TTSOps: A Closed-Loop Corpus Optimization Framework for Training Multi-Speaker TTS Models from Dark Data

Kentaro Seki, *Student Member, IEEE* Shinnosuke Takamichi, *Member, IEEE* Saeki Takaaki, *Member, IEEE*
and Hiroshi Saruwatari, *Member, IEEE*

Abstract—This paper presents TTSOps, a fully automated closed-loop framework for constructing multi-speaker text-to-speech (TTS) systems from noisy, uncurated web-scale speech data, often referred to as "dark data," such as online videos. Conventional TTS training pipelines require well-curated corpora with high acoustic quality and accurate text–speech alignment, which severely limits scalability, speaker diversity, and real-world applicability. While recent studies have proposed acoustic-quality-based data selection techniques, they often overlook two critical aspects: (1) the inherent robustness of modern TTS models to noise, and (2) the potential contribution of perceptually low-quality yet informative samples. To address these issues, TTSOps introduces a data-centric training pipeline that integrates three core components: (1) automated data collection from dark data sources, (2) utterance-level dynamic selection of data cleansing methods based on training data quality, and (3) evaluation-in-the-loop data selection using automatically predicted mean opinion scores (MOS) to estimate each utterance's impact on model performance. Furthermore, TTSOps jointly optimizes the corpus and the TTS model in a closed-loop framework by dynamically adapting both data selection and data cleansing processes to the characteristics of the target TTS model. Extensive experiments on Japanese YouTube data demonstrate that TTSOps outperforms conventional acoustic-quality-based baselines in both the naturalness and speaker diversity of synthesized speech.

Index Terms—Text-to-speech synthesis, dark data, multi-speaker TTS, data-centric AI, data selection, data cleansing, closed-loop training.

I. INTRODUCTION

NEURAL text-to-speech (TTS) systems have made remarkable advances in recent years, enabling the synthesis of speech with human-level naturalness and intelligibility [1]–[4]. These advancements are largely driven by deep generative models and the availability of large-scale, high-quality corpora. Multi-speaker TTS models have further enabled flexible

At the time this study was conducted, all authors were with the Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo-ku, Tokyo 113-8656, Japan. Shinnosuke Takamichi is currently with the Graduate School of Information Science and Technology, The University of Tokyo, and the Graduate School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawa 223-8522, Japan.

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

speaker control [5]–[8]. However, they still rely heavily on curated datasets with clean audio and accurate text–speech alignments. This reliance significantly limits speaker diversity and scalability in real-world applications.

In contrast, vast amounts of speech data already exist on the internet in the form of user-generated content, such as YouTube videos. This so-called "dark data" [9], [10] offers rich diversity in speaker identities, speaking styles, and recording environments, making it a potentially valuable resource for building multi-speaker TTS systems. However, these real-world recordings are typically noisy, weakly labeled, and often misaligned with their transcripts, rendering them unsuitable for direct use in TTS training. As a result, effective *data selection* and *data cleansing* are essential for leveraging dark data in high-quality multi-speaker TTS synthesis.

Conventional data selection methods for TTS often rely on predetermined metrics, especially acoustic quality (e.g., signal-to-noise ratio), to identify and retain "clean" utterances [11]–[14]. While this approach helps eliminate overtly corrupted or distorted samples, it assumes that acoustic quality is a sufficient proxy for the sample's effectiveness in training a TTS model. However, recent advances in neural TTS have demonstrated that models exhibit increasing robustness to acoustic imperfections [15], [16], and that even perceptually low-quality samples can meaningfully contribute to model performance. As a result, selection strategies based solely on predetermined, model-agnostic criteria often fail to capture what truly matters for model training, leading to an increasing discrepancy between perceived acoustic quality and the actual effectiveness of data for model training—a property we refer to as *training data quality*.

In addition to data selection, most existing pipelines apply a single, uniform data cleansing method—such as denoising or restoration—to all training data, regardless of the individual characteristics of each utterance [17], [18]. Although such approaches aim to enhance overall audio quality, they fail to account for the inherent heterogeneity of dark data, where the optimal data cleansing strategy may vary significantly across samples. Some utterances may benefit from noise suppression, others from speech restoration, and some may already be clean—potentially degraded by unnecessary data cleansing. This mismatch underscores the need for *utterance-level adaptive data cleansing*, in which the optimal method is dynamically selected to maximize each sample's utility for model training.

To address these limitations, we propose *TTSOps*, a fully

automated training pipeline for multi-speaker TTS that integrates both data selection and data cleansing into a unified framework. The core idea of TTSOps is to assess the training data quality of each utterance—defined as its expected contribution to downstream model performance—and to use this assessment to inform two key decisions: (1) whether to include the utterance in training, and (2) which data cleansing method, if any, should be applied. By embedding these decisions within a evaluation-in-the-loop loop, TTSOps facilitates a model-aware and performance-driven approach to corpus construction from dark data.

To implement training-data-quality-based selection, TTSOps adopts an evaluation-in-the-loop strategy. For each utterance, a TTS model is trained using that utterance as part of the training data, followed by speech synthesis using the resulting model. The perceptual quality of the generated speech is then evaluated using a pre-trained pseudo MOS predictor. The resulting pseudo MOS score serves as a proxy for the utterance's contribution to model output quality, thereby quantifying its utility as training data. By applying this process to a pool of candidate utterances, TTSOps selectively constructs a training corpus that prioritizes samples with the highest expected impact on learning.

In addition to data selection, TTSOps incorporates data-wise data cleansing switching into the training-evaluation loop. For each utterance, multiple data cleansing variants—such as denoising, restoration, or no processing—are generated. Each variant is then evaluated using the same pseudo MOS-based procedure described earlier: a TTS model is trained using the variant, synthetic speech is generated, and its training data quality is estimated. The version with the highest estimated training data quality is selected and included in the final training corpus. This unified loop enables TTSOps to jointly optimize both data selection and data cleansing in a model-aware manner, dynamically adapting to the characteristics of each utterance and the target TTS model.

We validate the effectiveness of TTSOps through experiments on actual YouTube data—a representative source of dark data characterized by high variability and minimal curation. The results demonstrate that TTSOps consistently outperforms conventional approaches, including acoustic-quality-based data selection and fixed data cleansing pipelines, when evaluated in terms of predicted training data quality. In particular, TTSOps improves the pseudo MOS of synthesized speech, increases the number of high-quality speakers, and enhances speaker diversity—all aligned with the optimization target used within the training loop. These findings highlight the value of training data quality as a guiding signal for both data selection and data cleansing, and demonstrate the promise of data-centric corpus construction, especially in noisy and heterogeneous settings.

This work extends our previous conference paper [19], which introduced an evaluation-in-the-loop framework for training data selection. The prior study focused solely on utterance selection based on pseudo MOS scores, assuming a fixed data cleansing pipeline and evaluating the method under limited conditions (single TTS model, single corpus size). In this journal extension, we substantially broaden the methodology and experimental scope by introducing utterance-level adaptive data cleansing switching, validating the framework across multiple TTS architectures and corpus sizes, and conducting large-scale experiments on real-world YouTube data. These extensions demonstrate the generalizability and robustness of TTSOps, and further underscore the central role of training data quality in data-centric TTS system design.

II. RELATED WORK

A. Multi-speaker and Noise-Robust TTS

Recent advances in neural TTS have enabled models to synthesize speech for multiple speakers by leveraging speaker embeddings such as x-vectors [20]. Models such as Deep Voice 3 [5] and various zero-shot adaptation techniques [7], [8], [21], [22] have demonstrated the ability to generalize to unseen speakers with limited data. In parallel, noise-robust training approaches have been proposed to improve synthesis quality under degraded conditions. These include methods based on frame-level noise modeling [15], fine-grained noise control [23], and degradation-robust representation learning [16].

