Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Toward Data-Efficient Speech Synthesis: Active Learning-Based Corpus Construction for Multi-Speaker Text-to-Speech Synthesis

KENTARO SEKI[1,2], (Student Member, IEEE), YUKI SAITO[1] (Member, IEEE), SHINNOSUKE TAKAMICHI[1,2] (Member, IEEE), TAKAAKI SAEKI[1] (Member, IEEE), and HIROSHI SARUWATARI[1] (Member, IEEE)

[1]Graduate School of Information Science and Technology, The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 133–8656 Japan
[2]Graduate School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawa 223-8522 Japan

Corresponding author: Kentaro Seki (e-mail: seki-kentaro922@g.ecc.u-tokyo.ac.jp).

This work was supported by JSPS KAKENHI 22H03639, 24KJ0860 and Moonshot R&D Grant Number JPMJPS2011.

ABSTRACT Text-to-speech (TTS) technology is rapidly expanding its applications, including speech-based human-AI interactions. While large-scale data has driven this progress, most previous studies have focused primarily on increasing data volume. However, as vast amounts of web-collected speech data become available, the next challenge is to determine which data is truly beneficial for learning and how to utilize it effectively. This paper proposes an active learning–based corpus construction method for multi-speaker TTS. The proposed framework automatically acquires and selects additional data based on the training and evaluation results of the TTS model, enabling efficient corpus expansion. Experimental results demonstrate that the proposed model-aware corpus construction approach outperforms conventional static sampling and evaluation-in-the-loop methods, achieving consistent improvements in both naturalness and recognition accuracy. Step-wise analysis further reveals that quality filtering and data acquisition contribute to learning improvements in complementary ways. Overall, this study advances TTS corpus construction from a static preparation process to a model-evolving learning process, laying the foundation for data-efficient and self-improving speech synthesis systems.

INDEX TERMS Active learning, automated corpus construction, data-efficient machine learning, text-to-speech synthesis.

I. INTRODUCTION

TEXT-to-speech (TTS) technology has rapidly advanced in recent years, driven by the development of large-scale neural architectures that enable natural and expressive speech synthesis across multi-speaker and multilingual settings [1]–[8]. The application scope of TTS continues to expand, particularly with the emergence of use cases such as voice-based interaction and integration with generative AI systems [9]–[11]. This progress has been supported by large-scale corpora built from a variety of sources, including studio-recorded datasets, ASR-oriented corpora, and web-scale speech resources such as YouTube and podcasts [12]–[16]. As TTS models become increasingly powerful, their performance depends not only on model design, but also on the training data. For this reason, corpus construction has become a critical component in the development of modern TTS systems.

Web-collected speech data is inherently noisy and heterogeneous, often containing misaligned, low-quality, or irrelevant utterances. To address this, recent web-oriented speech corpus construction efforts have primarily focused on quality filtering—removing undesirable samples to ensure a clean and reliable dataset [12]–[15]. While such methods have proven effective in improving overall data quality, these approaches often emphasize collecting as much data as possible and overlook the critical role of efficiency. In particular, web-scale data often exhibits biases in speaker demographics, content topics, and recording conditions. As a result, indiscriminately collecting large volumes of data can lead to significant redundancy and overlap among samples. If we can instead focus on identifying and acquiring data that more directly contributes to model improvement, training efficiency could be significantly enhanced. This becomes especially important in the current era of increasingly large models and datasets,

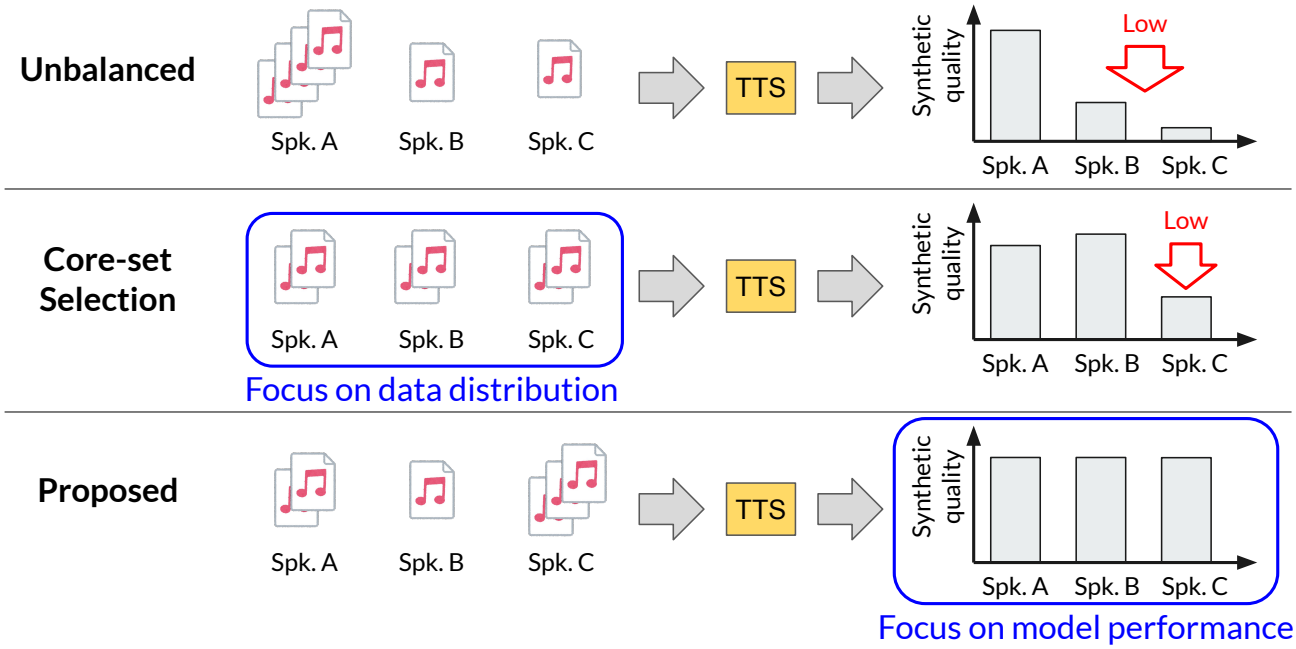


FIGURE 1: Three paradigms of corpus construction for multi-speaker TTS. Web-scale speech data is inherently unbalanced, and naive sampling results in biased corpora, leading to uneven synthesis quality across speakers. Core-set selection [17] mitigates this by enforcing coverage and diversity, but remains model-agnostic and cannot determine which balance is most effective for improving model performance, resulting in residual low-quality synthesis. In contrast, our proposed method directly incorporates model performance to construct corpora that are better aligned with the model's learning dynamics and synthesis quality.

where even modest improvements in data efficiency can yield substantial benefits.

To improve corpus efficiency, core-set selection methods have been proposed [17], which aim to extract representative samples that preserve diversity in linguistic content and speaker attributes. This approach constructs more balanced and compact datasets by minimizing redundancy and ensuring sufficient coverage across latent feature space, demonstrating improvements in training efficiency by reducing the dataset without degrading model performance. However, as illustrated in Fig. 1, these methods rely solely on distributional properties of the data and do not take into account the performance of the model itself. As a result, they are unaware of which aspects of the data the model already handles well, and which areas—such as specific speakers or phonetic contexts—it currently struggles with or lacks coverage for. Without access to this feedback loop, such methods cannot prioritize data that would most effectively improve synthesis performance across underrepresented or low-quality regions of the corpus.

In this study, we propose a corpus construction framework for multi-speaker TTS based on active learning, with the goal of directly optimizing model performance through data filtering and data acquisition. The proposed method extends existing evaluation-in-the-loop filtering approaches [14], [18] by incorporating a sampling loop that dynamically acquires additional data based on the model's training and evaluation outcomes. This enables the model to identify what kinds of data are currently lacking, and to actively expand the corpus in a way that improves learning efficiency. As a result, corpus construction becomes an adaptive and self-improving process that evolves in tandem with the model's training dynamics.

