TTL: Language-balanced Data Selection for LLM-based Text-to-Speech via Codec Token Scoring

Anonymous ACL submission

Abstract

Recent progress in large language model (LLM)-based text-to-speech (TTS) has enabled increasingly natural and versatile speech generation, including zero-shot voice synthesis and multilingual capability. However, these gains are often accompanied by a growing reliance on large-scale data, making training resource-intensive, costly, and time-consuming. In this paper, we introduce TTL, a novel data selection framework tailored for LLM-based TTS that leverages a probability gap computed on codec tokens. Our selection uses two scoring models of different sizes, each aggregating conditional probabilities of global and semantic tokens given the text. A larger score gap indicates more challenging and informative speech samples, which we prioritize to construct data-efficient training subsets. On top of that, we further show that naïve selection can severely bias the selected subset toward a high-resource language in bilingual settings, and propose a language-balanced strategy to mitigate this bias and effectively improve bilingual generalization. Extensive experiments on multiple benchmarks demonstrate consistent improvements in naturalness and intelligibility, with substantial gains in word error rate (WER) and strong bilingual performance under data-constrained settings. Source codes are available at https://anonymous.4open.science/r/TTL-F487.

1 Introduction

Text-to-speech (TTS) technology has made remarkable advancements in recent years, expanding its functionality and improving overall user experience. In particular, current models (Ju et al., 2024; Anastassiou et al., 2024; Chen et al., 2025) deliver increasingly natural speech and support zero-shot voice synthesis, often extending to multilingual generation. At the forefront of this progress is the incorporation of large language models (LLMs) into the speech synthesis pipeline. In many LLM-based TTS frameworks (Wang et al., 2025b; Ye et al., 2025; Wang et al., 2025a; Du et al., 2024), speech is represented as discrete token sequences, typically comprising semantic tokens for linguistic content and acoustic tokens for fine-grained speech details. Conditioned on text, the LLM generates these tokens, and a neural codec decoder reconstructs waveforms from them.

Though landmark advances in LLM-based TTS have highlighted the tremendous potential of modern architectures for speech synthesis, their success increasingly relies on large-scale data. Such scaling requires massive computational resources and makes reproduction costly and time-consuming, motivating the need for data-efficient training. However, existing approaches have typically focused on architectural design, tokenization schemes, and inference strategies, yet relatively less attention has been paid to how training data should be selected or prioritized for LLM-based TTS. From a data-centric perspective, improving the information density of training data can achieve comparable results with lower data demands, thereby reducing the computational burden. Despite its practical importance, this direction remains largely underexplored in the LLM-based TTS literature.

Motivated by this, we propose **T**TS data selection based on codec **T**oken scoring with **L**anguage balancing (TTL), a novel framework tailored for LLM-based TTS that leverages probability gaps on semantic and global tokens. Our selection criterion ranks each sample using the score gap between two scoring models of different sizes, where the smaller model matches the size of the target LLM-based TTS model we ultimately train. The score is computed by aggregating token-wise conditional probabilities of global and semantic tokens conditioned on the text, and the difference between the two aggregated scores is used for ranking. The

proposed framework prioritizes samples with large gaps, where the larger scoring model is confident but the smaller model remains uncertain on the global and semantic tokens, yielding more challenging and informative speech data. On top of that, we additionally highlight the language imbalance issue in recent bilingual LLM-based TTS training settings, where a specific language often dominates the training corpus and biases the selected subset toward the high-resource language. However, this issue remains largely unexplored in both the TTS and data selection literature. To address this issue, we newly introduce an adaptive language-balanced selection strategy, which identifies informative tail regions for each language using the inflection point of the difference score density. The proposed framework not only consistently improves bilingual TTS performance, but also enhances cross-language coverage without heuristic hyperparameters.

Our experimental results demonstrate that the proposed selection approach consistently selects more informative training subsets. In a challenging data-constrained setting, TTL improves both naturalness and intelligibility across multiple benchmarks, including Seed-TTS evaluations (en, zh, hard), LibriTTS, VCTK, and AISHELL-3, yielding substantial reductions in word error rate (WER) and consistent gains in bilingual settings. In addition, we provide an in-depth analysis of language balancing effects and further characterize the differences in acoustic and linguistic feature distributions across training subsets selected by each method.

Our contributions are summarized as follows:

- We introduce TTL, a novel data selection framework for LLM-based TTS that leverages probability differences on semantic and global tokens. *To the best of our knowledge, this is the first attempt at data selection tailored for LLM-based TTS.*

- To address the language imbalance issue in LLM-based TTS, we propose an adaptive language-balanced strategy that effectively improves bilingual performance without heuristic hyperparameters.

- Extensive experiments on multiple benchmarks show that TTL achieves consistent improvements under data-constrained settings, providing a practical path toward more efficient LLM-based TTS training.

2 Related Work

2.1 LLM-based TTS

Recent advances in speech tokenization bridge continuous waveforms and discrete token-based LLMs by quantizing speech into codec tokens. Many zero-shot TTS systems therefore adopt a codec-based LLM framework, where an LLM predicts speech token sequences and a codec decoder reconstructs the final waveform. For example, LLaSA (Ye et al., 2025) extended a LLaMA (Touvron et al., 2023) backbone with an expanded vocabulary that incorporates speech tokens and trains the model via next-token prediction. Spark-TTS (Wang et al., 2025a) introduced BiCodec, a single-stream codec that decomposes speech into low-bitrate semantic tokens for linguistic content and fixed-length global tokens for speaker attributes, utilizing Qwen2.5 (Yang et al., 2025).