However, conventional TTS works tend to focus exclusively on improving model architectures, treating the training dataset as a fixed, pre-defined resource. In such approaches, data curation and preprocessing are performed as offline steps prior to model training, typically in a model-agnostic manner—despite growing evidence that the quality, diversity, and relevance of training data have a substantial impact on downstream performance. In contrast, our work adopts a data-centric perspective that treats corpus construction not as a static prerequisite, but as a dynamic, model-aware process that co-evolves with training.

B. Data Selection for TTS Corpus Construction

The performance of TTS models is highly sensitive to the quality and diversity of the training data. Since collecting well-curated corpora is expensive and time-consuming, several automatic data selection methods have been proposed to construct TTS corpora from speech resources such as ASR datasets [11]–[14]. These methods typically rely on predetermined metrics such as signal-to-noise ratio (SNR) or perceptual quality indicators like NISQA [24] to select "clean" utterances.

However, as described in Section II-A, recent studies have shown that modern TTS models are increasingly robust to noisy inputs, and that even low-quality samples can contribute positively to learning. In contrast to approaches that depend on predefined, model-agnostic quality metrics, our method evaluates each utterance based on its estimated training data quality—defined as its expected contribution to downstream model performance.

C. Data Cleansing for TTS from Noisy Speech

Speech enhancement has long been studied as a means to improve the perceptual quality of degraded speech signals [25]. Related areas include dereverberation [26], [27],

speech separation or extraction from mixtures containing non-speech components [28], [29], and analysis-by-synthesis methods that aim to remove distortions caused by recording devices [30]. Recently, deep generative approaches—often referred to as generative speech enhancement or speech restoration—have attracted increasing attention for their ability to produce natural-sounding outputs in subjective evaluations [31]–[33].

Recently, several TTS studies have leveraged these methods as data cleansing techniques [17], [18], applying them to noisy corpora prior to model training. However, most existing pipelines adopt a single, fixed data cleansing method across the entire dataset. This static strategy is often suboptimal, especially as more diverse and specialized restoration methods continue to emerge. Selecting one best method in advance for every utterance is not only impractical, but also unnecessary.

In contrast, our framework eliminates the need for manual selection by dynamically combining multiple data cleansing methods in a model-aware manner. This allows us to construct an optimal TTS corpus in a fully automated fashion, while effectively adapting to the characteristics of each utterance and model.

D. Automatic Assessment of Synthetic Speech Quality

Perceptual evaluation of synthetic speech quality has traditionally relied on subjective mean opinion score (MOS) tests, which are labor-intensive and difficult to scale. To overcome these limitations, a growing body of work has proposed automatic MOS prediction models that estimate perceived naturalness directly from audio signals without requiring reference signals or human raters. Early approaches, such as AutoMOS [34], Quality-Net [35], and MOSNet [36], utilized raw waveforms or spectral features as input to recurrent or convolutional neural networks. Recent advances leverage self-supervised learning (SSL) representations, which are learned from large-scale unlabeled speech corpora, to dramatically enhance generalization and robustness [37]–[39].

The rapid progress in this field has been catalyzed by the annual VoiceMOS Challenge [40]–[42]. Each edition introduces new tracks that test increasingly challenging generalization scenarios, such as out-of-domain (OOD) evaluation where models trained on English are used to assess Chinese speech [40], and zoomed-in comparisons where models must distinguish between highly natural-sounding synthetic speech samples [42]. These models are expected to continue evolving in both predictive accuracy and domain generalizability, facilitating broader applicability across languages, synthesis methods, and evaluation scenarios.

E. Data-Centric AI and Its Application to Speech

Data-centric AI emphasizes improving model performance by focusing on the quality, quantity and diversity of training data, rather than solely optimizing model architectures [43]. In the field of speech technology, this perspective has motivated the creation of large-scale corpora from web-based sources, including GigaSpeech [44], JTubeSpeech [45], and J-CHAT [46]. However, these corpora are typically constructed
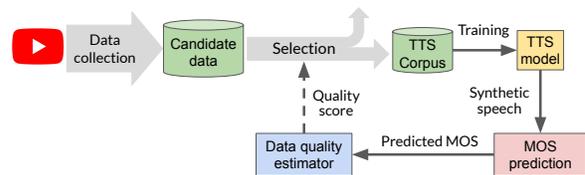

Fig. 1: Procedure of proposed data selection method based on training data quality. We evaluate training data quality of each utterance through loop of TTS model training and synthetic-speech evaluation. It finally builds a TTS corpus from dark data by utterance-wise filtering.

using static data filtering pipelines, without incorporating model-aware evaluation or adaptive preprocessing.

The proposed TTSOps framework operationalizes data-centric AI principles in the context of TTS. By integrating data cleansing, data selection, and training within a closed-loop process guided by training data quality, TTSOps offers a scalable and task-adaptive approach to corpus construction from noisy, real-world speech data.

III. DATA SELECTION BASED ON TRAINING DATA QUALITY

In this section, we describe a method for constructing a corpus from internet data using data selection based on training data quality. Our data selection is based on the *training data quality*, which we define as the expected contribution of each utterance to the performance of a given TTS model. Unlike conventional data selection approaches that rely on predetermined, model-agnostic criteria (e.g., acoustic quality scores), our method evaluates how much each utterance improves the perceptual quality of the synthesized speech generated by the model.

An overview of the proposed method is shown in Fig. 1. We perform data selection on speech data collected from YouTube using an evaluation-in-the-loop framework. In this loop, we train a TTS model with each candidate utterance and predict the perceptual naturalness of the resulting synthetic speech using a pseudo MOS model. Based on the pseudo MOS scores, we estimate the training data quality of each utterance. Finally, utterances with higher estimated quality are selected to construct the training corpus.

A. Data Collection and Pre-Screening

The first step is to collect text-audio pairs and pre-screen dark data to filter too low-quality data for TTS training. We collect a dataset from YouTube and combine data screening methods for ASR and ASV. We briefly describe each method below, but see the paper [45] for more detail.

1) Download Text-Audio Pairs from YouTube: Following the methodology in the previous study [45], we first compile a list of query terms, automatically extracted from two sources: (i) hyperlinked words appearing in Wikipedia articles, and (ii) trending keywords identified by Google Trends. Each term is then used to search YouTube, and only videos that include manually created subtitles are retained. For each selected
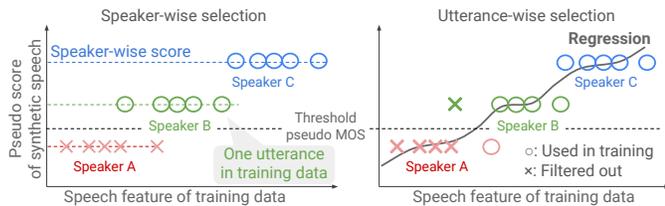
Fig. 2: Comparison of speaker-wise and utterance-wise selection. With regression, we filter out low-score utterances even if speaker's pseudo MOS is high

video, we download both the audio track and its corresponding subtitle file, thereby obtaining time-aligned text-audio pairs.

2) Pre-Screening via Text-Audio Alignment Accuracy: Since TTS training requires that speech segments be reasonably aligned with their transcriptions, we use the connectionist temporal classification (CTC) score [47] to assess the degree of alignment between audio and text. We apply CTC-based segmentation to split each audio file into utterances and compute alignment scores with respect to the corresponding YouTube subtitles. Utterances with notably low scores are filtered out to avoid cases that could severely disrupt the following training process.

3) Cleansing Based on Speaker Compactness: For TTS training, it is preferable to have multiple consistent utterances per speaker. To assess speaker consistency, we compute the variance of x-vectors [20] within each utterance group—defined as all utterances extracted from a single YouTube video. This variance serves as a proxy for the stability of speaker identity across the group. Groups exhibiting unusually high variance are filtered out, following the same principle as the alignment-based filtering.