This paper extends our previous conference publication, which introduced the core idea of informative sampling for TTS corpus construction [19]. The prior work demonstrated the potential of selecting data based on its expected contribution to model learning, but only included a limited set of preliminary experiments. In this journal version, we provide a more comprehensive description of the proposed method, clarify its relationship to prior work, and offer a broader evaluation including comparisons with core-set selection and detailed step-wise analysis.

The main contributions of this paper are summarized as follows:

- We propose a model-aware and data-efficient corpus construction method for multi-speaker TTS, based on active learning.
- We demonstrate the effectiveness of the proposed method across multiple TTS architectures.
- We analyze each step of the proposed framework and show how quality filtering and data acquisition contribute differently to corpus improvement.

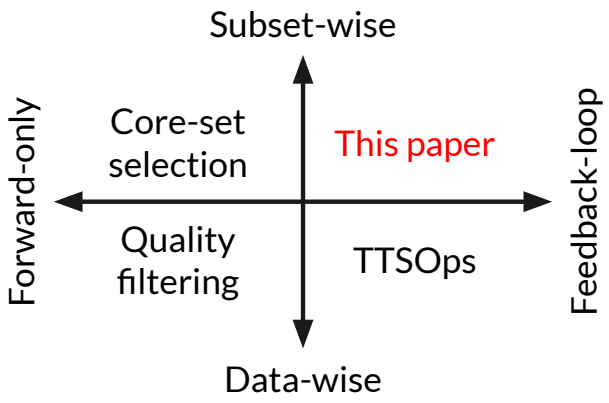


FIGURE 2: Taxonomy of data-selection methods in TTS.

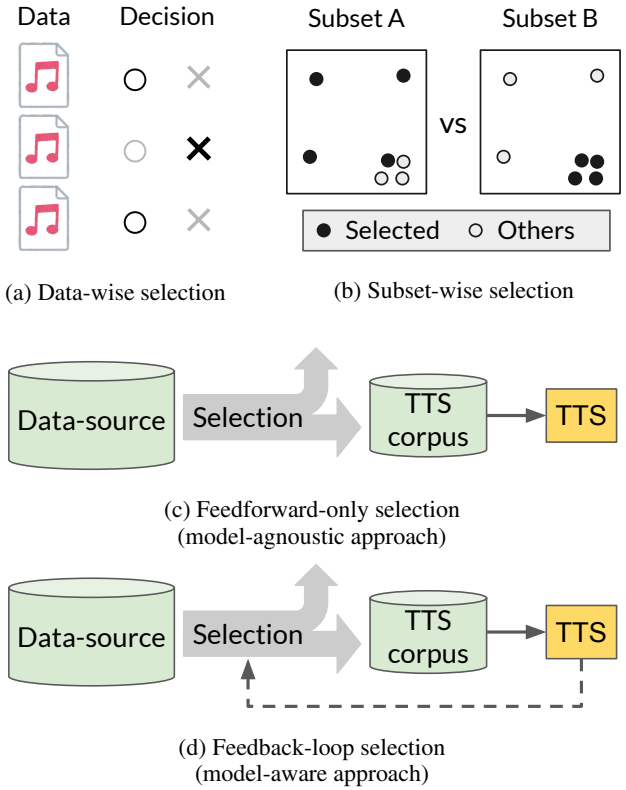


FIGURE 3: Each data selection method

II. RELATED WORK

A. CORPUS CONSTRUCTION FOR TTS

TTS systems have relied on studio-recorded corpora [20]–[23], which provide clean and well-aligned speech–text pairs. These datasets have significantly contributed to improving the quality of neural TTS systems, yet their scalability is limited due to the high cost of data collection. To reduce this cost and expand data coverage, approaches that repurpose existing resources originally developed for other speech tasks have been explored. For instance, LibriTTS [24] was derived from the automatic speech recognition (ASR) corpus LibriSpeech [25], offering a large multi-speaker dataset for training high-fidelity neural TTS models. Recently, ASR and related speech tasks have expanded to web-scale data collection, as seen in GigaSpeech [12] and JTubeSpeech [13]. Such corpora have demonstrated the feasibility of mining massive speech resources directly from online media, and similar approaches are now being adapted to speech generation tasks, including Emilia [16] and J-CHAT [15]. While these large-scale efforts have accelerated TTS research, they often emphasize quantity over quality and remain largely independent of the model and its learning dynamics. As the field continues to expand data scale, it becomes increasingly important to explore more efficient strategies for utilizing massive speech resources rather than relying solely on brute-force data accumulation.

B. DATA SELECTION FOR TTS

The taxonomy of data-selection methods in TTS is illustrated in Fig. 2, and the conceptual interpretation of each axis is presented in Fig. 3. Data-wise selection makes independent decisions for each sample without considering interactions between data, whereas subset-wise selection evaluates data as a set, aiming to optimize diversity and representativeness of the corpus. Feedforward-only selection methods determine data subsets according to predefined rules and are therefore model-agnostic, while feedback-loop approaches incorporate model evaluation into the data-selection process, achieving model-aware and adaptive data optimization.

The most basic approaches, located in the lower-left region of Fig. 2, are based on the simple idea of discarding low-quality data [24], [26]–[28]. For example, LibriTTS [24] constructs a TTS corpus by filtering data according to acoustic-quality indicators such as signal-to-noise ratio (SNR) and text–speech alignment accuracy. This type of quality-based filtering has been widely adopted in TTS research, and several studies have reported that using cleaner, higher-quality data improves the naturalness of synthesized speech [27], [28]. However, because these methods treat each sample independently, they do not consider redundancy or overlap that may exist across the dataset as a whole.

Subset-wise selection approaches have therefore been proposed to address corpus-level redundancy and imbalance (upper-left region in Fig. 2). For example, an unsupervised speaker-subset selection method is proposed to automatically identify a representative subset of speakers from a large multi-speaker corpus, showing that a carefully selected subset can achieve comparable synthesis quality to the full dataset [29]. Similarly, diversity-based core-set selection maximizes a diversity score to sample subsets that broadly cover the feature space, ensuring linguistic and acoustic variety while reducing data volume [17]. However, these methods remain forward-only approaches and depend heavily on human-designed criteria defined before training, and do not adapt to the evolving behavior of the model.

