

An Automated End-to-End Open-Source Software for High-Quality Text-to-Speech Dataset Generation

Ahmet Gunduz, Kamer Ali Yuksel, Kareem Darwish, Golara Javadi
Fabio Minazzi, Nicola Sobieski, Sébastien Bratières

aiXplain, Inc.

{ahmet, kamer, kareem, golara}@aixplain.com

Translated, Inc.

{fabio, nicola, sebastien}@translated.com

Abstract

Data availability is crucial for advancing artificial intelligence applications, including voice-based technologies. As content creation, particularly in social media, experiences increasing demand, translation and text-to-speech (TTS) technologies have become essential tools. Notably, the performance of these TTS technologies is highly dependent on the quality of the training data, emphasizing the mutual dependence of data availability and technological progress. This paper introduces an end-to-end tool to generate high-quality datasets for text-to-speech (TTS) models to address this critical need for high-quality data. The contributions of this work are manifold and include: the integration of language-specific phoneme distribution into sample selection, automation of the recording process, automated and human-in-the-loop quality assurance of recordings, and processing of recordings to meet specified formats. The proposed application aims to streamline the dataset creation process for TTS models through these features, thereby facilitating advancements in voice-based technologies.

Keywords: Text-to-Speech Dataset Generation, Automated Recording and Quality Assurance

1. Introduction

The advent of artificial intelligence (AI) has brought about transformative changes across a multitude of industries, from healthcare and finance to entertainment and communication. One of the most notable areas impacted by AI is voice-based technologies, which have seen significant advancements in recent years. Text-to-Speech (TTS) systems, in particular, have become increasingly sophisticated, finding applications in various sectors, including assistive technologies, content creation, and customer service. The recent achievement with data-centric approaches proved to be as crucial as model architecture (Mazumder et al., 2022). Especially for complex models, the data quality stands out as much as the quantity of it. Therefore, developing high-quality TTS models is contingent upon the availability of comprehensive and well-curated datasets. This process often involves labor-intensive tasks such as data selection, recording, and annotation.

Generating datasets suitable for TTS models involves several steps, each with its challenges. Sample selection, for instance, must consider the complete coverage of phonemes to make it possible for the resultant TTS model to reproduce all the sounds of the target language. The recording process, too, requires meticulous planning and execution to guarantee the audio quality is up to par. Furthermore, the assurance of recording quality often necessitates using Automatic Speech Recognition (ASR) models to validate the data. Finally,

preprocessing steps are essential to convert the raw recordings into a format amenable to training.

Given these complexities, we introduce an integrated tool to streamline the dataset generation process for training TTS models, which reduces manual effort and enhances the quality and reliability of the datasets produced. Our proposed open-source tool is unique because it enables the rapid preparation of text for recording, batch processing of recorded audio, and quality assurance. To our knowledge, no other integrated open-source tool with similar functionality is available. The contributions of this work are multifaceted, namely:

- A novel approach for sample selection that diversifies language-specific phoneme distribution, thereby enhancing the linguistic richness of the dataset.
- An automated recording process that minimizes human intervention, increasing efficiency and enabling the speakers to focus on the voice performance.
- Quality assurance mechanisms powered by ASR models are integrated into the system to validate the recording accuracy and quality.
- Preprocessing functionalities that prepare the recordings for subsequent model training.