After this step, each remaining group is assumed to contain utterances from a single, consistent speaker. It is worth noting that this method may inadvertently exclude some single-speaker groups with significant speaking-style variation, since x-vectors are sensitive to such variation even within the same speaker [48]. Nevertheless, we adopt this approach because our goal is to capture speaker diversity rather than style diversity.

B. Initial Training Using All the Candidate Data

We perform the initial training of the given TTS model using all the candidate data. To evaluate the quality of the training data for the desired TTS model, we consider it necessary to include a training step to capture the characteristics of the TTS model. In this step, we train a multi-speaker TTS model. Since the quality of synthesized speech is expected to vary across speakers in a multi-speaker TTS model, we infer the quality of the training data from the differences observed among speakers.

C. Evaluating Synthetic Speech Quality

We evaluate the quality of speech synthesized by the initially-trained TTS model. A simple evaluation method is to calculate the value of the loss function for each utterance during training (e.g., the distance between the ground-truth and the predicted features). However, the distance does not necessarily correspond to the perceptual quality of the synthesized speech [49], [50].

Since MOS directly reflects perceptual quality, subjective evaluation is a natural choice. However, it does not scale well when handling large volumes of data required for training. To address this, we use pseudo MOS predicted by an automatic MOS prediction model. These models have been reported to correlate well with human perceptual ratings [42], [51]. Therefore, we use the pseudo MOS score predicted by a pre-trained MOS prediction model as a proxy for naturalness in subjective evaluation.

This evaluation is used to determine the score of each training data, as described in the following Section III-D. In other words, this evaluation aims to estimate the difference in the effect of each data on the quality of the synthesized speech. The simplest method is to synthesize training data sentences and to filter out sentences with lower scores. However, this method uses different sentences among speakers. It is inappropriate because 1) the pseudo MOS score changes depending on the sentence to be synthesized [39] and 2) a sentence set greatly varies among speakers in dark data. Therefore, we evaluate the quality of synthesized speech for each speaker, using common sentences. In this way, it is possible to quantify the difference in synthetic speech quality of each speaker, without depending on the speakers' utterances. Therefore, we evaluate the quality of synthesized speech for each speaker using common sentences not included in the training data, and averages of the values are used for each speaker.

D. Quantifying Training Data Quality Score for Each Utterance

We filter the training data on the basis of the obtained speaker-wise pseudo MOS. The simplest method is to filter at the speaker level on the basis of the values, i.e., filtering out speakers with lower pseudo MOSs. However, since the data quality varies within the same speaker, filtering should be performed at the utterance level. In other words, we should filter out low-quality utterances even if speaker's pseudo MOS is high, and vice versa as shown in Fig. 2.

For this purpose, we train a regression model that predicts the speaker-wise pseudo MOS from each utterance in the training data. We assume that acoustically similar training data will achieve similar naturalness in the synthesized speech. Furthermore, we assume that a regression model predicts close values for acoustically similar data. These assumptions motivate us to use a regression model.

We train this regression model on all the pre-screened data. We confirm the effect of using a regression model in the experiment described in Section V-B1.

E. Training Data Selection and Re-Training

We evaluate the training data at the utterance level with the regression model. Then, utterances with lower scores are filtered out. Finally, we retrain the TTS model with the filtered data.
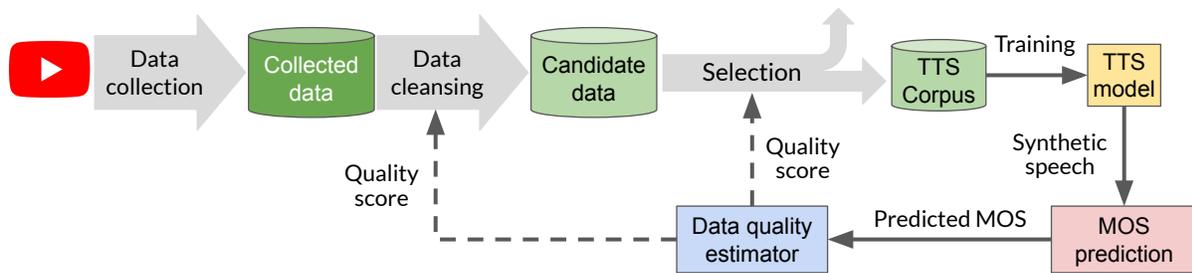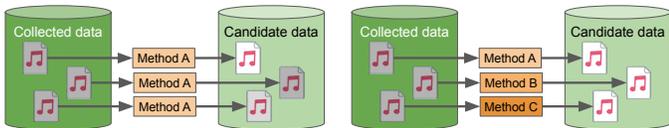
Fig. 3: Procedure of TTSOps. We obtain dark data from YouTube and evaluate each utterance through loop of TTS model training and synthetic-speech evaluation. It finally builds TTS corpus from dark data by utterance-wise filtering.

(a) In conventional methods, the same data cleansing method is uniformly applied to all data.

(b) In the proposed method, different data cleansing methods are applied to each data point.

Fig. 4: Comparison between conventional data cleansing procedures and the proposed data cleansing procedure.
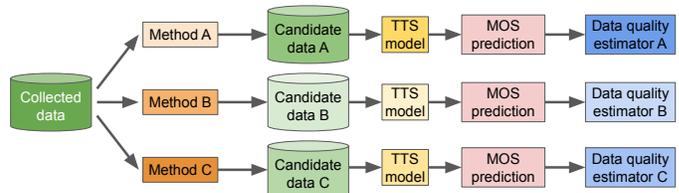

Fig. 5: A method for evaluating the quality of training data for each data cleansing method. The training and evaluation loop is applied to datasets obtained by uniformly applying each data cleansing technique, and the training data quality is assessed.

IV. TTSOps: Automatic Optimization Method of TTS Training Procedure

In this section, we propose TTSOps, a unified training pipeline that extends the methodology presented in Section III by integrating data cleansing strategies into the evaluation-in-the-loop framework. As illustrated in Fig. 3, TTSOps operates on speech data collected from the internet and estimates the training data quality—defined as each utterance's expected contribution to the performance of a TTS model—through a closed-loop process that combines data cleansing, data selection, model training, and pseudo MOS evaluation.

Unlike conventional pipelines that apply a fixed data cleansing method across the entire dataset, TTSOps performs utterance-wise adaptive switching of cleansing techniques, selecting the one that maximally improves the training data quality for each utterance as shown in Fig. 4. Based on this adaptive cleansing, the framework performs data selection guided by the estimated training data quality, ultimately constructing a high-quality corpus tailored to the characteristics of the target TTS model.

A. Data Collection

We collected the dataset following the method described in Section III-A. Specifically, we extracted text-audio pairs from YouTube videos and applied a pre-screening process based on CTC alignment scores and speaker compactness using x-vectors. This pre-screening was intended to prevent the initial training phase from failing by removing utterances with extremely low quality. For further details, refer to Section III-A.

B. Training Data Quality Evaluation

The method for evaluating the quality of training data for each data cleansing method is illustrated in Fig. 5. In the pro-

posed approach, each data cleansing method is uniformly applied to all data, and the quality of the training data is assessed through the training-evaluation loop described Section III. This method involves training a TTS model using candidate data, evaluating the quality of the synthesized speech generated by the trained TTS model for each speaker, and learning a regression model to predict the quality of synthesized speech based on the training data. By predicting the synthesis quality of a TTS model trained with specific data, this framework evaluates the quality of the training data.

This method leverages the fact that, even with a single multi-speaker TTS model, the quality of synthesized speech varies among speakers. Therefore, it evaluates the quality of each dataset through a single instance of model training. However, when combining multiple data cleansing methods, as in this study, a single model training instance does not allow for separate evaluations of data cleansing methods. To address this limitation, this study evaluates the quality of training data separately for each data cleansing method.