In contrast to these forward-only approaches, feedback-

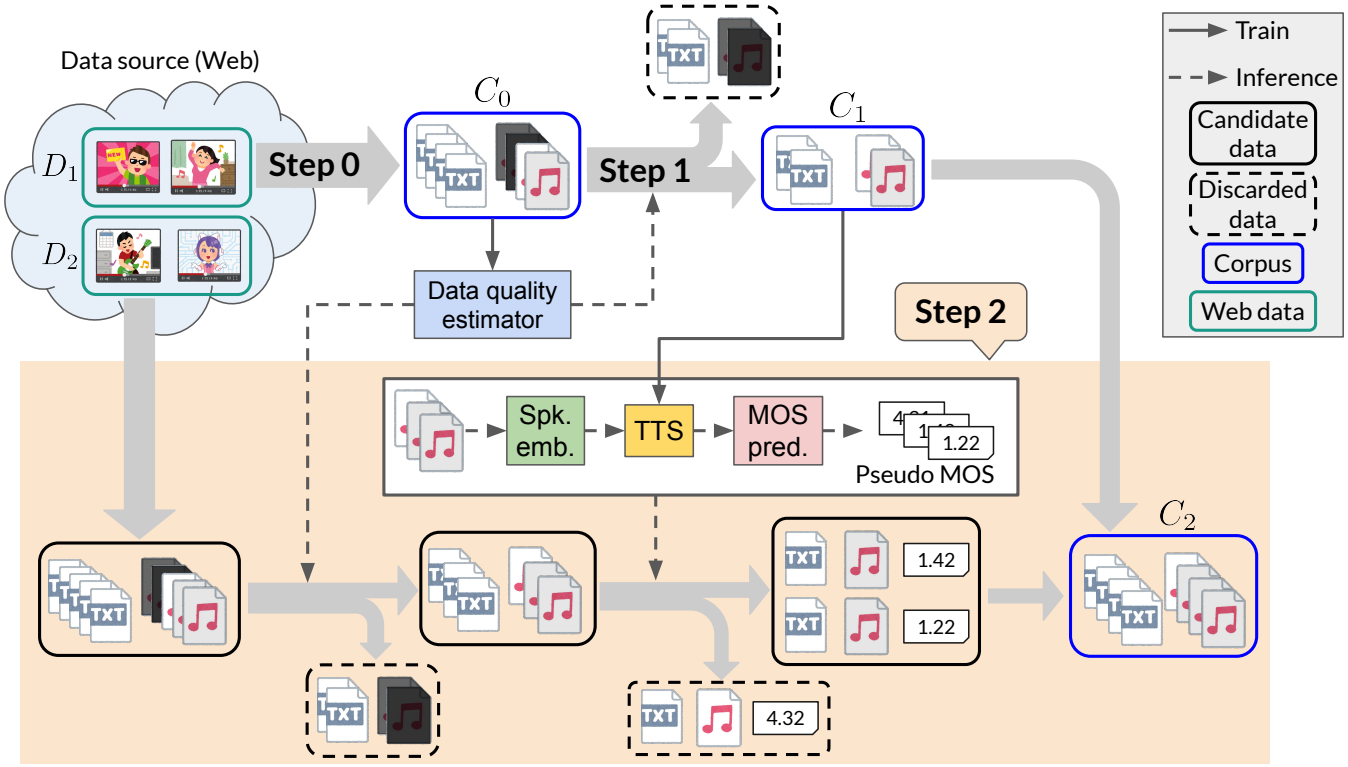


FIGURE 4: Overall procedure of proposed method.

loop strategies have been proposed to make data selection adaptive to model behavior, aiming to directly optimize model performance. Motivated by this idea, TTSOps [18] was introduced as a feedback-driven framework for data selection (lower-right region in Fig. 2). TTSOps trains a TTS model on candidate data, evaluates the synthesized outputs using a pseudo-MOS predictor, and selects samples that contribute most effectively to model learning. This evaluation-in-the-loop design quantifies data quality not by acoustic cleanness but by its contribution to model training, marking a shift toward model-aware corpus construction. Nevertheless, TTSOps still performs data-wise selection and does not yet address corpus-level optimization. To identify what kinds of data the model lacks and to acquire informative samples accordingly, a more adaptive and efficient strategy is required for future TTS data selection (upper-right region in Fig. 2).

C. ACTIVE LEARNING

Active learning is a widely used strategy for reducing annotation costs while maintaining model performance, and has been applied across various fields, including computer vision and natural language processing [30]–[32]. Typical approaches include uncertainty-based sampling, which selects samples based on the model's confidence [33], margin [34], and entropy [35]. Another class of methods is diversity-based sampling, which aims to extract a representative and diverse subset from the unlabeled data pool using clustering or coreset selection [36], [37]. Other approaches based on Bayesian inference have also been proposed [38].

In the speech domain, active learning has primarily been applied to automatic speech recognition (ASR) [39]–[44]. By selecting only the most informative utterances for annotation, these approaches have successfully reduced the word error rate (WER) under constrained labeling budgets. In contrast, its application to TTS corpus construction remains limited. Most existing approaches rely on static sentence selection or heuristic pre-filtering, without leveraging model feedback [45]. This study introduces a model-aware and evaluation-guided data acquisition strategy, extending the active learning paradigm to TTS synthesis.

III. PROPOSED METHOD

To improve the quality and generality of multi-speaker TTS models, it is essential to construct corpora that support stable speech synthesis across a wide range of speakers, including those with limited or noisy data. In this study, we focus on multi-speaker TTS models conditioned on x-vectors [46], and define a speaker as "synthesizable" if the model can generate speech that exceeds a predefined quality threshold. Our goal is to maximize the number of such synthesizable speakers through an efficient corpus construction process.

A. OVERVIEW

As shown in Fig. 4, our proposed method consists of Step 0, Step 1, and Step 2. Step 0 and Step 1 perform a data-wise, model-aware quality-filtering process on an initial subset of web-collected data. This process is based on the feedback-

loop strategy proposed in TTSOps [18], where training data is filtered not using predefined heuristics, but through automatic evaluation of synthetic speech quality, reflecting the model's actual learning behavior.

In contrast, Step 2 introduces a model-aware data acquisition mechanism. Based on the synthesis performance of the model trained on the filtered corpus from Step 1, Step 2 identifies speakers whose data remains insufficient or ineffective for achieving high-quality synthesis, and selectively adds utterances from the remaining data pool. Speakers who already have sufficient high-quality data are excluded, making the sampling subset-wise, and the use of model performance as a selection criterion ensures the process is model-aware.

Together, these procedures form a model-aware feedback loop that selects informative data subsets from a large web-scale candidate pool, enabling data-efficient and performance-oriented corpus construction.

B. PREPARATION

1) Creating a Video ID List

To initialize the corpus construction process, we first create a list of YouTube video IDs, denoted as D_{all}, using the search strategy described in our previous work [14]. This list serves as the global candidate pool for downstream sampling. Prefetching video IDs decouples the search process from data acquisition, ensuring that the candidate distribution remains fixed across multiple sampling stages. Since video IDs are lightweight compared to full audio-text pairs, it is feasible to collect them at large scale in advance.

2) Setting the Target Synthesis Quality

Our goal is to increase the number of speakers for whom high-quality synthetic speech can be generated. Since we expect that synthetic speech comparable to that produced by a model trained on studio-quality data would represent sufficiently high quality for practical purposes, we define a speaker as ''synthesizable'' if the quality of their synthesized speech exceeds a predefined threshold τ_{hq}. To determine this threshold, we train a TTS model on a studio-recorded multi-speaker corpus and generate synthetic speech for each speaker. Using a pseudo-MOS predictor, we evaluate the synthesis quality and define τ_{hq} as the minimum observed score among those speakers. This threshold is used throughout the corpus construction pipeline to assess whether a speaker is adequately covered, both during filtering and in later acquisition stages.

C. STEP 0: INITIAL FILTERING OF CANDIDATE DATA

As an initial step, we randomly sample a subset D_1 from the global video ID pool D_{all}. Following the approach introduced in our previous work [14], we collect audio-text pairs from videos in D_1 that contain manually created subtitles. To remove unreliable samples and ensure basic data quality, we apply two pre-screening criteria based on alignment accuracy and speaker consistency.

First, we compute the connectionist temporal classification (CTC) score [47] for each utterance to assess how well the audio aligns with the corresponding text. Utterances with low alignment scores are discarded. Second, we evaluate speaker consistency within each video by computing the intra-video variance of x-vectors [46], which represent speaker embeddings extracted from the audio. Videos with high intra-video variance are assumed to include multiple or unstable speakers and are removed from the candidate pool.

During this process, we also extract and store one representative x-vector per video, as they are lightweight and will be used in later stages for speaker-level sampling and analysis.

After applying these filtering steps, the remaining dataset is treated as the initial corpus C_0, which serves as the foundation for subsequent quality filtering and active acquisition. Each data sample $x_i \in C_0$ is represented as a triplet (a_i, t_i, s_i), where a_i denotes the audio segment, t_i is the corresponding transcription, and s_i is the speaker embedding extracted as an x-vector.

D. STEP 1: FILTER OUT LOW-QUALITY DATA

Although the initial corpus C_0 consists of utterances that meet basic structural conditions, it may still contain noisy or low-quality data. To address this, we apply a model-aware quality filtering approach based on our previous work [14], [18].