2.2 Data Selection

Data selection is the process of choosing an informative subset from a large dataset to make training more efficient. Recently, several selection techniques have been proposed in the LLM literature to enable data-efficient training (Xie et al., 2023; Zhang et al., 2025; Gu et al., 2025). Quad (Zhang et al., 2025) balanced data quality and diversity by iteratively selecting samples via a multi-armed bandit over clustered data. MiniPLM (Gu et al., 2025) introduced an offline knowledge distillation (KD)-based data selection framework that uses difference sampling. Though the proposed approach also adopts the KD-based scoring framework, it differs from MiniPLM in that it is explicitly tailored to LLM-based TTS. The scoring of TTL operates on codec tokens, focusing on semantic tokens for linguistic content and global tokens for speaker attributes. Additionally, it incorporates language balancing for bilingual training, yielding consistent gains across languages. In speech synthesis, Kurematsu et al. (1989) proposed phoneme-balanced corpus construction to alleviate sparsity for rare phonemes. Seki et al. (2024) introduced diversity-based core-set selection by using joint linguistic, speaker, and acoustic features, although it relies on an additional external dataset for feature extraction.

3 Proposed Method

In this section, we introduce TTL, a novel data selection framework designed for LLM-based TTS, and conduct experiments in a bilingual setting.

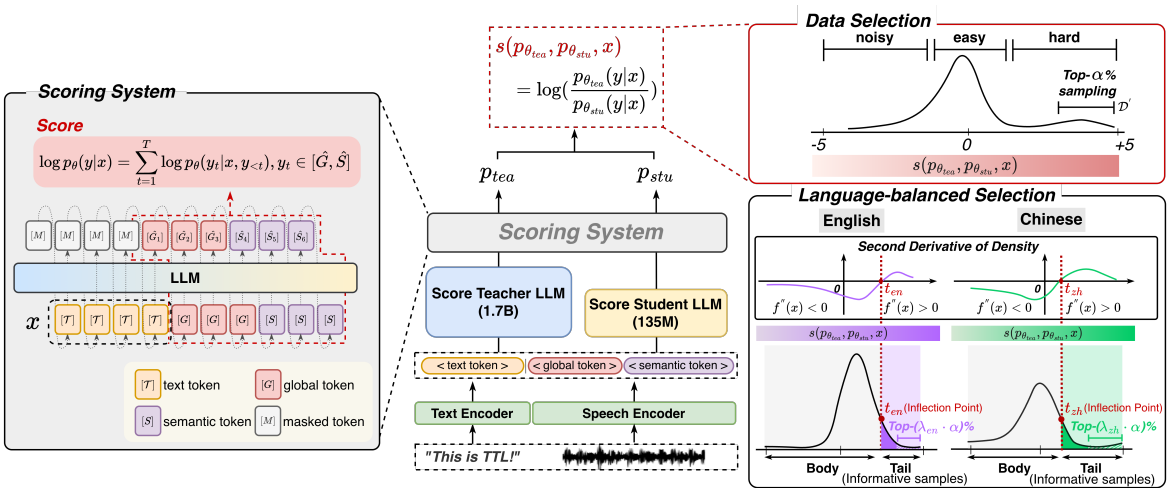

Figure 1: Overview of TTL. We utilize two scoring models of different sizes to compute the probability gap based on global and semantic tokens. This metric prioritizes informative samples where the teacher is confident while the student remains uncertain. For each language $l \in \{en, zh\}$, we identify the inflection point t_l of its score density to define the tail region, and set the allocation ratio λ_l to the normalized proportion of tail samples. We then select $\lambda_l \cdot \alpha \cdot |\mathcal{D}|$ samples from each language. Consequently, the final subset $\mathcal{D}'$ is constructed by aggregating the top-ranked samples independently selected from each language.

3.1 Codec Token Scoring

Scoring Model. As illustrated in Fig. 1, we adopt two scoring models of different sizes and rank each training sample by the probability gap computed on codec tokens, with semantic tokens capturing linguistic content and global tokens encoding speaker attributes. For ease of understanding, we refer to the larger model as the teacher and the smaller model as the student scoring model. The student's size is set to that of the target LLM-based TTS model we ultimately train and evaluate. We prioritize samples with large gaps, where the teacher is confident but the student is uncertain on these tokens, thereby selecting more challenging and informative speech data to train the target TTS model. Since training both teacher and student scoring models on the full dataset is highly costly, we train them on a small fraction (about 2%) of the data for a small number of epochs.

Score Computation. To calculate the score, we construct the input by concatenating the text token with the global tokens and semantic tokens. For each model $\theta \in \{\theta_{tea}, \theta_{stu}\}$, the score is defined as the sum of the token-wise conditional log-probabilities of global and semantic tokens conditioned on the text, which can be calculated as

$$\log p_\theta(y|x) = \sum_{t=1}^{T} \log p_\theta(y_t|x, y_{<t}) \qquad (1)$$

where x is text and $y = [\hat{G}, \hat{S}]$ represents the generated global and semantic tokens. Finally, the difference score $s(p_{\theta_{tea}}, p_{\theta_{stu}}, x)$ for each sample is computed as the gap between the teacher and student scores, which is given by

$$s(p_{\theta_{tea}}, p_{\theta_{stu}}, x) = \log\left(\frac{p_{\theta_{tea}}(y|x)}{p_{\theta_{stu}}(y|x)}\right). \qquad (2)$$

We interpreted $s(p_{\theta_{tea}}, p_{\theta_{stu}}, x)$ by categorizing it into three parts. First, for samples where $p_{\theta_{tea}} \gg p_{\theta_{stu}}$, they are more likely to be selected as they represent informative data for the target TTS model. Second, when $p_{\theta_{tea}} \simeq p_{\theta_{stu}}$, such samples are less likely to be selected, regarding them as already well predicted by smaller model. Lastly, under the assumption that the teacher scoring model provides a more accurate prediction, we discard samples with $p_{\theta_{tea}} \ll p_{\theta_{stu}}$ as potentially noisy, thereby improving data quality.

3.2 Language-balanced Selection

Motivation. Recent LLM-based TTS and neural codec systems are typically trained in bilingual Chinese-English settings (Wang et al., 2025b,a; Xie et al., 2025; Du et al., 2024). However, these training corpora are imbalanced toward English, making English the dominant language in both data scale and acoustic diversity. Under such conditions, we observe that conventional data selection approaches introduce severe selection bias toward high-resource English samples, consequently

degrading performance on Chinese benchmarks. Motivated by these observations, we introduce a language-balanced data selection strategy that effectively preserves multilingual diversity and improves bilingual performance.