The code used in this work is publicly available.¹

¹<https://github.com/aixplain/tts-q>

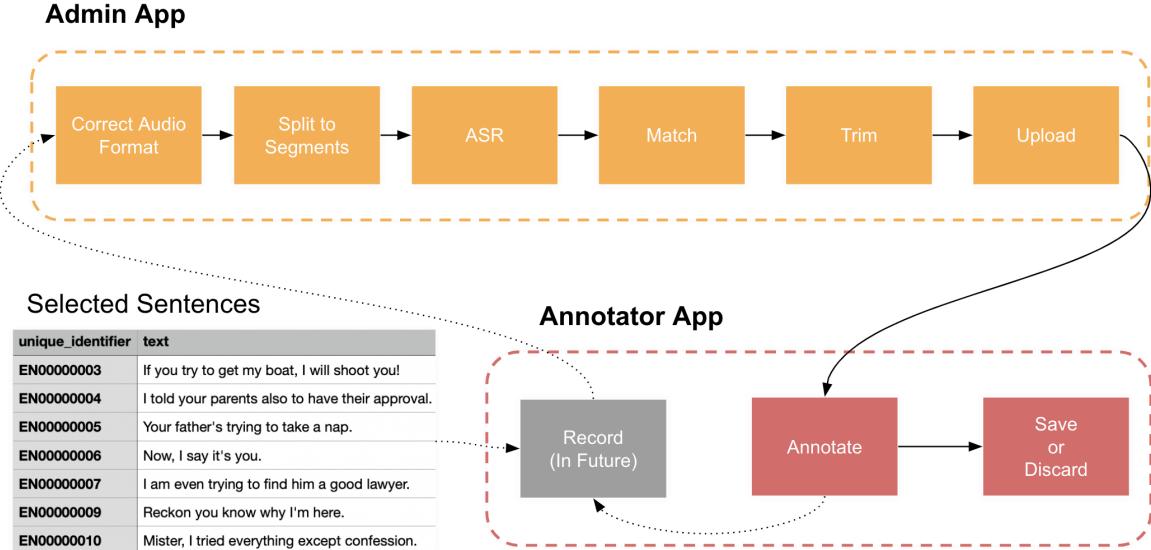


Figure 1: Workflow of the System

2. Related Work

Text-to-Speech (TTS) has seen significant advancements in recent years, primarily due to the application of deep learning techniques (Jeong et al., 2021; Zhang et al., 2023). Various architectures, such as Tacotron (Wang et al., 2017) and WaveNet (Oord et al., 2016), were proposed to improve the naturalness and intelligibility of synthesized speech.

Following the development of renowned generative modeling frameworks, like Generative Adversarial Networks and Normalizing Flows (Goodfellow et al., 2014; Rezende and Mohamed, 2015), their application in TTS engines became prominent. These frameworks facilitated parallel generation, ensuring the quality of synthesized speech remained consistent. After the WaveNet paper's release in 2016, there was a surge in efforts to develop a parallel non-autoregressive vocoder for high-quality speech synthesis. Architectures such as Parallel WaveNet and WaveGlow (Prenger et al., 2019; Oord et al., 2018), rooted in Normalizing Flows, not only accelerated the inference process but also upheld superior synthesis quality, especially evident on GPU devices (Popov et al., 2021).

Recently, a new category of generative models called Diffusion Probabilistic Models (DPMs), has demonstrated its proficiency in modeling intricate data distributions, encompassing areas like images, shapes, graphs, and handwriting. The fundamental concept of DPMs revolves around a two-step process: Initially, a forward diffusion process is constructed by progressively deconstructing the original data until a basic distribution is achieved. Subsequently, a reverse diffusion, parameterized by a neural network, is designed to trace the paths of

the forward diffusion in reverse time. Two vocoders representing the DPM family showed impressive results in raw waveform reconstruction: WaveGrad (Chen et al., 2020) and DiffWave (Kong et al., 2020) were shown to reproduce the fine-grained structure of human speech.

However, the quality of these models is highly dependent on the datasets used for training. Previous works have explored different aspects of dataset creation, including data augmentation techniques, phoneme-based selection, and quality assurance through manual annotation or semi-automated methods. Similarly, in machine translation (MT), approaches to dataset generation to optimize the annotation process have been adopted (Yuksel et al., 2022). This paper introduces a system with comprehensive features designed to expedite the generation of datasets for TTS applications.

3. Data and Preprocessing

Data were sourced from publicly available repositories, specifically the OPUS corpus². Six languages were targeted for this study: German, English, Mandarin, Italian, French, and Spanish. The datasets across different languages contained raw sentences (segments from the dataset not subjected to processing or cleaning). The study aimed to generate 30 hours of audio recordings for each language. Several text and audio constraints were established to ensure the data quality.