This approach runs a training-evaluation loop to estimate the training data quality of each utterance in C_0. We first train a multi-speaker TTS model on C_0 and synthesize speech for all speakers in C_0. These outputs are then evaluated using a pseudo-MOS prediction model to obtain quality scores. Using these scores as supervision, we train a training data quality (TQ) prediction module, denoted as $\mathrm{TQ}(\cdot;\theta_{\mathrm{TQ}})$, which takes an audio segment a_i as input and predicts its expected synthesis quality. Here, θ_{TQ} denotes the parameters of the training data quality estimator. Since TQ is trained to approximate the pseudo-MOS score, its outputs are directly comparable to the threshold τ_{hq} defined in Section III-B2.

Each utterance (a_i, t_i, s_i) is retained in the corpus C_1 if its estimated quality exceeds the threshold:

$$C_1 = \{(a_i, t_i, s_i) \mid \mathrm{TQ}(a_i;\theta_{\mathrm{TQ}}) > \tau_{\text{hq}}\}. \tag{1}$$

The resulting corpus C_1 consists of samples that are both structurally valid and expected to positively contribute to synthesis quality.

E. STEP 2: ADDITIONAL DATA ACQUISITION

Although the corpus C_1 consists of high-quality utterances, some speakers may still yield poor synthesis results due to insufficient or unbalanced data. To address this, we selectively acquire additional utterances for such speakers.

1) Formulation

Denoting the TTS model trained on C_1 as $\mathrm{TTS}(\cdot,\cdot;\theta_{C_1})$, we define a fixed set of test texts T, and evaluate the average synthesis quality (SQ) for each speaker s_i as:

$$\mathrm{SQ}(s_i;\theta_{C_1}) = \frac{1}{|T|}\sum_{t\in T}\mathrm{MOS}(\mathrm{TTS}(t, s_i;\theta_{C_1})), \tag{2}$$

where $\mathrm{MOS}(\cdot)$ denotes a pseudo-MOS prediction model that estimates the quality of synthesized speech.

We consider a new utterance $x_i = (a_i, t_i, s_i)$ for acquisition only if it satisfies the following two conditions:

$$\mathrm{TQ}(a_i; \theta_{\mathrm{TQ}}) > \tau_{\mathrm{hq}}, \tag{3}$$

$$\mathrm{SQ}(s_i; \theta_{C_1}) < \tau_{\mathrm{hq}}. \tag{4}$$

This ensures that only data from underperforming speakers is added, and only if its quality is expected to be sufficient for synthesis.

2) Procedure

To obtain candidate utterances, we sample from $D_2 = D_{\mathrm{all}} \setminus D_1$, which consists of videos not used in Step 0. We download audio-text pairs using the same procedure as in Section III-C, and apply the data-wise quality filtering described in Step 1 using $\mathrm{TQ}(\cdot; \theta_{\mathrm{TQ}})$ to remove low-quality utterances.

For each speaker in the resulting candidate set, we compute SQ using the current TTS model. If the score is below the threshold τ_{hq}, we include that speaker's utterances in the corpus. These additions form an auxiliary set of informative samples, which we denote as $C_{\mathrm{additional}}$.

The final training corpus is given by $C_2 = C_1 \cup C_{\mathrm{additional}}$.

F. (OPTIONAL) SWITCHING-BASED DATA CLEANSING

To further improve data quality, we optionally apply switching-based data cleansing, following the approach proposed in TTSOps [18]. In this strategy, each utterance is selectively enhanced using a predefined cleansing method (e.g., a speech enhancement model), but only if the enhancement improves its estimated synthesis quality. We consider the case where a single cleansing method is available, and the system must decide whether to apply it or not.

This step is optional because such preprocessing can be computationally expensive, and is intended for scenarios where additional computational cost is acceptable in exchange for further performance improvement.

1) Application in Step 1

In Step 1, we train two training data quality estimators: $\mathrm{TQ}(\cdot; \theta_{\mathrm{TQ}})$ on the original corpus C_0, and $\mathrm{TQ}(\cdot; \tilde{\theta}_{\mathrm{TQ}})$ on a cleansed version $\tilde{C}_0$ in which all audio samples have been preprocessed by a speech enhancement model.

For each utterance $x_i = (a_i, t_i, s_i)$, we compute the estimated quality both before and after cleansing:

$$\mathrm{TQ}(a_i; \theta_{\mathrm{TQ}}), \quad \mathrm{TQ}(\tilde{a}_i; \tilde{\theta}_{\mathrm{TQ}}).$$

If the cleansed version achieves higher quality, i.e.,

$$\mathrm{TQ}(\tilde{a}_i; \tilde{\theta}_{\mathrm{TQ}}) > \mathrm{TQ}(a_i; \theta_{\mathrm{TQ}}),$$

we replace (a_i, t_i, s_i) with $(\tilde{a}_i, t_i, s_i)$ and assign the higher score as the supervision label in constructing C_1.

This enables an adaptive application of data cleansing only when it is expected to improve synthesis quality, thereby avoiding unnecessary distortion.

2) Application in Step 2

A similar strategy can be applied to candidate utterances in Step 2. After downloading audio-text pairs from D_2, we apply the same switching rule based on TQ and its cleansed counterpart. If the quality of the cleansed version is higher than the original, we retain the enhanced sample for further quality filtering and acquisition.

G. INFERENCE PROCESS

At inference time, we sample from the set of x-vectors $\{s_i\}$ that were collected during the corpus construction process. Since the goal of our method is to enable high-quality synthesis for as many speakers in D_{all} as possible, we expect that sampling from these x-vectors will result in synthetic speech that exceeds the quality threshold τ_{hq}.

IV. EXPERIMENTAL EVALUATION

A. EXPERIMENTAL SETUP

1) Dataset

We construct our dataset based on the data collection script[1] provided in JTubeSpeech [13], which we use to retrieve Japanese YouTube videos with manually created subtitles. Following the original script, we first extract a list of $2,719$ video IDs, each corresponding to a unique speaker. This set serves as the global candidate pool D_{all}.

From this pool, we randomly sample 10% of the videos (271 videos) to form the subset D_1, which is used for initial corpus construction. The corresponding initial corpus C_0 is composed of approximately $5,000$ utterances from the 271 speakers in D_1.

The remaining $2,448$ videos (corresponding to unseen speakers) constitute the set $D_2 = D_{\mathrm{all}} \setminus D_1$, from which additional data is obtained during Step 2. This step yields approximately $54,000$ additional utterances, which are selectively filtered and added to the corpus based on model-aware acquisition criteria described in Section III-E.

2) Model and Training

We built a multi-speaker TTS system by combining an acoustic model with the pre-trained HiFi-GAN vocoder [48] UNIVERSAL_V1[2].

Two acoustic models were used: FastSpeech 2 [4] and Matcha-TTS [6]. We followed the official implementations for both models[3][4] with default hyperparameters. To provide speaker conditioning, we used 512-dimensional x-vectors extracted from a pre-trained model[5]. Each x-vector was averaged per speaker and mapped to the model's encoder input through a linear projection.

Only the acoustic model was trained, while the vocoder remained fixed. To initialize the acoustic model, we used a

[1] https://github.com/sarulab-speech/jtubespeech
[2] https://github.com/jik876/hifi-gan
[3] FastSpeech 2: https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[4] Matcha-TTS: https://github.com/shivammehta25/Matcha-TTS
[5] https://github.com/sarulab-speech/xvector_jtubespeech

TABLE 1: Quantitative comparison across different corpus construction strategies. Our proposed active learning method outperforms both the random baseline and core-set selection across all metrics for both FastSpeech 2 and Matcha-TTS.