Inflection Point-guided Tail Estimation. By ranking samples based on the difference scores in Eq. 2 over the full dataset $\mathcal{D}_l$ for each language $l \in \{en, zh\}$, we can obtain a language-wise score distribution. As aforementioned, samples with larger difference scores are considered more informative and tend to reside in the upper tail of this distribution. To identify the informative tail region, we determine the inflection point t_l that separates the body and tail of the distribution for each language. Specifically, we convert the difference score distribution into a smooth density using kernel density estimation (KDE) and identify the inflection point t_l where the second derivative of the density changes from negative to positive.

Given the inflection point t_l, λ_l is defined as the normalized proportion of samples whose difference scores exceed t_l for language l, which is as follows:

$$\lambda_l = \frac{|\{s(p_{\theta_{tea}}, p_{\theta_{stu}}, x) > t_l \mid x \in \mathcal{D}_l\}|}{\sum_l |\{s(p_{\theta_{tea}}, p_{\theta_{stu}}, x) > t_l \mid x \in \mathcal{D}_l\}|}. \tag{3}$$

The proposed method enables language balancing without heuristic hyperparameters by deriving t_l directly from the difference score distribution. Consequently, λ_l is determined automatically from the data, enabling adaptive selection of informative samples across languages.

Language-balanced Subset Construction. For each language $l \in \{en, zh\}$, a language-specific subset $\mathcal{D}'_l$ is constructed by selecting the top $\lambda_l \cdot \alpha \cdot |\mathcal{D}_l|$ samples based on the difference scores, which is defined as

$$\mathcal{D}'_l = \text{top-}(\lambda_l \alpha)\% \left\{ s(p_{\theta_{\text{tea}}}, p_{\theta_{\text{stu}}}, x) \;\middle|\; x \in \mathcal{D}_l \right\} \tag{4}$$

where $\mathcal{D}_l$ denotes the full dataset for language l. The final subset is then obtained by merging the language-specific subsets, which is given by

$$\mathcal{D}' = \bigcup_{l \in \{en, zh\}} \mathcal{D}'_l. \tag{5}$$

The target LLM-based TTS model is then trained on $\mathcal{D}'$ to generate semantic and global tokens, where $\mathcal{D}'$ consists of high-scoring and informative samples selected for each language.

Language	Duration (hr)	# of Samples
English	8,998.93	7,710,721
Chinese	1,099.21	887,685

Table 1: Statistics of the bilingual corpus used in our experiments.

4 Experimental Setup

Model. We used Spark-TTS (Wang et al., 2025a) as the baseline LLM-based TTS framework and adopted SmolLM2 (Allal et al., 2025) as its LLM backbone. We used the 1.7B SmolLM2 model as the teacher θ_{tea} and 135M as the student scoring model θ_{stu}. For each selection method, we trained a Spark-TTS model with a 135M SmolLM2 backbone, using the same architecture.

Datasets. We conducted experiments on bilingual datasets comprising English and Chinese. As shown in Table 1, the training corpus totaled approximately 10,000 hours, with a detailed list provided in the appendix. For evaluation, we used benchmarks in both languages: LibriTTS (Zen et al., 2019) and VCTK (Yamagishi et al., 2019) for English, and AISHELL-3 (Shi et al., 2021) for Chinese. In addition, we adopted the Seed-TTS (Anastassiou et al., 2024) test sets, including test-en for English and test-zh and test-hard for Chinese.

Conventional Methods. Given the lack of data selection methods specifically designed for LLM-based TTS, we compared against the following model-agnostic selection approaches:

- Random sampling: Samples were selected uniformly at random.
- Phoneme-based sampling (Kurematsu et al., 1989): A text-centric strategy that prioritized samples to maximize phoneme coverage and diversity.
- MiniPLM (Gu et al., 2025): A text-centric data selection method that uses the teacher–student log-likelihood gap based on text inputs, originally proposed for LLM pre-training.

Note that, for a fair comparison, we selected competing baselines that do not rely on external models used to extract features or scores, nor on additional external databases. Additionally, we trained the LLM-based TTS model on subsets selected by each method with the same sampling rate α.

Method	*SEED-TTS test-en*				*SEED-TTS test-zh*			
	UTMOS(↑)	MOS(↑)	WER(↓)	SECS(↑)	UTMOS(↑)	MOS(↑)	CER(↓)	SECS(↑)
Ground Truth	2.809	3.855 ± 0.656	2.14	0.870	2.988	3.895 ± 0.569	1.254	0.852
Random	2.750	3.705 ± 0.519	19.370	0.809	2.663	2.705 ± 0.793	83.802	0.922
Phoneme	2.665	3.637 ± 0.553	21.024	0.770	2.772	2.274 ± 0.943	30.666	0.915
MiniPLM	2.767	$\mathbf{3.870 \pm 0.644}$	**8.553**	0.785	2.779	3.385 ± 0.776	39.372	0.921
Ours, TTL	**2.788**	3.735 ± 0.535	13.902	**0.810**	**2.785**	$\mathbf{3.535 \pm 0.859}$	**22.287**	**0.928**
Method	***SEED-TTS test-hard***				***LibriTTS***			
	UTMOS(↑)	MOS(↑)	CER(↓)	SECS(↑)	UTMOS(↑)	MOS(↑)	WER(↓)	SECS(↑)
Ground Truth	3.016	4.200 ± 0.126	49.509	0.851	3.180	4.050 ± 0.429	6.123	1.000
Random	2.596	2.730 ± 0.693	89.086	**0.922**	**3.155**	3.745 ± 0.394	23.754	0.825
Phoneme	2.629	3.330 ± 0.841	40.311	0.903	2.908	3.585 ± 0.605	40.990	0.791
MiniPLM	2.673	3.388 ± 0.734	61.325	0.916	3.128	3.750 ± 0.489	16.717	0.803
Ours, TTL	**2.679**	$\mathbf{3.615 \pm 0.594}$	**36.702**	0.920	3.147	$\mathbf{3.905 \pm 0.421}$	**16.705**	**0.826**
Method	***VCTK***				***AISHELL-3***			
	UTMOS(↑)	MOS(↑)	WER(↓)	SECS(↑)	UTMOS(↑)	MOS(↑)	CER(↓)	SECS(↑)
Ground Truth	3.722	4.250 ± 0.444	1.992	1.000	2.751	3.750 ± 0.786	3.931	1.000
Random	3.309	3.740 ± 0.277	8.807	0.753	2.827	3.735 ± 0.556	68.727	0.911
Phoneme	2.892	3.285 ± 0.749	25.90	0.723	2.281	2.778 ± 0.850	> 100	0.849
MiniPLM	3.434	3.905 ± 0.245	**3.701**	0.744	2.862	3.986 ± 0.472	26.148	0.917
Ours, TTL	**3.435**	$\mathbf{3.990 \pm 0.255}$	5.802	**0.756**	**2.904**	$\mathbf{4.028 \pm 0.448}$	**9.456**	**0.926**