C. Corpus Determination

To determine final corpus, data cleansing and selection are performed using the method shown in Fig. 6, based on training data quality evaluation. In the data cleansing process, the data cleansing method that results in the highest training data quality is selected and applied for each dataset. Furthermore, since the training data quality has already been evaluated using the method described in Section III, data selection is conducted accordingly.

Here, by saving the preprocessed data, the need to repeat data cleansing is eliminated, enabling data cleansing selection simply by gathering the relevant files. Specifically, the training
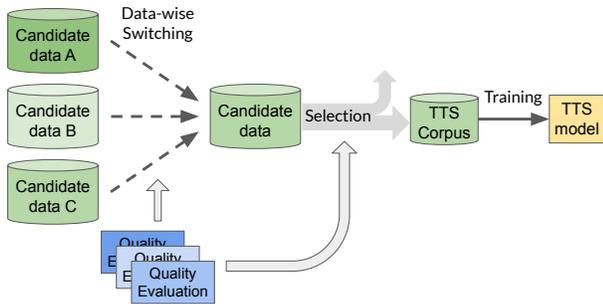
Fig. 6: Corpus determination procedure. Preprocessing methods are switched data-wise based on training data quality, and the final corpus is determined through data selection based on the training data quality of each dataset.

data quality obtained from the adopted data cleansing method is directly regarded as the quality of each dataset. Data with higher quality is selected in descending order to construct the TTS corpus. The constructed corpus is then used to train the TTS model.

V. EXPERIMENTAL EVALUATION OF DATA SELECTION

A. Experimental Setting

1) Dataset: We followed JTubeSpeech [45] scripts[1] to obtain dark data from YouTube; the amount was approximately $3,500$ hours. Pre-screening with a CTC threshold of -0.3 and speaker compactness threshold of $[1,7]$[2], we obtained approximately 66 hours ($60,000$ Japanese utterances) of $2,719$ speakers as the pre-screened data. The sentences used for calculating the pseudo MOS were 100 phoneme-balanced sentences from the JVS corpus [52]. The test data used to evaluate the finally trained TTS models was 324 sentences from the ITA corpus[3]. There was no overlap in text among the pre-screened data, sentences for the pseudo MOS, and test data.

2) Model and Training: We trained an acoustic model and combined it with the pre-trained HiFi-GAN vocoder [53] UNIVERSAL_V1[4] as a multi-speaker TTS system. We used two acoustic models, FastSpeech 2 [54] and GlowTTS [55]. We followed the model size and hyperparameters of the open-sourced implementation[56] except for the speaker representation. Instead of the one-hot speaker representation implemented in the repository, we used an open-sourced x-vector extractor[7], and a 512-dimensional x-vector was used to condition the TTS model. The x-vector was added to the output of the FastSpeech 2 encoder via a 512-by-256 linear layer. The x-vector was averaged for each speaker; one x-vector corresponded to one unique speaker. The TTS model was pre-trained using $10,000$ utterances from the JVS corpus [52], the

100-speaker Japanese TTS corpus. We performed 300k steps with a batch size of 16 in this pre-training. TTS training in this paper started from this pre-trained model with 100k steps with a batch size of 16.

We used a pre-trained UTMOS [39] strong learner[8] to obtain a five-scale pseudo MOS on naturalness from synthetic speech. The regression model for predicting the pseudo MOS from the training data was 1-layer 256-unit bi-directional long short-term memory [56], followed by a linear layer, ReLU activation, and another linear layer. We used frame-level self-supervised learning (SSL) features[9] obtained with a wav2vec 2.0 model [57][10], as the input. The frame-level outputs were aggregated to predict the pseudo MOS. The number of training steps, minibatch size, optimizer, and training objective were 10k, 12, Adam [58] with a learning rate of 0.0001, and mean squared error, respectively.

3) Compared Methods: We compared the following data selection methods.

- **Unselected**: All the pre-screened data was used for the TTS training; the training data size was approximately $60,000$ utterances.
- **Acoustic-quality (utterance-wise)**: The training data was selected in terms of the acoustic quality of the data. We used NISQA [24], a recent deep learning-based model to predict scores on naturalness, noisiness, coloration, discontinuity, and loudness of the speech data. Each score takes $[1,5]$. We chose multiple threshold $\theta = 4.0, 3.5, 3.0, 2.5$. That is data for which all the scores were higher than θ were selected. For $\theta = 4.0, 3.5, 3.0, 2.5$, the TTS training data size n was $5128, 12172, 20390$ and 31038, respectively.
- **Ours-Utt (evaluation-in-the-loop utterance-wise selection)**: Our evaluation-in-the-loop data selection. For each data, we estimate the speech quality synthesized by the TTS model from an initial training with pre-screened data. The threshold for selecting training data was set to have the resulting training data be the same in size as "Acoustic-quality".
- **Ours-Spk (evaluation-in-the-loop speaker-wise selection)**: Our data selection, but the data selection was performed per speaker as described in Section III-C. The threshold for selecting training data was set to have the resulting in the size of the training data be almost the same as "Acoustic-quality".

4) Evaluation: We evaluated the selection methods to clarify the following:

- **Does our method obtain more "high-quality speakers?": pseudo MOS comparison.** Our TTS model is expected to reproduce voices for a higher number of speakers. We define a "high-quality speaker" as a speaker with a higher pseudo MOS than the threshold. We in advance trained an high-quality multi-speaker TTS model using the JVS corpus [52] and calculated the speaker-wise pseudo MOS scores. We set the lowest score among the JVS speakers

---

[1]https://github.com/sarulab-speech/jtubespeech

[2]These values are the same as the experiments in the JTubeSpeech paper [45].

[3]https://github.com/mmorise/ita-corpus

[4]https://github.com/jik876/hifi-gan

[5]FastSpeech 2: https://github.com/Wataru-Nakata/FastSpeech2-JSUT

[6]GlowTTS: https://github.com/trgkpc/glow-tts_JSUT

[7]https://github.com/sarulab-speech/xvector_jtubespeech

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

[9]We compared the SSL features and the NISQA [24] features (used in the baseline) in the preliminary evaluation. The result demonstrated that the SSL features performed better.

[10]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec
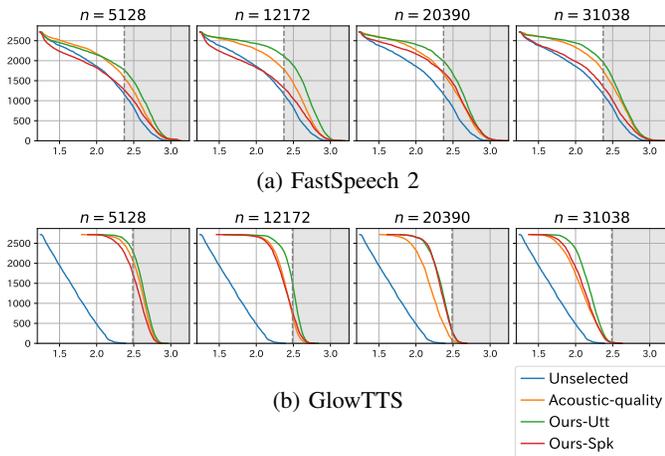
(a) FastSpeech 2

(b) GlowTTS

Fig. 7: Cumulative histograms of pseudo MOS. Y-axis value indicates number of speakers with higher score than x-axis value. The shaded area corresponds to high-quality speakers.

TABLE I: Number of overall, seen and unseen speakers. Values of each cell are number of high-quality speakers, all speakers, and ratio of two values, respectively.