The initial step involved the extraction of sentences from the scripts collected for each target language. Subsequently, a filtering process was

²<https://opus.nlpl.eu>

applied to these sentences to eliminate those containing numerical values or abbreviations, so that the dataset can be used independently from the normalization rules that will be adopted in the training and inference phases. Additionally, sentences were filtered based on predetermined length constraints to account for the typical attention span of the current TTS engines. The sentences that met these criteria were fed into our text analysis system for further processing. Tables 1 and 2 outline the specific criteria set for the text and audio samples.

4. Methodology

The methodology section provides an in-depth description of the proposed end-to-end tool for TTS dataset generation. The tool is designed to be modular, allowing for customization at each stage of the dataset creation process. Figure 1 indicates the overall workflow of the system. The first module focuses on sample selection, employing language-specific phoneme distribution algorithms to ensure a balanced and comprehensive dataset. The second automates the recording process, providing a user-friendly interface for capturing high-quality audio samples. The third leverages Automatic Speech Recognition (ASR) models for quality assurance, flagging recordings that do not meet predefined quality criteria. The final module preprocesses the recordings into formats compatible with existing TTS training pipelines.

The system employs the aiXplain SDK³ and platform⁴ for model-specific tasks such as Automatic Speech Recognition (ASR) and Voice Activity Detection (VAD). Streamlit is utilized for the user interface due to its straightforward design and ease of implementation. Each service is containerized using Docker, and the entire project can be deployed using a Docker Compose file for reliability.

4.1. Sample Selection

Within the text analysis system, the sentences underwent a phonemization process to capture the distinct phonetic elements inherent to each language. For this purpose, the Espeak Phonemizer⁵ was employed to generate monophones, diphones, and triphones for each sentence. After this, frequencies of the generated phonemes were computed for each sentence and aggregated into a dictionary object, forming an initial corpus-level distribution of phonetic elements. This served as a foundational phonetic reference for each language.

To construct a representative subset of the corpus, an iterative selection process was employed, guided by multiple criteria: the type of sentence, the length of the sentence, and the phonetic distribution relative to the corpus-level distribution. The algorithm for sentence prioritization employs stochastic selection, but assigns higher probability values to sentences that contribute to aligning the subset's phonetic distribution more closely with that of the overall corpus. This alignment is quantified by calculating the divergence between the phonetic distribution of the current subset and the corpus-level distribution. In this iterative process, sentence type and length constraints are applied as filters to guide the selection of subsequent samples, ensuring that the resulting subset remains within predefined percentage boundaries for these criteria. Figure 3 shows the workflow of the sentence selection process. Each sentence was weighted according to the frequency of its constituent phonemes. Sentence selection for the final dataset was performed through iterative sampling to align the selected sentences' phonetic elements with the corpus-level distribution. This selection process was initialized randomly and refined iteratively.

To quantify the dataset, a target of a minimum of 600,000 words per language was set. Utilizing a reference rate of 2.75 words per second, it was determined that this word count would be sufficient to produce more than 30 hours of audio recordings.

4.2. Preparing Recordings

Voice actors were permitted to utilize their preferred audio recording and editing software for self-directed recording sessions, provided they adhered to the criteria specified in Table 2. We also gave the voice actors the choice of saving the recording of each sentence in a separate file or batch recording one sentence after another into one file. If a voice actor decides to record in batch, they need to adhere to the following guidelines:

- A single file should have a maximum of 500 sentences.
- File naming should follow the "start_ID-end_ID.wav" convention (e.g., DE00000037-DE00000720.wav).
- A minimum of 2 seconds should be maintained between each sentence. This is essential to perform accurate Voice Activity Detection (VAD) for identifying sentence audio segments.
- In case of errors, the voice actor can re-read a sentence (as many times as (s)he likes), with the condition that the last iteration is correct.

³<https://docs.aixplain.com/main.html>

⁴<https://platform.aixplain.com>

⁵<https://github.com/rhasspy/espeak-phonemizer>

Criteria	Description
Sentence Types	The dataset should encompass a variety of sentence structures, including declarative (80%), interrogative (10-15%), and exclamatory (5-10%) sentences, as indicated by ending with period, question mark, and exclamation mark, respectively.
Sentence Length	Each sentence should be no fewer than five and no more than 13 words.
Normalization	The dataset should exclude acronyms, abbreviations, digits, or symbols necessitating text normalization.