Table 2: Objective (UTMOS, WER, SECS) and subjective (MOS) results on Seed-TTS (test-en/test-zh/test-hard) and LibriTTS/VCTK/AISHELL-3 with $\alpha = 6.25\%$. Seed-TTS uses English for test-en and Chinese for test-zh and test-hard. LibriTTS and VCTK are English benchmarks, while AISHELL-3 is a Chinese benchmark. > 100 indicates that WER/CER exceeded 100%, meaning the ASR transcription fails. Bold indicates the best result.

Evaluation Metrics. Objective evaluations were performed using the following metrics:

- Word error rate (WER, %) and character error rate (CER, %): To measure pronunciation accuracy and intelligibility, we computed WER for English and CER for Chinese. Transcriptions were obtained using Whisper large-v3 (Radford et al., 2023).
- Speaker embedding cosine similarity (SECS): This metric quantified speaker similarity by computing the cosine similarity between embeddings extracted from the synthesized speech and the reference audio. We used WavLM (Chen et al., 2022) as the speaker encoder.
- UTMOS: We used UTMOSv2 (Baba et al., 2024) to predict the perceptual quality of synthesized speech via a neural MOS prediction network, serving as a proxy for objective naturalness.

Subjective evaluations were conducted via human listening tests on a 5 point scale (1: bad, 5: excellent) using the following metric:

- Mean opinion score (MOS): It evaluated the naturalness and perceptual quality of the synthesized speech.

Details of the MOS test are provided in the Appendix A.3.

5 Experimental Results

5.1 Main Results

Zero-shot TTS Performance. Table 2 shows zero-shot TTS performance on Seed-TTS, LibriTTS, VCTK, and AISHELL-3. As mentioned above, Seed-TTS included test-en, test-zh, and test-hard, where test-zh and test-hard were Chinese test sets. Both LibriTTS and VCTK were English benchmarks, while AISHELL-3 was a Chinese benchmark dataset. We compared the proposed TTL with random sampling, referred to as Random, phoneme-based sampling, referred to as Phoneme, and MiniPLM. For all selection methods, we selected the top $\alpha = 6.25\%$ of samples, representing a challenging data-constrained setting.

From the results, TTL demonstrated consistently strong performance across all evaluation metrics and benchmarks. In terms of MOS and UTMOS, which reflect speech quality, TTL demonstrated the best overall performance across the benchmarks in both English and Chinese evaluation sets. Specifically, TTL achieved the highest MOS on almost all benchmarks, except for SEED-TTS test-en. On this benchmark, although MiniPLM achieved a slightly

higher MOS, TTL obtained the highest UTMOS, indicating that it still maintained strong objective speech quality. Moreover, TTL maintained consistently competitive UTMOS scores across all benchmarks. While Random achieved a slightly higher UTMOS on LibriTTS, TTL obtained the highest MOS score of 3.905, indicating that human listeners perceived TTL-generated speech as higher quality. In the case of SECS, which measures speaker similarity, TTL consistently achieved strong and stable performance across all benchmarks, demonstrating its ability to preserve speaker characteristics while maintaining high speech quality.

More notably, TTL demonstrated substantial intelligibility improvements on the Chinese evaluation sets. On SEED-TTS test-zh, TTL achieved the lowest CER of 22.287%, substantially outperforming the other selection methods. Even on the challenging SEED-TTS test-hard benchmark, TTL achieved the best CER among all baseline methods. In contrast, TTL exhibited slightly higher WER than MiniPLM on VCTK and SEED-TTS test-en, likely because MiniPLM utilized substantially more English samples during training. Nevertheless, despite using far fewer English samples, TTL maintained competitive English performance and achieved better WER results on LibriTTS. Considering the substantial intelligibility improvements observed on the Chinese evaluation sets, these results suggest that TTL achieved a more balanced bilingual performance overall. This can be attributed to TTL's language-balanced selection strategy. Even with fewer English samples, TTL still maintained competitive English performance, suggesting that TTL effectively identified more informative samples. Overall, these results demonstrate that TTL successfully mitigated high-resource language bias while maintaining strong performance across both languages under a fixed data budget.

Comparison of Levenshtein Distance Components. Since TTL achieved particularly strong gains in intelligibility, we further analyzed WER and CER by decomposing errors into insertion, deletion, and substitution counts, as shown in Table 3. On VCTK, the performance improvement of TTL was largely driven by a sharp drop in deletion and substitution errors. In particular, deletions decreased from 3039 to 1159 and substitutions decreased from 1800 to 1162. While insertions also decreased from 1097 to 831, most of the improvement came from fewer deletions and substitutions, suggesting fewer missing words and clearer pronunciation, directly improving ASR-based intelligibility. The effect was even more significant on the Chinese test sets. On Seed-TTS test-zh, TTL lowered CER to 22.287% from 83.802% with Random and 30.666% with Phoneme. This dramatic reduction aligned with a substantial decrease in substitution errors, which fell from 23553 to 4364, alongside large reductions in deletions from 11342 to 4231. On the AISHELL-3, TTL achieved a CER of 9.456%, a dramatic reduction from 68.727% with Random and over 100% with Phoneme. This improvement was largely driven by a sharp drop in substitutions from 139124 to 10597, with insertions also decreasing from 14323 to 3935 and deletions from 21515 to 864. Overall, this error decomposition showed that TTL improved intelligibility primarily by reducing deletions and substitutions, with the largest benefits on Chinese test sets.