	FastSpeech 2			Matcha-TTS		
metric	UTMOS (↑)	# HQ-speaker (↑)	CER (↓)	UTMOS (↑)	# HQ-speaker (↑)	CER (↓)
Baseline	2.4192	1467	16.9047	2.1457	784	26.6982
Core-set	2.5055	1780	16.5465	2.2899	1111	25.6426
Active Learning	**2.6019**	**2185**	**16.0223**	**2.3846**	**1577**	**22.9432**

model pre-trained on 10,000 utterances from the JVS corpus [23], a 100-speaker Japanese TTS dataset. The pre-training was performed for 300k steps with a batch size of 16. For training on each corpus (C_0, C_1, or C_2), we fine-tuned the model for 100k steps from the same JVS pre-trained checkpoint. Although continual fine-tuning from the previous model (e.g., from C_0 to C_1) was possible, we found that restarting from the JVS checkpoint at each step yielded better synthesis performance.

For pseudo-MOS estimation, we used a publicly available strong UTMOS model [49][6]. The threshold τ_{hq} was set to 2.4156 for FastSpeech 2 and 2.3693 for Matcha-TTS, respectively. Although these values may appear low in absolute terms, they are consistent within the pseudo-MOS framework and serve as meaningful reference points for relative evaluation. [18].

3) Compared Methods

We compare our method against two baseline strategies, each corresponding to one of the corpus construction paradigms illustrated in Fig. 1.

Baseline: This setting corresponds to the "Unbalanced" configuration shown in Fig. 1. We select training samples at random from the pool of utterances whose estimated training data quality exceeds the threshold τ_{hq}. This ensures that all selected samples satisfy the quality filtering condition, but the resulting corpus distribution remains identical to that of the original web data, and is therefore expected to be biased. The total number of selected samples is matched to that of the proposed method, ensuring a fair comparison under equal data budgets.

Core-set Selection: This setting corresponds to the "Core-set Selection" configuration in Fig. 1. From the same quality-filtered pool, we apply a core-set selection strategy [17] to construct a representative subset. This method performs subset-wise selection by maximizing dataset diversity, but it is model-agnostic and does not take into account synthesis performance.

Active Learning: Our proposed method corresponds to the "Active Learning" configuration in Fig. 1. It performs model-aware, subset-wise selection by actively identifying under-performing speakers and acquiring additional data that is predicted to improve synthesis quality.

[6]https://github.com/sarulab-speech/UTMOS22

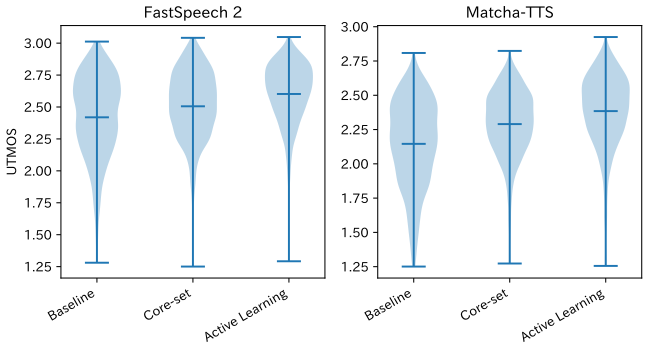


FIGURE 5: Distribution of UTMOS scores for each method. Violin plots show the speaker-wise UTMOS distributions under three corpus construction strategies, with three horizontal lines indicating the maximum, mean, and minimum values respectively. Our proposed active learning method improves the average score and lifts the lower tail of the distribution, indicating enhanced quality for previously low-performing speakers.

4) Evaluation Metrics

Conducting human listening tests for more than one thousand speakers would be prohibitively costly. Therefore, as a limited yet reasonable proxy for large-scale evaluation, we adopt the following three automatic evaluation metrics:

UTMOS: (↑) The average pseudo-MOS across speakers, predicted by a pre-trained UTMOS model [49]. We evaluate the synthetic speech for all $2,719$ speakers and report both the mean and its distribution.

 # HQ-Speaker: (↑) The number of speakers whose average synthetic quality exceeds the threshold τ_{hq}. This metric reflects the effective speaker coverage of the constructed corpus.

CER: (↓) Character error rate (CER) of the synthesized speech, measured using the Whisper large-v3 model[7] as a fixed ASR backend. This metric captures the intelligibility of generated speech.

B. RESULTS

1) Comparison with Baseline

The results are shown in Fig. 5 and Table 1. Figl. 5 visualizes the distribution of speaker-wise UTMOS scores using violin plots, while Table 1 summarizes the quantitative metrics for each corpus construction strategy.

[7]https://github.com/openai/whisper

TABLE 2: Normalized diversity scores $\bar{V}(S)$ for each corpus construction strategy. The score measures pairwise variance in speaker embeddings and reflects the diversity of the resulting training corpus. While the core-set selection achieves the highest diversity, our proposed active learning method shows lower diversity than the random baseline. Despite this, the proposed method achieves superior synthesis quality, suggesting that diversity alone is not sufficient to ensure optimal TTS performance.

	FastSpeech 2	Matcha-TTS
Baseline	0.6677	0.6529
Core-set	**1.2050**	**1.1120**
Active Learning	0.6222	0.5613

Across all metrics and both model architectures, our proposed active learning method outperforms the baselines. Compared to the random baseline ("Baseline"), it achieves substantial gains in UTMOS, #HQ-speaker, and CER, demonstrating that model-aware, subset-wise selection is effective in improving both naturalness and intelligibility. The advantage is particularly evident in the lower tail of the UTMOS distribution, where our method noticeably lifts low-performing speakers, suggesting better data coverage for difficult cases.

Compared to core-set selection ("Core-set"), our method also shows consistent improvements, especially in average UTMOS and #HQ-speaker. This indicates that while diversity-based selection helps mitigate data imbalance, explicitly incorporating model feedback offers additional benefits in synthesis quality.

2) Comparison with Baseline Methods

The results are shown in Fig. 5 and Table 1. Fig. 5 shows violin plots of speaker-wise UTMOS distributions, and Table 1 summarizes quantitative results. Our proposed active learning-based method improved the performance of both models across all metrics, confirming its effectiveness. In particular, when compared with the core-set selection method, the lower part of the distribution shifts upward, indicating that the addition of data in Step 2 effectively raised the overall TTS quality. Furthermore, a comparison between the baseline and core-set selection shows that the core-set method consistently outperforms the baseline under all conditions, confirming its validity.

3) Diversity Analysis

To analyze the characteristics of each method, we analysed the objective function used in "Core-set" approaches for subset construction. Specifically, we consider the following diversity score [17]:

$$V(S) = \sum_{x,y \in S} \|x - y\|^2 \tag{5}$$

Here, S denotes the selected subset, and x, y are fixed-dimensional feature vectors representing the data samples.

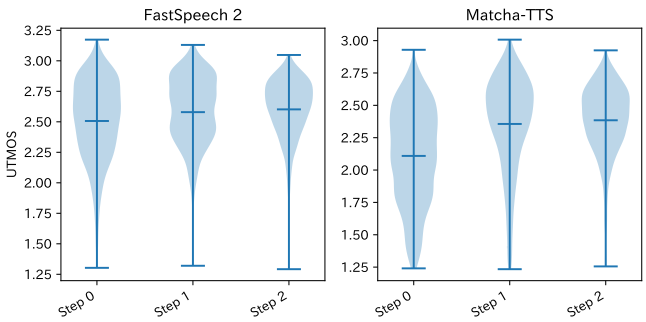


FIGURE 6: Distribution of speaker-wise UTMOS scores at each step of the "Active Learning" method.

Since this score grows quadratically with the subset size $|S|$, we instead compute the normalized diversity score:

$$\bar{V}(S) = \frac{1}{|S|^2} V(S) \tag{6}$$

This allows for a fair comparison of diversity across corpora of different sizes.

Table 2 presents the normalized diversity score $\bar{V}(S)$ for each method. As expected, the core-set selection method achieves the highest diversity. Interestingly, the active learning method yields a lower diversity score than the random baseline.