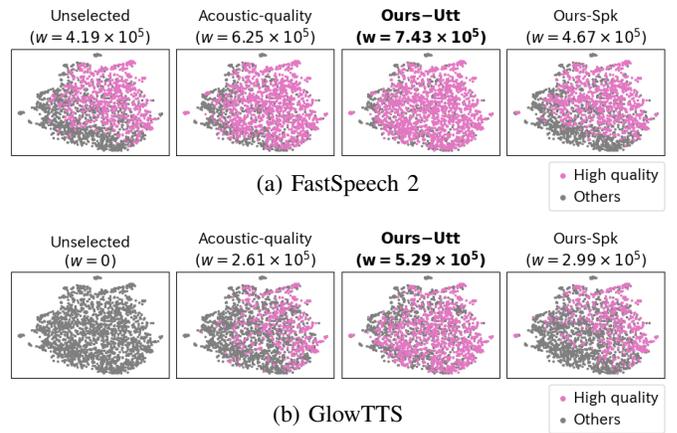| n	Method	FastSpeech 2	GlowTTS
58500	Unselected	1157 (42.6%)	0 (0.0%)
5128	Acoustic-quality	1572 (57.8%)	2068 (76.1%)
	Ours-Utt	**1764** (**64.9%**)	**2305** (**84.8%**)
	Ours-Spk	1274 (46.9%)	1735 (63.8%)
12172	Acoustic-quality	1786 (65.7%)	706 (26.0%)
	Ours-Utt	**2114** (**77.7%**)	**1510** (**55.5%**)
	Ours-Spk	1321 (48.6%)	798 (29.3%)
20390	Acoustic-quality	1661 (61.1%)	90 (3.3%)
	Ours-Utt	**1989** (**73.2%**)	**305** (**11.2%**)
	Ours-Spk	1712 (63.0%)	297 (10.9%)
31038	Acoustic-quality	1748 (64.3%)	19 (0.7%)
	Ours-Utt	**1955** (**71.9%**)	**29** (**1.1%**)
	Ours-Spk	1347 (49.5%)	25 (0.9%)

(a) FastSpeech 2

(b) GlowTTS

Fig. 8: Distributions of high-quality speakers by each data selection method ($n = 12172$). Higher w indicates more diversity.

as the threshold[11]. Speakers with a higher score than the threshold were considered to be high-quality speakers. For each data selection method, we calculated 1) the distribution of the pseudo MOSs for synthetic speech by the trained TTS model and 2) the number of high-quality speakers.

- **Does our method increase speaker variation?** We evaluated whether our method obtains diverse (i.e., sounding different) high-quality speakers. To quantify the speaker variation, we calculated the cost of a Euclidean minimum spanning tree [59] of x-vectors of the high-quality speakers. The calculation is similar to the g2g (median of the distances to the nearest x-vector) [60], but we used summation instead of a representative value (i.e., median) because our purpose is to evaluate how widely speakers spread.
- **Does synthetic speech of so-called "high-quality speakers" truly sound natural?: actual MOS evaluation.** We evaluated whether our data selection based on pseudo MOS is truly effective in synthesizing perceptually natural speech. To this end, we conducted a subjective evaluation of synthetic speech quality using different data selection methods, and investigated their performance in relation to pseudo MOS.

Note that seen and unseen speakers are different among the data selection methods. All the speakers were seen ones for "Unselected," but only parts of them were seen in the other methods. Also, seen speakers were different among "Acoustic-quality," "Ours-Utt," and "Ours-Spk." Speakers not used for training were considered to be unseen speakers for each method. Unless otherwise noted, we describe results aggregating those of both seen and unseen speakers.

B. Results

1) Number of High-Quality Speakers: Fig. 7 is a cumulative histogram of the pseudo MOSs. Our method had the highest values among the methods, demonstrating that a TTS

---

[11]The JVS corpus was constructed in a well-designed environment, and we confirmed that it was not an outlier.

---

model trained on our data selection can synthesize multi-speaker voices with higher quality than the other methods. The numbers of high-quality speakers for "Unselected," "Acoustic-quality," "Ours-Utt," and "Ours-Spk" were $924, 1737, 1942$, and 1367, respectively. We see that the proposed method increased the number of high-quality speakers, compared with the other methods. Specifically, the increment was approximately 1.2 times from that of "Acoustic-quality." From these results, we can say that the proposed method work better than the conventional methods for the purpose of increasing the speaker variety of the multi-speaker TTS model. Note that this TTS corpus is much larger than the previous Japanese multi-TTS corpus (JVS) composed of 100 speakers.

In addition, comparing "Ours-Utt" and "Ours-Spk", the proposed method was significantly better in terms of both the pseudo MOS distribution and the number of the high-quality speaker. This indicates that utterance-wise selection significantly contributes to enhancing the performance, rather than speaker-wise selection.

2) Evaluation of Speaker Variation: To validate the effectiveness of dark data, Fig. 8 shows the distributions of
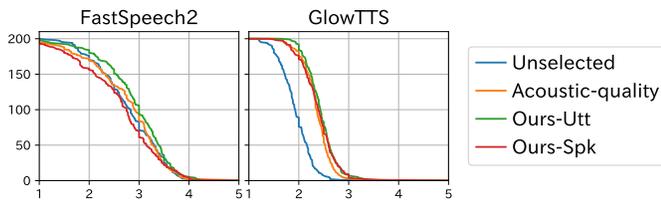
Fig. 9: Cumulative histograms of actual MOS. Y-axis value indicates number of speakers with higher score than x-axis value. The shaded area corresponds to high-quality speakers.
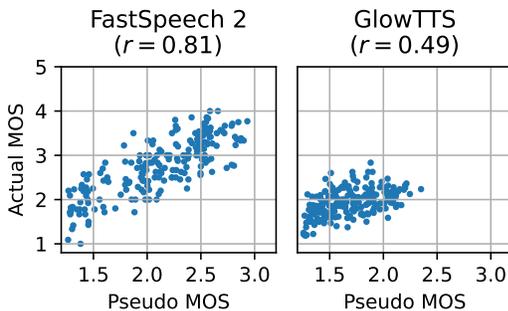

Fig. 10: Scatter plots showing the relationship between pseudo MOS and actual MOS for each speaker. Although a consistent correlation is observed in both models, the correlation is weaker for GlowTTS, leading to reduced effectiveness of the proposed method.

the high-quality speakers and speaker variation scores w. Both qualitatively and quantitatively, "Ours-Utt" increased the speaker variation compared with the other methods, indicating that our method contributes to speaker variation.

3) Validation on Actual MOS: We conducted a five-point MOS test to assess the naturalness of synthetic speech. A total of 500 listeners participated, each listening to 20 samples. For both FastSpeech 2 and GlowTTS, we evaluated the synthetic speech generated by TTS models trained on data selected with each data selection method. For "Acoustic-quality," "Ours-Utt," and "Ours-Spk," the data set size n was set to $12,172$. To reduce the evaluation cost, we randomly sampled 200 speakers from the total of $2,719$ for evaluation. The actual MOS scores were then aggregated for each speaker.

Fig. 9 shows the evaluation results. In both models, "Ours-Utt" achieved the highest scores among all methods, indicating that the proposed method is more effective than conventional approaches even in terms of perceptual speech quality.

However, the improvement observed in GlowTTS was less pronounced than in FastSpeech 2. To investigate the cause, we examined the relationship between pseudo MOS and actual MOS on "Unselected." Fig. 10 presents the scatter plots and the correlation coefficients r. While FastSpeech 2 showed a high correlation, GlowTTS exhibited lower correlation, possibly because the narrower distribution range of pseudo MOS in GlowTTS amplified the impact of prediction errors. Since our proposed method estimates training data quality using the pseudo MOS scores from "Unselected," its effectiveness in actual MOS is likely reduced in the case of GlowTTS, where

pseudo MOS shows a lower correlation with actual MOS. The limitations of the proposed method arising from the accuracy of pseudo MOS are discussed in Section VII-C.