VCTK				
Method	**WER(↓)**	**#INS**	**#DEL**	**#SUB**
Random	8.807	1097	3039	1800
Phoneme	25.90	13091	3051	1300
Ours, TTL	**5.802**	831	1159	1162
SEED-TTS test-zh				
Method	**CER(↓)**	**#INS**	**#DEL**	**#SUB**
Random	83.802	613	11342	23553
Phoneme	30.666	2523	7398	2970
Ours, TTL	**22.287**	915	4231	4364
AISHELL-3				
Method	**CER(↓)**	**#INS**	**#DEL**	**#SUB**
Random	68.727	14323	21515	139124
Phoneme	> 100	167590	9414	56056
Ours, TTL	**9.456**	3935	864	10597

Table 3: Comparison of Levenshtein distance components on LibriTTS, test-zh, and AISHELL-3 with $\alpha = 6.25\%$, reporting insertion, deletion, and substitution counts. Bold indicates the best result.

Zero-shot TTS Performance under Extremely Small Selection Ratio. When decreasing the selection ratio to $\alpha = 3.125\%$, TTL continued to outperform other methods on both perceptual quality and intelligibility across VCTK, Seed-TTS test-zh, and AISHELL-3, as shown in Table 4. On VCTK, TTL achieved the lowest WER of 42.874% while maintaining a competitive UTMOS of 3.214 and SECS of 0.746. On Seed-TTS test-zh, TTL delivered the lowest CER at 54.064%, corresponding to about a 42% relative reduction from

VCTK			
Method	**UTMOS(↑)**	**WER(↓)**	**SECS(↑)**
Random	**3.317**	44.699	**0.746**
Phoneme	2.613	75.848	0.685
Ours, TTL	3.214	**42.874**	**0.746**
SEED-TTS test-zh			
Method	**UTMOS(↑)**	**CER(↓)**	**SECS(↑)**
Random	2.523	93.815	0.906
Phoneme	2.551	98.30	0.822
Ours, TTL	**2.760**	**54.064**	**0.921**
AISHELL-3			
Method	**UTMOS(↑)**	**WER(↓)**	**SECS(↑)**
Random	2.729	92.179	0.901
Phoneme	2.264	> 100	0.769
Ours, TTL	**2.838**	**22.122**	**0.921**

Table 4: Objective (UTMOS, WER, SECS) results on test-hard and VCTK/AISHELL-3 with $\alpha = 3.125\%$. > 100 indicates that WER/CER exceeded 100%, meaning the ASR transcription fails. Bold indicates the best result.

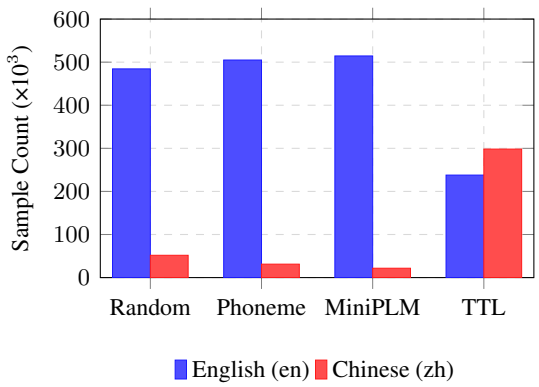

Figure 2: Number of selected samples per language across sampling methods. TTL adopts the proposed language balancing, while TTL w/o balancing denotes the proposed selection without language balancing.

Random, while achieving the highest UTMOS of 2.760 and SECS of 0.921. On AISHELL-3, TTL achieved a WER of 22.122%, a dramatic reduction from 92.179% with Random and over 100% with Phoneme, again with the highest UTMOS and SECS. Notably, TTL achieved the highest SECS on both Chinese benchmarks, indicating that speaker similarity was well preserved even under an extremely data-constrained setting. This confirms that TTL consistently selected more informative and higher-quality samples.

Language Distribution. Since the training corpus was English-heavy, English naturally appeared as the high-resource language. As shown in Fig. 2, conventional selection methods inherited this imbalance, producing subsets dominated by English with relatively few Chinese samples. TTL without language balancing also exhibited a highly biased language distribution, indicating that score–based selection alone can further amplify the high-resource language bias. This observation motivated the balancing strategy as a necessary component of TTL for bilingual training. In the proposed framework, TTL explicitly applied language balancing, enforcing adaptive sampling ratio between English and Chinese so that both languages are consistently represented in the selected subset. Interestingly, TTL selected substantially fewer English samples than the other methods, yet still achieved strong English performance. This suggests that TTL selected more informative English samples, enabling improved English performance even with reduced English data.

Method	***SEED-TTS test-en***	***SEED-TTS test-zh***
	WER(↓)	**CER(↓)**
Random	20.849	83.519
Phoneme	20.387	30.932
Ours, TTL	**17.142**	**23.131**
Method	***SEED-TTS test-hard***	***LibriTTS***
	CER(↓)	**WER(↓)**
Random	91.606	23.469
Phoneme	54.888	32.135
Ours, TTL	**47.202**	**18.485**
Method	***VCTK***	***AISHELL-3***
	WER(↓)	**CER(↓)**
Random	9.176	68.497
Phoneme	22.151	75.511
Ours, TTL	**5.908**	**12.613**

Table 5: WER/CER results evaluated with the SenseVoiceSmall ASR model under a fixed subset budget of $\alpha = 6.25\%$.