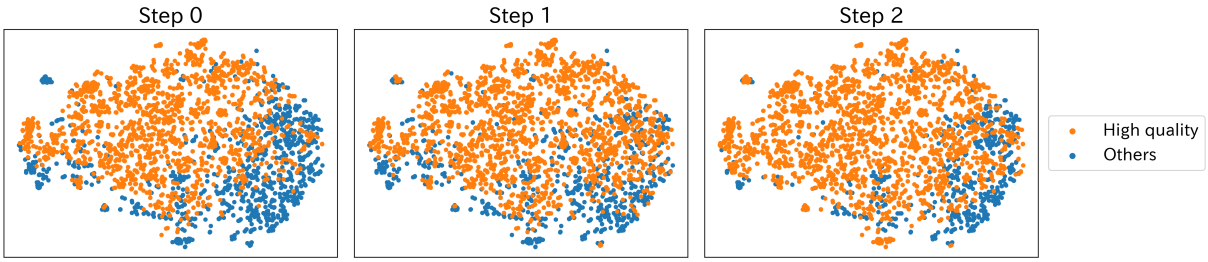
To investigate this behavior, we also compute $\bar{V}(S)$ for the subset constructed in Step 1, obtaining 0.5594 for FastSpeech 2 and 0.5165 for Matcha-TTS—both lower than the random baseline. This indicates that the diversity reduction in active learning is likely due to Step 1 being restricted to the initial randomly sampled speakers, which limits the overall diversity.

It is important to note that the diversity score is an auxiliary indicator of corpus structure, and does not directly reflect TTS performance. A lower diversity score does not necessarily imply a lower-quality corpus. In fact, as shown in Table 1, our active learning method outperforms core-set selection in all synthesis-related metrics, despite achieving a lower diversity score in Table 2. This suggests that diversity, while useful for analyzing corpus characteristics, may not be sufficient on its own to optimize synthesis performance.

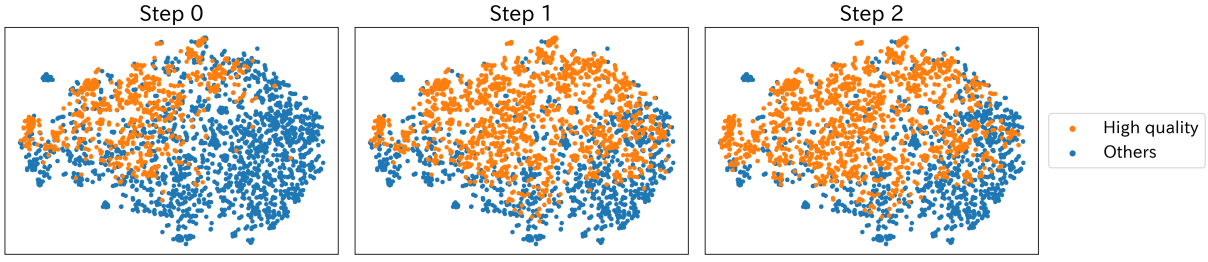
4) Step-wise Improvement Analysis

To analyze how the proposed method improves the corpus at each stage, we evaluate the corpora obtained after Step 0, Step 1, and Step 2, along with the TTS models trained on them.

As shown in the violin plots in Fig. 6, the overall speaker-wise UTMOS distribution improves with each step. From Step 0 to Step 1, the entire distribution shifts upward, indicating that quality filtering effectively removes low-quality utterances. From Step 1 to Step 2, the lower end of the distribution improves significantly, suggesting that the additional data acquired in Step 2 helps boost performance for underrepresented or low-quality speakers.

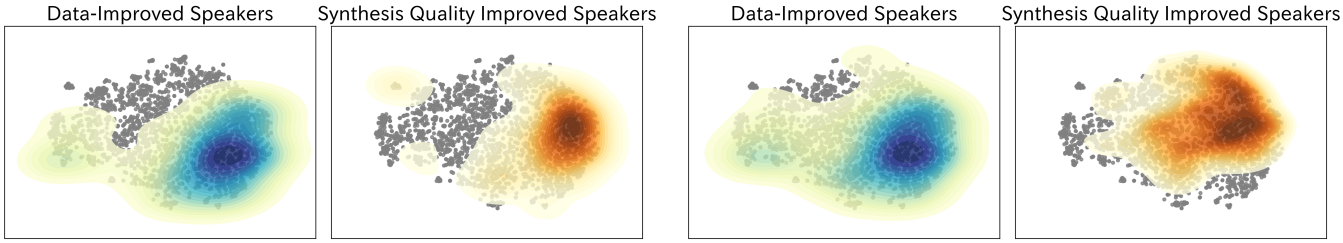


(a) FastSpeech 2



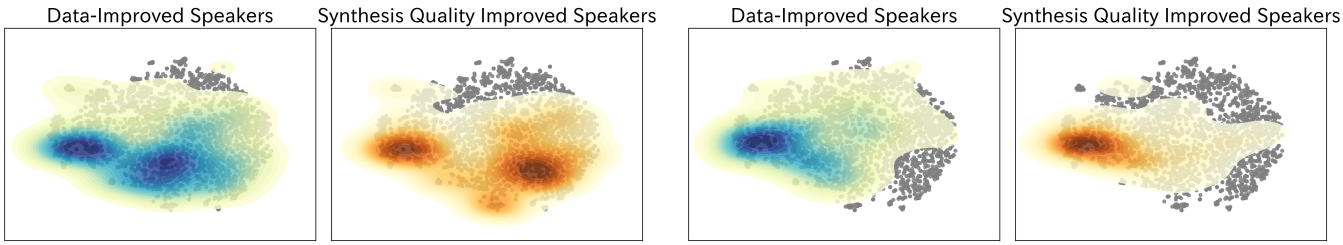
(b) Matcha-TTS

FIGURE 7: Scatter plot of high-quality speakers in the x-vector space at each step.



(a) Improvement from Step 1 (FastSpeech 2)

(b) Improvement from Step 1 (Matcha-TTS)



(c) Improvement from Step 2 (FastSpeech 2)

(d) Improvement from Step 2 (Matcha-TTS)

FIGURE 8: Kernel density estimation (KDE) visualization of speaker distributions in the x-vector space. These visualizations illustrate how data refinement and additional acquisition contribute to synthesis quality improvement.

To analyse how our method improves speaker coverage throughout the corpus construction process, we plot the distribution of high-quality speakers in the speaker embedding space. As described in Section III, a speaker is regarded as high-quality if their pseudo-MOS score exceeds the threshold τ_{hq}. Fig. 7 shows the scatter plot of these high-quality speakers projected onto the x-vector space using t-SNE [50]. These plots show that the distribution of high-quality speakers gradually spreads across the speaker space at each step.

To analyze how each step contributes to improving the corpus, we examined the relationship between training data updates and synthesis quality gains. We define the following two speaker sets:

- **Data-Improved Speakers:** Speakers whose training data were affected by corpus updates. In Step 1, these are speakers whose low-quality utterances were removed. In Step 2, they are speakers for whom new utterances were added.
- **Synthesis Quality Improved Speakers:** Speakers whose UTMOS crossed the threshold τ_{hq} as a result of

TABLE 3: Quantitative evaluation of our proposed active learning method under different data cleansing strategies. "w/o Cleansing" corresponds to no cleansing, "w/ Cleansing" to always applying cleansing, and "Switching" to switching-based data cleansing.

	FastSpeech 2			Matcha-TTS		
metric	UTMOS (↑)	# HQ-speaker (↑)	CER (↓)	UTMOS (↑)	# HQ-speaker (↑)	CER (↓)
w/o Cleansing	2.6019	2185	16.0223	2.3846	1577	22.9432
w/ Cleansing	2.6096	2167	16.4616	2.2620	1069	24.7111
Switching	**2.6261**	**2274**	**16.0137**	**2.4238**	**1739**	**21.7617**

TABLE 4: Quantitative evaluation of each corpus construction method under switching-based data cleansing.

