4$]	-3.0

Table 12: Synthetic-real joint-loss gap inside multilingual training, evaluated separately by language. Each value is relative to the matched multilingual real-only model, with 95% confidence intervals. The final column subtracts the corresponding single-language gap from Table 10 at 490M and 2B tokens; negative values therefore indicate a smaller synthetic-real gap under multilingual training.

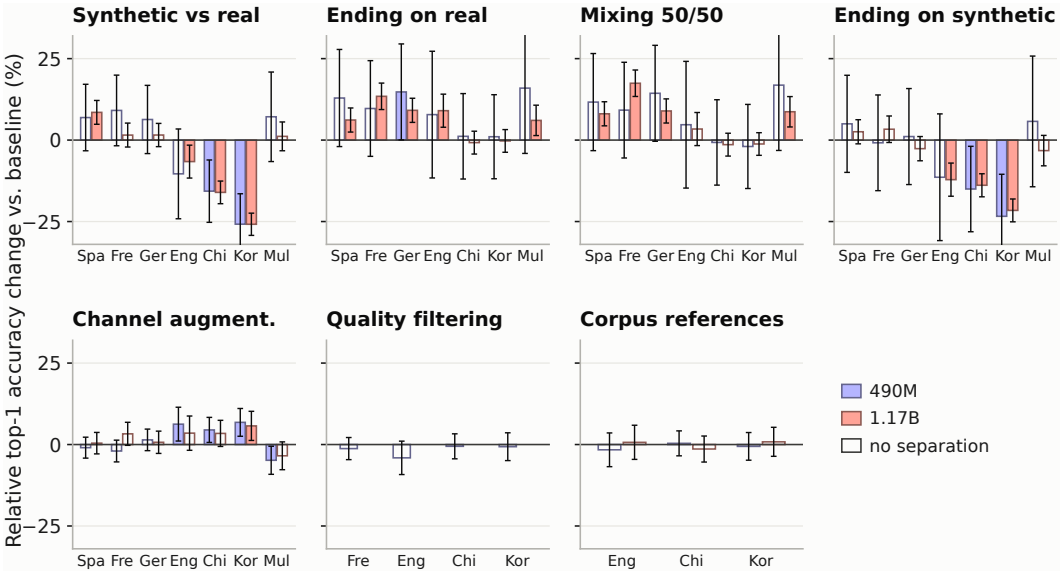


Figure 12: One headline contrast per experiment, reported as the relative change in transcript-token accuracy against that contrast's own baseline at a 2B-token budget. Bars show the 490M and 1.17B models; error bars are 95% confidence intervals. A solid bar indicates that the interval excludes zero, while an open bar indicates that no separation was detected, following the convention of Figure 4. Positive values favor the intervention. The contrasts are: synthetic-only minus real-only; each of the three schedules minus the full matched real-only baseline; channel augmentation minus plain synthetic speech; composite quality filtering minus a random drop of the same size; and real-corpus reference speakers minus Common Voice references. Twenty-six of the eighty contrasts separate from zero. The three schedules reproduce the loss result: ending on real speech and mixing improve accuracy by up to 17% in the European languages, whereas ending on synthetic speech reduces it by 12%-23% in English, Chinese, and Korean. Quality filtering and the reference-speaker source separate nowhere at either scale.

Transcript accuracy does not simply reproduce the loss ordering. Real speech yields lower joint loss in every language (Figure 1), yet synthetic-only pretraining produces higher transcript accuracy for Spanish, French, German, and multilingual training; for Spanish at 1.17B, the gain is 8.5%. The largest downstream degradations occur in the same two languages with the largest loss gaps: Korean loses 26% of its accuracy and Chinese 16% at both scales. They are also the only languages that separate at 490M; all other detected differences appear at 1.17B.

For the intervention comparisons beyond synthetic-only versus real-only, transcript accuracy generally agrees with the loss results whenever it detects a difference. The sole exception is multilingual channel augmentation: it improves speech loss but lowers transcript accuracy by 4.8% at 490M. The agreement also appears in the null results, since quality filtering and the reference-speaker source

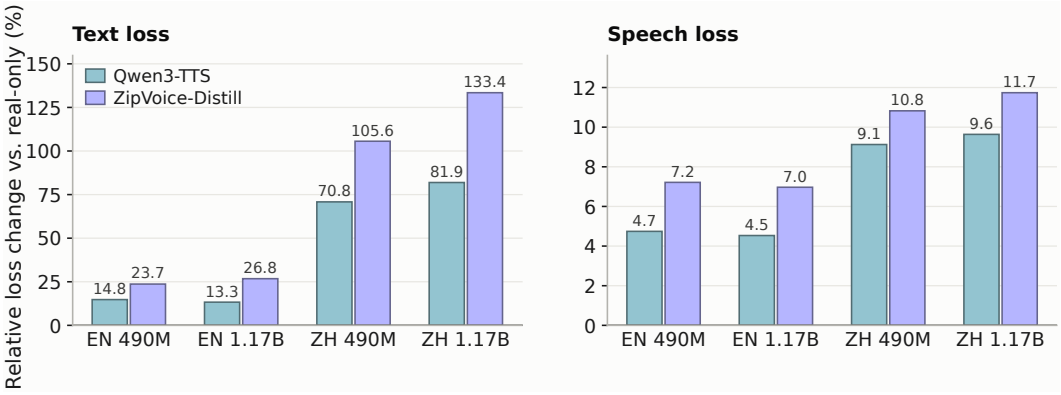


Figure 13: Relative changes (%) in text and speech loss for two TTS systems against real-only training at 2B tokens. Lower is better; 0% denotes parity with real-only training. Bars show 95% confidence intervals. Qwen3-TTS is closer to real speech than ZipVoice-Distill in every setting and on both axes, and the ordering holds at both model sizes. The speech-side cost of either synthesizer is around twice as large in Chinese as in English, while the text-side cost is an order of magnitude larger.

produce no detectable change in either transcript accuracy or loss. Greedy decoding is considerably less discriminative: among the 470 word-error-rate contrasts computed across all experiments, only 81 separate from zero, and the mixture sweep alone accounts for 36 of them through repeated comparisons of neighboring shares against a common baseline. We therefore treat error rates as a coarse sanity check and use top-1 accuracy as the more discriminative downstream measure.

C Controls

C.1 Replication with a Smaller TTS System

We test whether the synthetic-speech result is specific to Qwen3-TTS by repeating the 2B-token experiment with the smaller ZipVoice-Distill synthesizer Zhu et al. (2025), in English and Chinese at both model sizes. The two synthetic corpora use the same transcript rows and the same Common Voice reference speaker assigned to each row, so only the synthesis system changes.

Qwen3-TTS is closer to real-only training in all four configurations, and the paired confidence intervals separate the two synthesizers on text, speech, and joint loss. Measured against the full real-only baseline of each configuration, the joint loss of ZipVoice-Distill sits 2.6 and 2.8 percentage points above Qwen3-TTS for English at 490M and 1.17B, and 2.2 and 2.9 points above it for Chinese. The qualitative synthetic-real gap therefore persists with a second TTS system, while its magnitude depends on the synthesizer. Whether a faster, lower-quality synthesizer can compensate for a larger per-token gap by generating more unique training data is not tested here.