5.2 Analysis

Robustness Across Different ASR Models. To further validate the robustness of our evaluation, we additionally conducted ASR experiments with SenseVoiceSmall (An et al., 2024). As shown in Table 5, with $\alpha = 6.25\%$, TTL achieved the best performance across all evaluation datasets, including Seed-TTS test-en, test-zh, test-hard, LibriTTS, VCTK, and AISHELL-3. In particular, TTL significantly reduced CER on the Chinese datasets compared to other methods. Similar trends were observed for the smaller subset budget of $\alpha = 3.125\%$, which will be provided in Appendix A.4.

Acoustic and Linguistic Feature Distribution. Fig. 3 analyzed how different sampling strategies

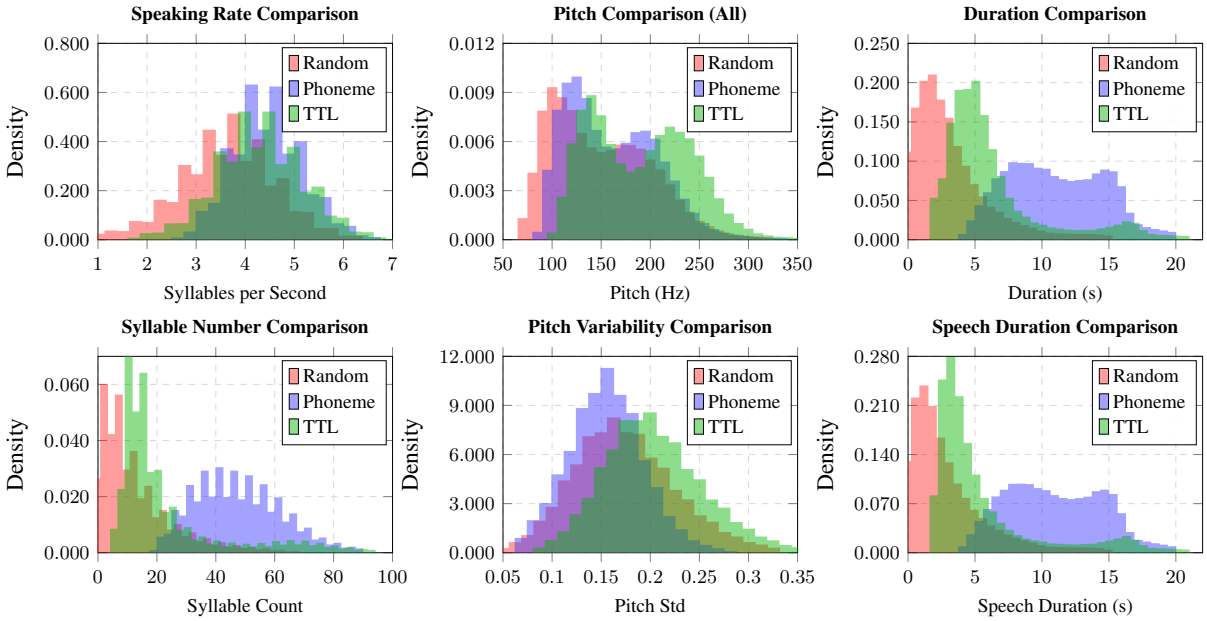

Figure 3: Comparison of acoustic and linguistic feature distributions for training subsets selected by three sampling strategies, including Random, Phoneme, and TTL. Each panel shows the density of a feature computed on the selected subsets: speaking rate, pitch, utterance duration, syllable count, pitch variability, and speech duration. The histograms correspond to Random (red), Phoneme (blue), and TTL (green).

shaped the distributions of acoustic and linguistic features. For speaking rate, the random sampling included more slow utterances, whereas the phoneme-balanced sampling concentrated more tightly around mid-to-fast speaking rates. TTL covered a wider range around typical speaking rates, suggesting that it preserved diversity without overemphasizing unusually slow or fast speech. For duration and syllable count, the random sampling was heavily skewed toward very short utterances and low syllable counts, which could limit exposure to sentence-level structure. In contrast, the phoneme-balanced sampling was biased toward much longer utterances and higher syllable counts, which could increase difficulty and exacerbate failures such as omissions or alignment issues. However, the proposed TTL lay between these extremes, emphasizing mid-length utterances while still retaining a meaningful tail of longer samples. Overall, by avoiding the short-utterance bias of the random sampling and the long-utterance bias of the phoneme-balanced approach, TTL selected training samples from the bilingual corpus that were both informative and learnable.

Computational Efficiency. LLM-based TTS typically requires large-scale training, resulting in substantial computational cost. TTL improved training efficiency, while introducing additional overhead from dataset scoring. To analyze training cost, Table 6 reports the GPU-hours required for each stage, including dataset scoring, and final TTS model training. TTL reduced the end-to-end training time from 144 hours to 48 hours, resulting in a 3× speedup. Overall, despite the additional overhead, TTL significantly reduced the total training time.

Task		Time	GPU
Scoring	Teacher Scoring (1.7B)	33h	8
	Student Scoring (135M)	7h	
Final Model Training (135M)		8h	
Full Dataset Training (135M)		144h (≈ 6d)	

Table 6: Breakdown of GPU-hours for each stage of TTL, including dataset scoring, final model training with $\alpha = 6.25\%$, and full data training with $\alpha = 100\%$.

6 Conclusions

In this paper, we introduced TTL, a data selection framework for LLM-based TTS that leveraged a probability gap computed on codec tokens. We also addressed the overlooked language imbalance issue in bilingual TTS training through an adaptive language-balanced selection strategy. Extensive experiments demonstrated that TTL consistently improved naturalness, intelligibility, and bilingual TTS performance in data-constrained settings.

Limitations

The proposed selection method has a limitation. The current evaluation is limited in bilingual scope. This is consistent with prior studies on codec and LLM-based TTS, which commonly focus on English and Chinese, and it would be valuable to validate TTL under more diverse multilingual settings with more languages, varying resource levels, and different domain distributions. Despite this limitations, TTL represents a first attempt at data selection tailored to LLM-based TTS that explicitly accounts for codec-token structure and bilingual training. Empirically, it achieves strong improvements in both perceptual quality and intelligibility under data-constrained settings.