VI. EXPERIMENTAL EVALUATION OF TTSOPS

A. Experimental Setting

1) Dataset, Model and Training: We utilized the same data collected from YouTube as described in Section V-A1. Also, we utilized the same model and the same training method as described in Section V-A2.

2) Compared Methods: As data cleansing methods, the following were considered:

- **Uncleansed**: The data was used as is. This method is expected to be suitable for high-quality audio, such as studio-recorded speech.
- **Demucs**: As a preprocessing method, we applied a pre-trained speech enhancement model, Demucs [29], to all data[12]. Demucs is a type of source separation model trained to remove background music. Therefore, it is expected to reduce additive noise and improve data quality.
- **Miipher**: As a preprocessing method, we applied a pre-trained audio restoration model, Miipher [33], to all data[13]. This model is trained to restore degraded audio. Consequently, it is expected to reduce environmental distortions caused by recording devices or reverberation, thereby enhancing data quality.
- **Switching (ours)**: For each dataset, one of "Uncleansed," "Demucs," or "Miipher" was selected and applied. The selection was based on training data quality, aiming to choose the method that maximized overall quality.

Furthermore, to validate the proposed metric, training data quality, we investigated an alternative approach that uses acoustic quality evaluation instead of training data quality. In this approach, the preprocessing method that achieves the highest acoustic quality is selected, and the corpus is constructed by selecting data in descending order of acoustic quality. The acoustic quality is evaluated using the deep learning model NISQA [24], where the minimum value among five metrics is used as the evaluation score.

3) Evaluation: We performed the pseudo MOS evaluation according to the procedure described in Section V-A4, aiming to answer the following research questions:

- **Is the "Switching" framework effective in improving model performance?**
- **Which criterion is more suitable for determining the data cleansing method: acoustic quality or training data quality?**

This evaluation was conducted by calculating 1) the distribution of the pseudo MOSs for synthetic speech generated by the trained TTS models, and 2) the number of high-quality speakers, as defined in Section V-A4.

In addition, to evaluate whether the synthetic speech of high-quality speakers truly sounded natural, we conducted an actual MOS test via human evaluation, following the

---

[12]https://huggingface.co/spaces/Wataru/Miipher/tree/main
[13]https://github.com/facebookresearch/demucs

(a) Acoustic quality-based, FastSpeech 2

(b) Acoustic quality-based, GlowTTS

(c) Training data quality-based, FastSpeech 2
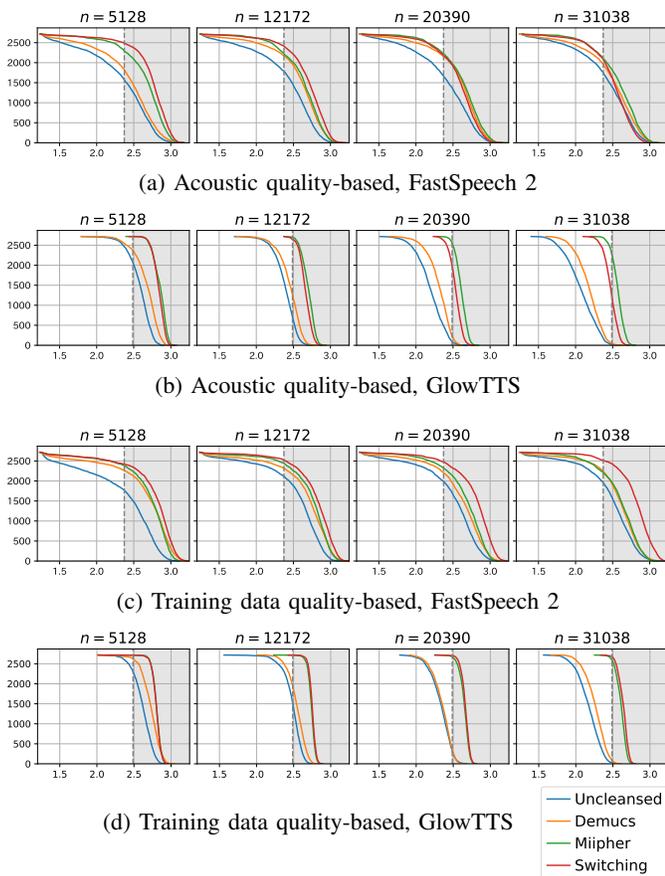
(d) Training data quality-based, GlowTTS

Fig. 11: Cumulative histograms of pseudo MOS. Y-axis value indicates number of speakers with higher score than x-axis value. The shaded area corresponds to high-quality speakers.

procedure defined in Section V-B3. This allowed us to verify the effectiveness of the "Switching" strategy not only in terms of pseudo MOS but also from a perceptual standpoint.

Finally, we analyzed the proportion of data samples that were processed by each cleansing method in the "Switching" scenario. This analysis reveals how each method was selected depending on data characteristics and model robustness, providing further insight into the behavior of the proposed framework.

B. Result

1) Pseudo MOS Evaluation: Fig. 11 shows the cumulative histograms of pseudo MOS for different data cleansing methods and selection metrics. When the training data quality is used as the selection metric, the proposed "Switching" method consistently outperforms the baseline methods across all corpus sizes and model architectures. This consistent superiority demonstrates the effectiveness and robustness of our approach.

In contrast, when acoustic quality is used as the selection metric, the benefit of "Switching" becomes less reliable. In some conditions, such as $n = 31038$, "Switching" even underperforms compared to applying "Miipher" uniformly. This inconsistency indicates that acoustic quality does not always reflect the actual utility of data for training TTS models.

TABLE II: Number of overall, seen and unseen speakers. Values of each cell are number of high-quality speakers, all speakers, and ratio of two values, respectively.

(a) Acoustic Quality based

n	Method	FastSpeech 2	GlowTTS
5128	Uncleansed	1572 (57.8%)	2068 (76.1%)
	Demucs	1828 (67.2%)	2371 (87.2%)
	Miipher	2290 (84.2%)	**2718 (100.0%)**
	Switching	**2484 (91.4%)**	2717 (99.9%)
12172	Uncleansed	1786 (65.7%)	706 (26.0%)
	Demucs	2177 (80.1%)	1257 (46.2%)
	Miipher	2224 (81.8%)	**2700 (99.3%)**
	Switching	**2413 (88.7%)**	2650 (97.5%)
20390	Uncleansed	1661 (61.1%)	90 (3.3%)
	Demucs	2155 (79.3%)	236 (8.7%)
	Miipher	**2223 (81.8%)**	**2551 (93.8%)**
	Switching	2182 (80.3%)	1989 (73.2%)
31038	Uncleansed	1748 (64.3%)	19 (0.7%)
	Demucs	1958 (72.0%)	46 (1.7%)
	Miipher	**2103 (77.3%)**	**2297 (84.5%)**
	Switching	2078 (76.4%)	1238 (45.5%)

(b) Training data Quality based

| n	Method	FastSpeech 2	GlowTTS
5128	Uncleansed	1764 (64.9%)	2305 (84.8%)
	Demucs	2263 (83.2%)	2623 (96.5%)
	Miipher	2379 (87.5%)	**2716 (99.9%)**
	Switching	**2420 (89.0%)**	2715 (99.9%)
12172	Uncleansed	2114 (77.7%)	1510 (55.5%)
	Demucs	2320 (85.3%)	1947 (71.6%)
	Miipher	2437 (89.6%)	2715 (99.9%)
	Switching	**2524 (92.8%)**	**2718 (100.0%)**
20390	Uncleansed	1989 (73.2%)	305 (11.2%)
	Demucs	2205 (81.1%)	335 (12.3%)
	Miipher	2318 (85.3%)	2670 (98.2%)
	Switching	**2473 (91.0%)**	**2703 (99.4%)**
31038	Uncleansed	1955 (71.9%)	29 (1.1%)
	Demucs	2241 (82.4%)	91 (3.3%)
	Miipher	2204 (81.1%)	2590 (95.3%)
	Switching	**2516 (92.5%)**	**2677 (98.5%)**
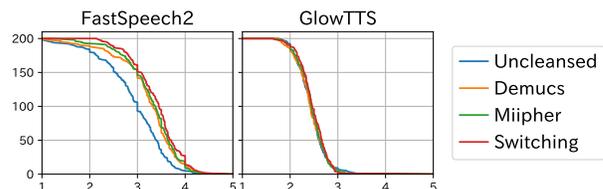
Fig. 12: Cumulative histograms of actual MOS. Y-axis value indicates number of speakers with higher score than x-axis value. The shaded area corresponds to high-quality speakers.