	FastSpeech 2			Matcha-TTS		
metric	UTMOS (↑)	# HQ-speaker (↑)	CER (↓)	UTMOS (↑)	# HQ-speaker (↑)	CER (↓)
Baseline	2.5220	1888	16.7674	2.2200	934	25.2385
Core-set	2.4802	1704	16.3252	2.3883	1584	23.0718
Active Learning	**2.6261**	**2274**	**16.0137**	**2.4238**	**1739**	**21.7617**

the step. That is, their score was below the threshold before the step and above it afterward.

Fig. 8 visualizes the distribution of these speaker sets using kernel density estimation (KDE) over the *x*-vector space projected by t-SNE.[8]. Across both models and steps, we observe that "Data-Improved Speakers" and "Synthesis Quality Improved Speakers" occupy similar regions, suggesting that improvements in training data directly lead to gains in synthesis quality. Furthermore, comparing Step 1 and Step 2 shows that quality improvements occur in different regions of the speaker space. These results indicate the complementary nature of the two steps and highlight the model-aware effect of Step 2's targeted data acquisition.

C. EXTENSION TO SWITCHING-BASED DATA CLEANSING

To explore effective ways of incorporating data cleansing into our framework, we investigated the effectiveness and impact of switching-based data cleansing, an extension introduced in Section III-F.

1) Data Cleansing Method

As a data cleansing method, we adopted Demucs [51][9] as the speech enhancement backend. This model is designed for source separation, and is expected to reduce background music and other types of noise. As described in Section III-F, this enhancement is applied only when it improves the estimated training data quality.

2) Effectiveness of Switching-Based Data Cleansing

To evaluate the effectiveness of switching-based data cleansing, we compare our proposed method under three conditions: without any data cleansing ("w/o Cleansing"), with data cleansing applied uniformly to all samples ("w/ Cleansing"), and with data cleansing applied selectively based on the switching strategy described in Section III-F ("Switching").

[8]This is the same projection space as in Fig. 7.
[9]https://huggingface.co/spaces/Wataru/Miipher/tree/main

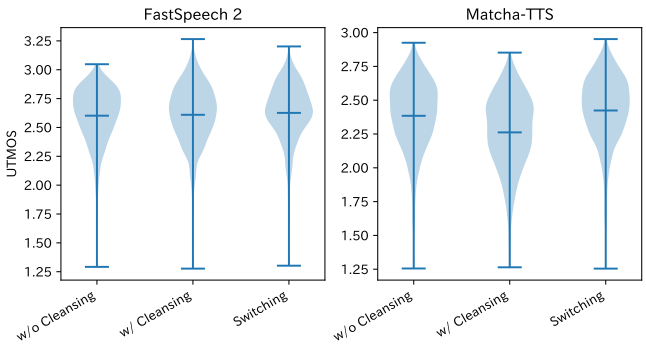


FIGURE 9: Distribution of speaker-wise UTMOS scores in our proposed active learning method under different data cleansing strategies.

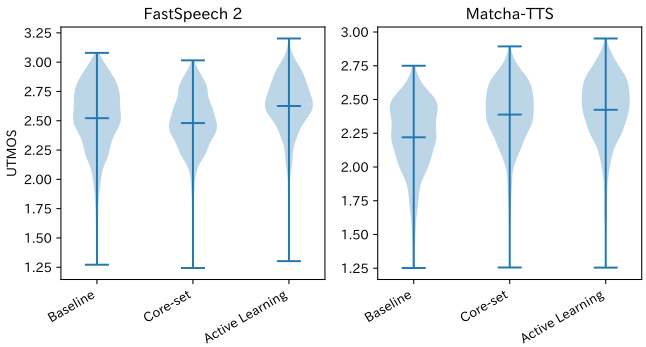


FIGURE 10: Distribution of speaker-wise UTMOS scores for each corpus construction method under switching-based data cleansing.

The results are shown in Fig. 9 and Table 3. "Switching" achieves the best performance across all metrics, demonstrating its effectiveness in both FastSpeech 2 and Matcha-TTS. Interestingly, the CER of "w/o Cleansing" is lower than that of "w/ Cleansing" in both models. This suggests that the speech enhancement model may suppress not only background noise but also phonetic cues that are important

for intelligibility, highlighting a trade-off between naturalness and recognition accuracy. In contrast, switching-based data cleansing applies speech enhancement only when it improves training data quality, allowing the model to benefit from noise reduction while preserving important speech characteristics. As a result, it improves synthesis quality without sacrificing intelligibility.

3) Comparison with Baseline Methods under Switching-Based Data Cleansing

To evaluate the effectiveness of different corpus construction strategies under a shared data cleansing condition, we compared ''Baseline'', ''Core-set Selection'', and our proposed ''Active Learning'' method, all incorporating the switching-based data cleansing strategy described in Section III-F.

The results are shown in Fig. 10 and Table 4. Our proposed method consistently achieved the best performance across all metrics and both models. These results demonstrate that our active learning approach remains effective even when combined with switching-based data cleansing, indicating that the active learning approach can effectively leverage the benefits of switching-based data cleansing.

V. CONCLUSION

This paper presented a method for data-efficient construction of multi-speaker TTS corpora from large-scale web data. Unlike conventional approaches that primarily focus on increasing data volume, our method selectively acquires necessary data based on an active learning framework, resulting in significantly improved data efficiency. Compared to core-set selection, which is model-agnostic and relies on manually designed diversity metrics, our method employs a feedback loop to make model-aware sampling decisions without relying on predefined evaluation criteria. This enables the corpus construction process to account for model difficulty and target underrepresented or low-performing regions in the data.

Experimental results with multiple TTS models demonstrated that our method consistently outperforms both random and core-set selection approaches across naturalness and intelligibility metrics under the same dataset size. In addition, we integrated a switching-based data cleansing mechanism that applies enhancement only when it improves estimated training data quality, leading to further improvements in the reliability of the constructed corpus.

Together, these components form a scalable and effective pipeline for model-guided corpus construction. Future work includes extending the proposed framework to other speech generation tasks beyond TTS.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI 22H03639, 24KJ0860 and Moonshot R&D Grant Number JPMJPS2011.

REFERENCES

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, ''Tacotron: Towards end-to-end speech synthesis,'' in *Interspeech 2017*, 2017, pp. 4006–4010.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, ''Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2018, pp. 4779–4783.

[3] J. Kim, S. Kim, J. Kong, and S. Yoon, ''Glow-TTS: A generative flow for text-to-speech via monotonic alignment search,'' *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 8067–8077, 2020.

[4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, ''FastSpeech 2: Fast and high-quality end-to-end text to speech,'' in *Proc. Int. Conf. Learn. Representation*, 2021.

[5] J. Kim, J. Kong, and J. Son, ''Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,'' in *Int. Conf. Mach. Learn.*, 2021, pp. 5530–5540.

[6] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. E. Henter, ''Matcha-TTS: A fast TTS architecture with conditional flow matching,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2024, pp. 11 341–11 345.

[7] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, E. Liu, Y. Leng, K. Song, S. Tang *et al.*, ''NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models,'' in *Int. Conf. Mach. Learn.*, 2024, pp. 22 605–22 623.

[8] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, ''NaturalSpeech: End-to-end text-to-speech synthesis with human-level quality,'' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4234–4245, 2024.

[9] K. Mitsui, Y. Hono, and K. Sawada, ''Towards human-like spoken dialogue generation between ai agents from written dialogue,'' *arXiv preprint arXiv:2310.01088*, 2023.

[10] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, ''SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities,'' in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2023.

[11] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, ''Moshi: a speech-text foundation model for real-time dialogue,'' *arXiv preprint arXiv:2410.00037*, 2024.

[12] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, ''GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio,'' in *Proc. Interspeech*, 2021, pp. 3670–3674.

[13] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, ''JTube-Speech: corpus of Japanese speech collected from YouTube for speech recognition and speaker verification,'' *arXiv preprint arXiv:2112.09323*, 2021.