Table 1: Text Data Criteria

Criteria	Description
File Format	WAV, Mono channel
Sampling Rate	88 kHz
Sample Format	16-bit, PCM
Peak Volume Levels	from -3 dB to -6 dB
Signal-to-Noise Ratio	Not less than 35 dB
Silence Duration	Leading and trailing silences should not exceed 100 ms; internal silences should not exceed 0.5 seconds.
Audio Artifacts	The recordings should be free from lip-smacking, echo, and breath sounds.
Recording Length	Each recording should be no longer than 15 seconds and no shorter than 2 seconds.
Speech Rate	Recordings should be made at a natural speed.
Accent	The accent in the recordings should align with the target language.
Punctuation Accuracy	The audio should accurately reflect the punctuation in the text.

Table 2: Audio Data Criteria

The batch recording is initially split into segments using VAD. Each segment is then transcribed using an ASR, and the Levenshtein edit distance is computed between the ASR output and all the sentences between the start and end IDs in the recording name. Initially, Whisper⁶ was employed as the default ASR system, owing to its language-agnostic capabilities. However, due to the auto-correction features inherent to Whisper, which led to discrepancies in the dataset, it became necessary to transition to alternative ASR systems available on the aiXplain platform. Though we used Azure, one of the advantages of using the aiXplain platform is that we can swap in and out different ASR systems with no changes to the code. The sentence with the lowest edit distance is picked given if: a) the ratio of edit distance to the minimum length of ASR output and sentence is less than 0.2, and the difference in length between the ASR output and the sentence does not exceed 20% of the length of the shorter of the two. Given these conditions, we achieved an average match rate of 99%. Subsequently, matched segments are trimmed using VAD to remove leading and trailing silences exceeding 100 ms. We also ensured there was at least 25 ms of silence to avoid any speech truncation.

⁶<https://github.com/openai/whisper>

4.3. Automated Processing of Recordings

This module aims to help with data creation for text-to-speech (TTS), which involves two primary components: the user interface and the back-end, which facilitate audio data annotation and storage.

4.3.1. User Interface

The user interface is bifurcated into an Admin Dashboard and an Annotator Dashboard, as depicted in Figures 4 and 6, respectively. For privacy considerations, some user-specific information has been redacted from these figures.

Admin Dashboard: This interface is responsible for various administrative tasks such as uploading recordings from voice actors, dataset creation, visualization of annotation progress, user account management, and task assignment.

Annotator Dashboard: This interface is designed to annotate recordings. As can be seen in the screenshot, annotation involves multiple tasks:

- Verifying that the audio matches the corresponding sentence and or editing the text as