These observations support our central claim: *Training data quality*—defined as the expected contribution of each utterance to downstream model performance—is a more reliable and task-aware criterion for data selection than predetermined criterion such as acoustic quality.

2) Validation on Actual MOS: We conducted a five-point MOS test following the procedure described in Section V-B3. In this evaluation, we compared "Uncleansed," "Demucs," "Miipher," and "Switching," all based on the training-data-based selection strategy with a fixed data set size of $n =$
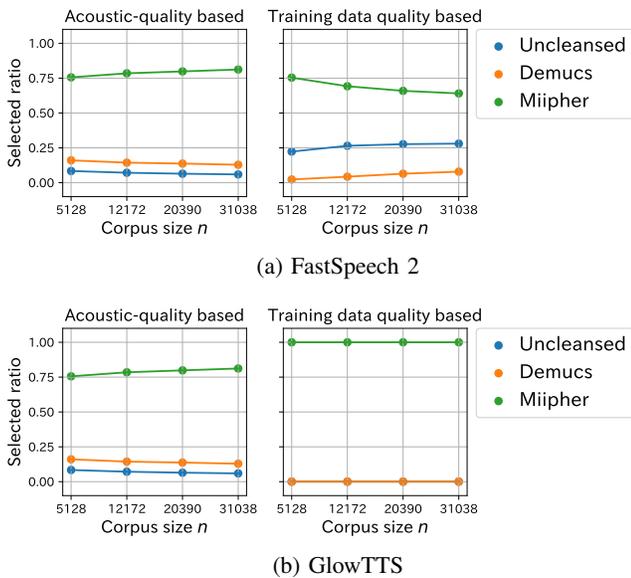
(a) FastSpeech 2

(b) GlowTTS

Fig. 13: The proportion of each preprocessing method selected in the "Switching" approach.

12172.

Fig. 12 shows the results. For FastSpeech 2, among the individual data cleansing methods, "Miipher" performed the best, followed by "Demucs" and then "Uncleansed." The proposed method, which dynamically leverages the strengths of each, achieved the best overall performance. This result indicates that the proposed method is more effective than the conventional methods even in terms of perceptual speech quality.

On the other hand, for GlowTTS, "Uncleansed," "Demucs," and "Miipher" exhibited similar performance in the subjective evaluation, and the proposed method did not yield additional improvements. This suggests that further improvements in preprocessing techniques may be needed to fully realize the benefits of the "Switching" approach in models such as GlowTTS.

3) Investigation of the Switching Method: Fig. 13 presents the selection rates of each preprocessing method used in the "Switching" strategy, across various dataset sizes and model architectures.

Since fewer than 10% of the collected utterances achieved a NISQA score above 4.0, indicating that the majority of the data were acoustically degraded, "Miipher" was selected for most samples due to its strong restoration performance.

However, for FastSpeech 2, over 20% of the utterances were selected without any preprocessing, despite being sourced from noisy YouTube recordings. This suggests that FastSpeech 2 has a certain degree of robustness to noise and may benefit from preserving relatively clean segments without risking degradation from unnecessary processing.

In contrast, GlowTTS consistently selected "Miipher" across all dataset sizes. This preference can be attributed to GlowTTS's reliance on internal alignment learning, which makes it more sensitive to noise and alignment errors. Consequently, aggressive restoration becomes necessary to stabilize

training, leading to an almost exclusive selection of Miipher.

These findings highlight that the optimal data cleansing strategy is highly dependent on the architecture and noise tolerance of the target TTS model. Notably, the selection patterns remained stable across different corpus sizes, underscoring the adaptability and consistency of the "Switching" framework in meeting model-specific requirements.

VII. DISCUSSION

A. Computation Cost

One practical concern regarding our framework is the computational cost associated with the training-evaluation loops. As described in Section III, our data selection method requires two TTS training cycles: one to estimate training data quality using pseudo MOS scores, and another for final training with the constructed corpus. In the case of TTSOps (Section IV), the number of training runs increases to $n + 1$, where n is the number of candidate preprocessing methods.

While this may seem costly at first glance, it reflects the reality of TTS system development. In practical scenarios, model training is rarely a one-shot process: developers typically go through many rounds of data cleaning, retraining, and hyperparameter tuning to achieve satisfactory results. TTSOps automates this trial-and-error process by integrating data selection and data cleansing into a closed-loop framework, significantly reducing the need for manual intervention. As a result, the overall development cycle can become more efficient—even if each run of the framework incurs some additional computation.

As a reference, under our experimental conditions using a single NVIDIA RTX 3090 GPU, each training run of FastSpeech 2 within the TTSOps framework required approximately 2 hours. Given that the framework involved four training runs in total—corresponding to three candidate preprocessing methods and one final training—the overall training time amounted to approximately 8 hours.

B. Scalability and Language-Agnostic Design

A key strength of the proposed framework lies in its fully automated and language-agnostic design. Because our method does not rely on language-specific heuristics or human intervention, it can be easily scaled to larger corpora and adapted to various languages. In our experiments, we successfully applied the method to a large-scale Japanese dataset, demonstrating its practicality under real-world conditions. In particular, the MOS prediction model used in our method was trained on English and Chinese data. Achieving promising results on Japanese—an out-of-domain language for the model—serves as experimental evidence for the language-independence of our approach. Further validation across different languages and datasets remains an important direction for future work.

C. Limitations of Pseudo MOS and Future Prospects

While the correlation between pseudo MOS and actual MOS was limited for GlowTTS, this issue is expected to diminish as

MOS prediction models continue to advance. Indeed, international competitions such as the VoiceMOS Challenge are held annually [40]–[42], promoting rapid progress in prediction accuracy. These developments have enabled more fine-grained evaluations, such as distinguishing subtle differences in high-MOS regions and conducting zero-shot assessments.

A notable feature of our TTSOps framework is its flexibility: the evaluation-in-the-loop mechanism is not restricted to pseudo MOS. It can be extended to incorporate alternative evaluation metrics (e.g., prosody, emotion), thereby broadening its applicability according to the specific training objectives.

D. Importance of Switching Data Cleansing Method

Although "Miipher" was selected for the majority of utterances in our experiments, this does not suggest that applying Miipher uniformly constitutes an optimal strategy. Indeed, the documentation for LibriTTS-R [17] reports that approximately 5% of samples failed to be successfully restored using Miipher[14]. Our switching strategy enables the automatic exclusion of such degraded samples and the selection of more suitable alternatives on a per-utterance basis.

As new enhancement models continue to emerge—targeting specific issues such as background noise, codec artifacts, or reverberation—the importance of dynamic, utterance-specific preprocessing is expected to increase. Our framework is designed to accommodate this evolving landscape, maintaining robustness in the face of both the strengths and limitations of future enhancement techniques.

From this perspective, the proposed "Switching" mechanism represents a forward-compatible design that anticipates the growing diversity and complexity of speech restoration models. It offers a principled and scalable solution for optimizing data cleansing strategies automatically, without reliance on static or manually crafted pipelines.