Ethics Statements

This work does not raise any ethical concerns.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, and 1 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, and 27 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *Preprint*, arXiv:2406.02430.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2024. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. In *Proc. IEEE SLT*.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-fi multi-speaker english tts dataset. In *Proc. INTERSPEECH*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, and 1 others. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Guoguo Chen, Shuzhou Chai, Guan-Bo 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, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Proc. INTERSPEECH*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proc. ACL*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *Preprint*, arXiv:2407.05407.

Yuxian Gu, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2025. Miniplm: Knowledge distillation for pre-training language models. In *Proc. ICLR*.

Sanaul Haq, Philip J. B. Jackson, and James D. Edge. 2008. Audio-visual feature selection and reduction for emotion classification. In *Proc. AVSP*.

Jesin James, Li Tian, and Catherine Inez Watson. 2018. An open source emotional speech corpus for human robot interaction applications. In *Proc. INTERSPEECH*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *Proc. ICML*.

Yuma Koizumi, Heiga Zen, Shigeki Karita, and 1 others. 2023. Libritts-r: A restored multi-speaker text-to-speech corpus. In *Proc. INTERSPEECH*.

Akira Kurematsu, Kazuya Takeda, Hisao Kuwabara, and Kiyohiro Shikano. 1989. Atr japanese speech database as a tool of speech recognition and synthesis. In *Proc. SIOA*.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *Proc. ICASSP*.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2023. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proc. ACM MM*.

Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. 2024. Generative expressive conversational speech synthesis. In *Proc. ACM MM*.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess). *PLOS ONE*, 13(5):e0196391.

Ltd. Magic Data Technology Co. 2019. Magicdata mandarin chinese read speech corpus. `http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101`.

Luz Martinez, Mohammed Abdelwahab, and Carlos Busso. 2020. The msp-podcast corpus: An annotated multimodal dataset of emotionally rich natural speech. In *Proc. INTERSPEECH*.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Proc. INTERSPEECH*.

Kari Ali Noriy, Xiaosong Yang, and Jian Jun Zhang. 2023. Emns /imz/ corpus: An emotive single-speaker dataset for narrative storytelling. *Preprint*, arXiv:2305.13137.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *Proc. ICASSP*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. ACL*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.

Kentaro Seki, Shinnosuke Takamichi, Takaaki Saeki, and Hiroshi Saruwatari. 2024. Diversity-based core-set selection for text-to-speech with linguistic and acoustic features. In *Proc. ICASSP*.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. In *Proc. INTERSPEECH*.

Jianhua Tao, Fangzhou Liu, Meng Zhang, and Huibin Jia. 2008. Design of speech corpus for mandarin text to speech. In *The Blizzard Challenge Workshop*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Proc. ECCV*.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, and 6 others. 2025a. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *Preprint*, arXiv:2503.01710.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2025b. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. In *Proc. ICLR*.

Hanke Xie, Dake Guo, Chengyou Wang, Yue Li, Wenjie Tian, Xinfa Zhu, Xinsheng Wang, Xiulin Li, Guanqiong Miao, Bo Liu, and Lei Xie. 2025. Dialospeech: Dual-speaker dialogue generation with llm and flow matching. *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–6.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. In *Proc. NeurIPS*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yike Guo, and Wei Xue. 2025. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *Preprint*, arXiv:2502.04128.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2022. Vector-quantized image modeling with improved vqgan. In *Proc. ICLR*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Proc. INTERSPEECH*.

Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Jiantao Qiu, Lei Cao, Ju Fan, Ye Yuan, Guoren Wang, and Conghui He. 2025. Harnessing diversity for important data selection in pretraining large language models. In *Proc. ICLR*.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. In *Proc. ACL*.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *Proc. ICASSP*.

Shuoyi Zhou, Yixuan Zhou, Weiqin Li, Jun Chen, Runchuan Ye, Weihao Wu, Zijian Lin, Shun Lei, and Zhiyong Wu. 2024. The codec language model-based zero-shot spontaneous style tts system for covoc challenge 2024. In *Proc. IEEE ISCSLP*.

A Appendix

A.1 Model Configuration

To process speech, we employed BiCodec, a single-stream speech codec that effectively decomposes audio into low-bitrate semantic tokens(50 TPS) for linguistic content and fixed-length global tokens(32 tokens) for speaker attributes. Building upon this tokenization, our speech language model was based on SmolLM2 architecture, utilizing a decoder-only transformer framework. The model was trained with the AdamW optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.96$, to optimize the autoregressive prediction of speech tokens.

A.2 Algorithm of TTL : Language-Balanced Data Selection via Probability Gap

The pseudocode of the proposed language-balanced data selection strategy is presented in Algorithm 1.

Algorithm 1 TTL : Language-Balanced Data Selection via Probability Gap

Require: Full dataset $\mathcal{D}$, Sampling rate α, Language balancing ratio λ_l, Teacher θ_{tea}, Student θ_{stu}

Ensure: Refined subset $\mathcal{D}'$

$\mathcal{D}' \leftarrow \emptyset$
for each language $l \in \{en, zh\}$ **do**
  $\mathcal{D}_l \leftarrow \{x \in \mathcal{D} \mid \text{Lang}(x) = l\}$
  for each sample $x \in \mathcal{D}_l$ **do**
    $s(p_{\theta_{tea}}, p_{\theta_{stu}}, x) \leftarrow \log p_{\theta_{tea}}(y|x) - \log p_{\theta_{stu}}(y|x)$
  end for
  Estimate density of $\{s(p_{\theta_{tea}}, p_{\theta_{stu}}, x) \mid x \in \mathcal{D}_l\}$ via KDE
  Identify inflection point t_l where f'' transitions from negative to positive
  $\lambda_l \leftarrow \frac{|\{x \in \mathcal{D}_l | s(x) > t_l\}|}{\sum_l |\{x \in \mathcal{D}_l | s(x) > t_l\}|}$
  $N_l \leftarrow \lfloor \lambda_l \cdot \alpha \cdot |\mathcal{D}| \rfloor$
  Sort $\mathcal{D}_l$ in descending order based on score $s(p_{\theta_{tea}}, p_{\theta_{stu}}, x)$
  $\mathcal{D}'_l \leftarrow$ Select top N_l samples from sorted $\mathcal{D}_l$
  $\mathcal{D}' \leftarrow \mathcal{D}' \cup \mathcal{D}'_l$
end for
return $\mathcal{D}'$