[14] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, ''Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection,'' in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[15] W. Nakata, K. Seki, H. Yanaka, Y. Saito, S. Takamichi, and H. Saruwatari, ''J-CHAT: Japanese large-scale spoken dialogue corpus for spoken dialogue language modeling,'' *arXiv preprint arXiv:2407.15828*, 2024.

[16] H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, L. Liu, C. Yang, J. Li, P. Shi *et al.*, ''Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation,'' in *Proc. IEEE Spoken Lang. Technol. Workshop*. IEEE, 2024, pp. 885–890.

[17] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, ''Diversity-based core-set selection for text-to-speech with linguistic and acoustic features,'' in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 12 351–12 355.

[18] ——, ''TTSOps: A closed-loop corpus optimization framework for training multi-speaker tts models from dark data,'' *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[19] ——, ''Active learning for text-to-speech synthesis with informative sample collection,'' *arXiv preprint arXiv:2507.08319*, 2025.

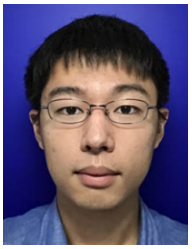
[20] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, ''Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit,'' 2016.

[21] K. Ito and L. Johnson, ''The lj speech dataset,'' https://keithito.com/LJ-Speech-Dataset/, 2017.

[22] R. Sonobe, S. Takamichi, and H. Saruwatari, ''JSUT corpus: Free large-scale japanese speech corpus for end-to-end speech synthesis,'' *arXiv preprint arXiv:1711.00354*, 2017.

[23] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, ''JVS corpus: Free Japanese multi-speaker voice corpus,'' *arXiv preprint arXiv:1908.06248*, 2019.

[24] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech*, 2019, pp. 1526–1530.
[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an asr corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.
[26] E. Cooper, A. Chang, Y. Levitan, and J. Hirschberg, "Data selection and adaptation for naturalness in hmm-based speech synthesis." in *Proc. Interspeech*, 2016, pp. 357–361.
[27] F.-Y. Kuo, "Data selection for improving naturalness of tts voices trained on small found corpuses," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 319–324.
[28] E. Tesfaye Biru, Y. Tofik Mohammed, D. Tofu, E. Cooper, and J. Hirschberg, "Subset selection, adaptation, gemination and prosody prediction for amharic text-to-speech synthesis," in *Proc. ISCA Speech Synthesis Workshop*, 2019.
[29] P. O. Gallegos, J. Williams, J. Rownicka, and S. King, "An unsupervised method to select a speaker subset from large multi-speaker speech synthesis datasets," in *Interspeech 2020*, 2020, pp. 1758–1762.
[30] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," *ACM Computing Surveys*, vol. 54, no. 9, pp. 1–40, 2021.
[31] A. Albalak, Y. Elazar, S. M. Xie, S. Longpre, N. Lambert, X. Wang, N. Muennighoff, B. Hou, L. Pan, H. Jeong *et al.*, "A survey on data selection for language models," *arXiv preprint arXiv:2402.16827*, 2024.
[32] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, and X. Hu, "Data-centric artificial intelligence: A survey," *ACM Computing Surveys*, vol. 57, no. 5, pp. 1–42, 2025.
[33] D. Wang and Y. Shang, "A new active labeling method for deep learning," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2014, pp. 112–119.
[34] D. Roth and K. Small, "Margin-based active learning for structured output spaces," in *European conference on machine learning*. Springer, 2006, pp. 413–424.
[35] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2372–2379.
[36] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *International Journal of Computer Vision*, vol. 113, no. 2, pp. 113–127, 2015.
[37] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489*, 2017.
[38] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1183–1192.
[39] D. Hakkani-Tür, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, 2002, pp. IV–3904.
[40] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 504–511, 2005.
[41] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech and Language*, vol. 24, no. 3, pp. 433–444, 2010.
[42] K. Malhotra, S. Bansal, and S. Ganapathy, "Active learning methods for low resource end-to-end speech recognition." in *Proc. Interspeech*, 2019, pp. 2215–2219.
[43] J. Luo, J. Wang, N. Cheng, and J. Xiao, "Loss prediction: End-to-end active learning approach for speech recognition," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2021, pp. 1–7.
[44] O. Kundacina, V. Vincan, and D. Miskovic, "Combining x-vectors and bayesian batch active learning: Two-stage active learning pipeline for speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, 2025.
[45] A. H. Azeemi, I. A. Qazi, and A. A. Raza, "A survey on data selection for efficient speech processing," *IEEE Access*, 2025.
[46] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5329–5333.
[47] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "CTC-segmentation of large corpora for german end-to-end speech recognition," in *Int. Conf. Speech and Computer*, 2020, pp. 267–278.
[48] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 17 022–17 033, 2020.
[49] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.
[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, 2008.
[51] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023.



KENTARO SEKI received the B.E. and the M.E. degree in 2022 and 2024, respectively from the University of Tokyo, Tokyo, Japan, where he is currently working toward the Ph.D. degree. His research interests include speech synthesis, speech enhancement, and audio signal processing. He is a student member of several organizations including IEEE, IEEE Signal Processing Society (SPS), and Acoustical Society of Japan (ASJ). He was the recipient of several awards and grants, including IEEE SPS Travel Grant for IEEE ICASSP 2023, Google Travel Grants for Students in East Asia, the best student presentation award of ASJ in 2022. He was also the recipient of the Research Fellowship for Young Scientists (DC1) from the Japan Society for the Promotion of Science (JSPS) in 2024.

YUKI SAITO received the Ph.D. degree in Information Science and Technology in 2021 from the Graduate School of Information Science and Technology, The University of Tokyo, Japan. His research interests include speech synthesis, voice conversion, and machine learning. He was the recipient of eight paper awards including the 2020 IEEE SPS Young Author Best Paper Award. He is a Member of the Acoustical Society of Japan, a Member of IEEE SPS, and a Member of Institute of Electronics, Information and Communication Engineers.

SHINNOSUKE TAKAMICHI received the B.E. degree from Nagaoka University of Technology, Nagaoka, Japan, in 2011, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan, in 2013 and 2016, respectively. He is currently a Lecturer at The University of Tokyo. He has received more than 20 paper/achievement awards including the 2020 IEEE Signal Processing Society Young Author Best Paper Award.

TAKAAKI SAEKI received the Ph.D. degree from the University of Tokyo, Japan, in 2024. His research interests include speech synthesis, speech representation learning, and machine learning. He is currently a Research Scientist at Google DeepMind, USA, where he works on multimodal language modeling. He was the recipient of several awards, including Yamashita SIG Research Award from the Information Processing Society of Japan, the Best Paper Award from the Institute of Electronics, Information and Communication Engineers (IEICE), Japan.

HIROSHI SARUWATARI received the B.E., M.E., and Ph.D. degrees from Nagoya University, Nagoya, Japan, in 1991, 1993, and 2000, respectively. In 1993, he joined SECOM IS Laboratory, Tokyo, Japan, and in 2000, Nara Institute of Science and Technology, Ikoma, Japan. Since 2014, he has been a Professor with The University of Tokyo, Tokyo, Japan. His research interests include statistical audio signal processing, blind source separation, and speech enhancement. He has put his research into the world's first commercially available independent-component-analysis-based BSS microphone in 2007. He was the recipient of several paper awards from IEICE in 2001 and 2006, from TAF in 2004, 2009, 2012, and 2018, from IEEE-IROS2005 in 2006, and from APSIPA in 2013 and 2018, and also the DOCOMO Mobile Science Award in 2011, Ichimura Award in 2013, Commendation for Science and Technology by the Minister of Education in 2015, Achievement Award from IEICE in 2017, and Hoko-Award in 2018. He has been professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ. Since 2018, he has been an APSIPA Distinguished Lecturer.