VIII. CONCLUSION

This paper presented **TTSOps**, a fully automated, closed-loop corpus optimization framework for training multi-speaker text-to-speech (TTS) models from large-scale, noisy, and uncurated web data. Unlike conventional pipelines that rely on static data cleansing and model-agnostic selection criteria, TTSOps unifies evaluation-in-the-loop data selection and utterance-wise data cleansing switching, both guided by a novel criterion: the training data quality—defined as the estimated contribution of each utterance to downstream TTS performance.

By integrating model-aware evaluation and adaptive data cleansing into the training loop, TTSOps effectively constructs high-quality, speaker-diverse corpora from real-world "dark data" such as YouTube videos. Extensive experiments using two TTS architectures demonstrated that TTSOps consistently outperforms acoustic-quality-based baselines in terms of naturalness, speaker coverage, and speaker diversity. Furthermore, the effectiveness of the framework was validated not only

[14]https://www.openslr.org/141/

through pseudo MOS but also by human-rated MOS evaluations.

Importantly, the framework adapts to the robustness characteristics of the target TTS model and flexibly selects the most effective data cleansing method for each utterance, enabling principled and scalable corpus construction. While modest discrepancies between pseudo and actual MOS were observed for certain models (e.g., GlowTTS), the increasing accuracy of automatic speech quality predictors promises further gains.

In summary, TTSOps offers a practical and forward-compatible approach to data-centric TTS development, bridging the gap between uncurated web data and high-quality synthetic speech. Future work includes scaling the framework to multilingual and cross-domain settings, incorporating alternative quality metrics (e.g., prosody or expressiveness), and exploring lightweight evaluation strategies to further reduce computational cost.

REFERENCES

[1] 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.

[2] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Int. Conf. Mach. Learn.*, 2021, pp. 8599–8608.

[3] 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.

[4] 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.

[5] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: 2000-speaker neural text-to-speech," in *Proc. Int. Conf. Learn. Representation*, 2018.

[6] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[7] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6184–6188.

[8] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Int. Conf. Mach. Learn.*, 2022, pp. 2709–2720.

[9] D. Trajanov, V. Zdraveski, R. Stojanov, and L. Kocarev, "Dark data in Internet of Things (IoT): Challenges and opportunities," in *7th Small Systems Simulation Symposium*, 2018, pp. 1–8.

[10] B. Schembera and J. M. Durán, "Dark data as the new challenge for big data science and the introduction of the scientific data officer," *Philosophy & Technology*, vol. 33, pp. 93–115, 2020.

[11] 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.

[12] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-Fi multi-speaker English TTS dataset," in *Proc. Interspeech*, 2021, pp. 2776–2780.

[13] P. Puchtler, J. Wirth, and R. Peinl, "HUI-Audio-Corpus-German: A high quality TTS dataset," in *KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, September 27–October 1, 2021, Proceedings 44*, 2021, pp. 204–216.

[14] A. Rousseau, P. Deléglise, and Y. Esteve, "TED-LIUM: An automatic speech recognition dedicated corpus," in *Proc. LREC*, 2012, pp. 125–129.

[15] C. Zhang, Y. Ren, X. Tan, J. Liu, K. Zhang, T. Qin, S. Zhao, and T.-Y. Liu, "DenoiSpeech: Denoising text to speech with frame-level noise modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 7063–7067.

[16] T. Saeki, K. Tachibana, and R. Yamamoto, "DRSpeech: Degradation-robust text-to-speech synthesis with frame-level and utterance-level acoustic representation learning," in *Proc. Interspeech*, 2022, pp. 793–797.

[17] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "LibriTTS-R: A restored multi-speaker text-to-speech corpus," in *Proc. Interspeech*, 2023, pp. 5496–5500.

[18] A. Watanabe, S. Takamichi, Y. Saito, W. Nakata, D. Xin, and H. Saruwatari, "Coco-Nut: Corpus of Japanese utterance and voice characteristics description for prompt-based control," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2023, pp. 1–8.

[19] 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.

[20] 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.

[21] 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.* IEEE, 2024, pp. 11 341–11 345.

[22] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 705–718, 2025.

[23] K. Nikitaras, G. Vamvoukakis, N. Ellinas, K. Klapsas, K. Markopoulos, S. Raptis, J. S. Sung, G. Jho, A. Chalamandaris, and P. Tsiakoulis, "Fine-grained noise control for multispeaker speech synthesis," in *Proc. Interspeech*, 2022, pp. 828–832.

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

[25] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.

[26] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[27] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4628–4632.

[28] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[29] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023.

[30] T. Saeki, S. Takamichi, T. Nakamura, N. Tanji, and H. Saruwatari, "SelfRemaster: Self-supervised speech restoration for historical audio resources," *IEEE Access*, vol. 11, pp. 144 831–144 843, 2023.

[31] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2021, pp. 166–170.

[32] H. Liu, X. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: A unified framework for high-fidelity speech restoration," in *Proc. Interspeech*, 2022, pp. 4232–4236.

[33] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, Y. Zhang, W. Han, A. Bapna, and M. Bacchiani, "Miipher: A robust speech restoration model integrating self-supervised speech and text representations," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5.

[34] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," *arXiv preprint arXiv:1611.09207*, 2016.

[35] S.-w. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. Interspeech*, 2018, pp. 1873–1877.

[36] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.

[37] W.-C. Tseng, C.-y. Huang, W.-T. Kao, Y. Y. Lin, and H.-y. Lee, "Utilizing self-supervised representations for mos prediction," in *Proc. Interspeech*, 2021, pp. 2781–2785.

[38] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 8442–8446.

[39] 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.

[40] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS challenge 2022," in *Proc. Interspeech*, 2022, pp. 4536–4540.

[41] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2023, pp. 1–7.

[42] W.-C. Huang, S.-W. Fu, E. Cooper, R. E. Zezario, T. Toda, H.-M. Wang, J. Yamagishi, and Y. Tsao, "The VoiceMOS challenge 2024: Beyond speech quality prediction," in *2024 IEEE Spoken Lang. Technol. Workshop*, 2024, pp. 803–810.

[43] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""everyone wants to do the model work, not the data work": Data cascades in high-stakes ai," in *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15.

[44] 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.

[45] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, "JTubeSpeech: corpus of Japanese speech collected from YouTube for speech recognition and speaker verification," *arXiv preprint arXiv:2112.09323*, 2021.

[46] 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.

[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. Williams and S. King, "Disentangling style factors from speaker representations," in *Proc. Interspeech*, 2019, pp. 3945–3949.

[49] T. Hayashi, R. Yamamoto, T. Yoshimura, P. Wu, J. Shi, T. Saeki, Y. Ju, Y. Yasuda, S. Takamichi, and S. Watanabe, "ESPnet2-TTS: Extending the edge of TTS research," *arXiv preprint arXiv:2110.07840*, 2021.

[50] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 5679–5683.

[51] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "A review on subjective and objective evaluation of synthetic speech," *Acoustical Science and Technology*, vol. 45, no. 4, pp. 161–183, 2024.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997.

[57] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 12 449–12 460, 2020.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representation*, 2015.

[59] W. B. March, P. Ram, and A. G. Gray, "Fast euclidean minimum spanning tree: algorithm, analysis, and applications," in *Proc. the 16th ACM SIGKDD Int. Conf. Knowledge discovery and data mining*, 2010, pp. 603–612.

[60] D. Stanton, M. Shannon, S. Mariooryad, R. Skerry-Ryan, E. Battenberg, T. Bagby, and D. Kao, "Speaker generation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7897–7901.

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.

Shinnosuke Takamichi received the Ph.D. degree from the Graduate School of Information Science, Nara Institute of Science and Technology, Japan, in 2016. He is currently an Associate Professor at Keio University, Japan. 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.