A.3 MOS Test Details

We recruited 10 evaluators to assess 600 samples, with 100 samples from each dataset. Participants rated the overall naturalness on a scale from 1 to 5, taking into account speech quality, intonation, and pronunciation. All samples were presented blindly and in a randomized order to eliminate potential bias.

A.4 Robustness Across Different ASR Models

As mentioned in Section 5.2, we also evaluated TTL with SenseVoiceSmall under the smaller subset budget of $\alpha = 3.125\%$. Table 7 reports the results. TTL consistently showed the lowest WER/CER on most evaluation datasets, with the exception of test-en, where Random achieved a lower WER. Despite this, TTL maintained the best intelligibility across all other test sets, indicating the robustness of the results across different ASR models.

Method	*SEED-TTS test-en* WER(↓)	*SEED-TTS test-zh* CER(↓)
Random	**62.443**	91.399
Phoneme	76.151	93.289
Ours, TTL	72.697	**55.439**
Method	***SEED-TTS test-hard*** **CER(↓)**	***LibriTTS*** **WER(↓)**
Random	96.541	67.285
Phoneme	96.763	79.034
Ours, TTL	**80.507**	**62.991**
Method	***VCTK*** **WER(↓)**	***AISHELL-3*** **CER(↓)**
Random	55.905	86.980
Phoneme	69.226	> 100
Ours, TTL	**52.78**	**24.374**

Table 7: WER/CER results evaluated with the SenseVoiceSmall ASR model under a fixed subset budget of $\alpha = 3.125\%$. > 100 indicates that WER/CER exceeded 100%, meaning the ASR transcription fails.

A.5 Data Statistics

The full dataset $\mathcal{D}$ comprises a total of 10,098.14 hours of speech data, consisting of approximately 8.6 million utterances. As summarized in Table 8, the dataset exhibits a severe linguistic imbalance:

- **English Subset (Dominant):** English samples dominate the corpus, accounting for 8,998.93 hours (approximately 89.1% of the total volume) sourced from 19 distinct datasets.
- **Chinese Subset (Minority):** In sharp contrast, Chinese samples are limited to 9 datasets totaling 1,099.21 hours, representing only 10.9% of the total volume.

Such a skewed distribution poses a challenge for bilingual modeling, as naïve training may lead to performance degradation on the under-represented language.

Data	Language	#Utterance	Male (h)	Female (h)	Total (h)
AISHELL-3 (Shi et al., 2021)	Chinese	88,035	16.01	69.61	85.62
CASIA (Tao et al., 2008)	Chinese	857	0.25	0.20	0.44
ESD (Zhou et al., 2022)	Chinese	16,101	6.69	7.68	14.37
HQ-Conversations (Zhou et al., 2024)	Chinese	50,982	35.77	64.23	100.00
M3ED (Zhao et al., 2022)	Chinese	253	0.04	0.06	0.10
MAGICDATA (Magic Data Technology Co., 2019)	Chinese	609,474	360.31	393.81	754.13
MER2023 (Lian et al., 2023)	Chinese	1,667	0.86	1.07	1.93
NCSSD-CL-CN (Liu et al., 2024)	Chinese	98,628	53.83	59.21	113.04
NCSSD-RC-CN (Liu et al., 2024)	Chinese	21,688	7.05	22.53	29.58
Total	Chinese	887,685	480.81	618.40	1,099.21
CREMA-D (Cao et al., 2014)	English	809	0.30	0.27	0.57
DailyTalk (Lee et al., 2023)	English	23,754	10.79	10.86	21.65
EMNS (Noriy et al., 2023)	English	918	0.00	1.49	1.49
EmoV-DB (James et al., 2018)	English	3,647	2.22	2.79	5.86
Expresso (Nguyen et al., 2023)	English	11,595	5.47	5.39	10.86
Gigaspeech (Chen et al., 2021)	English	6,619,339	4,310.19	2,885.66	7,195.85
Hi-Fi TTS (Bakhturina et al., 2021)	English	323,911	133.31	158.38	291.68
IEMOCAP (Busso et al., 2008)	English	2,423	1.66	1.31	2.97
JL-Corpus (James et al., 2018)	English	893	0.26	0.26	0.52
Librispeech (Panayotov et al., 2015)	English	230,865	393.95	367.67	761.62
LibriTTS-R (Koizumi et al., 2023)	English	363,270	277.87	283.03	560.90
MEAD (Wang et al., 2020)	English	3,767	2.26	2.42	4.68
MELD (Poria et al., 2019)	English	5,100	2.14	1.94	4.09
MSP-Podcast (Martinez et al., 2020)	English	796	0.76	0.56	1.32
NCSSD-CL-EN (Liu et al., 2024)	English	62,107	36.84	32.93	69.77
NCSSD-RL-EN (Liu et al., 2024)	English	10,032	4.18	14.92	19.09
RAVDESS (Livingstone and Russo, 2018)	English	950	0.49	0.48	0.97
SAVEE (Haq et al., 2008)	English	286	0.15	0.15	0.31
TESS (Yu et al., 2022)	English	1,956	0.00	1.15	1.15
VCTK (Yamagishi et al., 2019)	English	44,283	16.95	24.51	41.46
Total	English	7,710,721	5,202.79	3,796.14	8,998.93
Overall Total		8,598,406	5,683.6	4,414.54	10,098.14

Table 8: Statistical overview of the full dataset D used in our experiments.