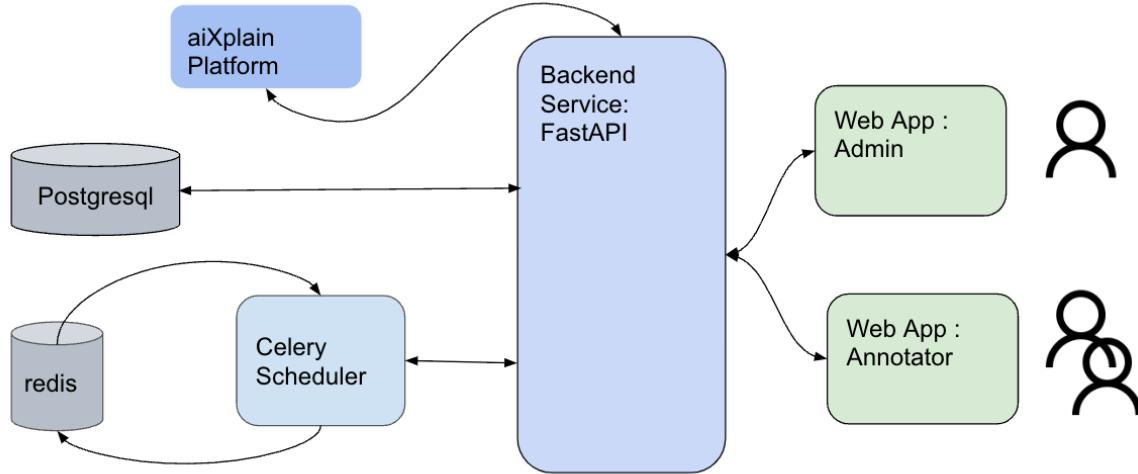


Figure 2: System Architecture of the Tool

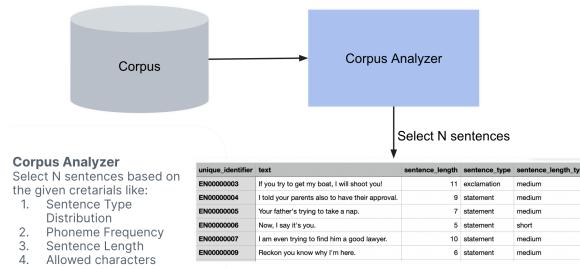


Figure 3: Sentence Selection for TTS Recordings

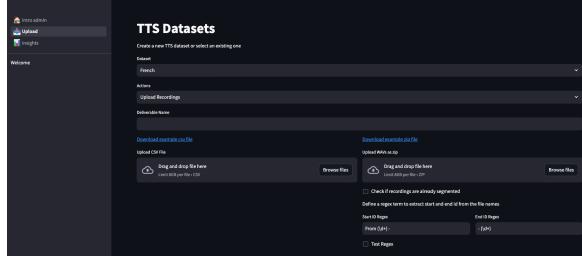


Figure 4: Screenshot of Admin Dashboard for Recording Upload

necessary to ensure they match entirely. The annotators can post-edit the original text or the ASR output.

- Marking recordings as having specific problems such as repetitions and incorrect prosody.

More details about annotations are provided in the Quality Assurance section (Section 4.4).

4.3.2. Backend

The system architecture, illustrated in Figure 2, comprises a PostgreSQL database, a Celery

scheduler⁷, and a Redis backend orchestration service and in-memory database⁸. Using Celery and Redis helps manage various asynchronous tasks, such as audio segmentation and running ASR. The processed and annotated recordings are stored in the PostgreSQL database and S3 buckets. Both raw and processed audio files are also saved to local directories to ease subsequent processing.

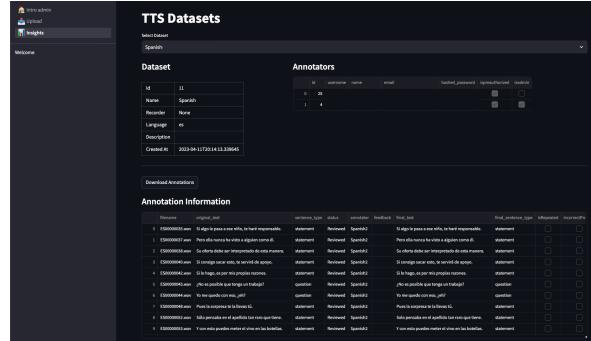


Figure 5: Screenshot of Admin Dashboard for Insights of the annotations done.

4.4. Quality Assurance

After the initial preprocessing and uploading of audio recordings, the first Quality Assurance is performed by a team of annotators native in each target language. The annotation phase ensures that each recording meets the criteria for high-quality samples. Figure 6 displays a screenshot of the annotation interface. Administrators, utilizing the Admin Dashboard, allocate specific datasets to individual annotators. A user-level authentication

⁷<https://docs.celeryq.dev/en/stable/>

⁸<https://redis.io/>

system monitors annotator activities and adds a layer of security. Access is restricted to accounts created using pre-authorized email addresses.

Upon logging in, annotators are presented with the datasets assigned to them. Individual recordings are sequentially retrieved for annotation when a dataset is selected, prioritizing those with the highest Word Error Rate (WER), as computed by the ASR system against the reference text. Annotators can listen to the audio, make text edits to ensure exact alignment with the audio (including pauses for punctuation), modify the sentence structure, and discard recordings for various reasons. These reasons may include word or phrase repetition, improper prosody, inconsistency between audio and text, incorrect pronunciation, or auditory artifacts such as noise or lip-smacking. An additional feedback field is available for annotators to elucidate the rationale behind discarding a sample.

The annotation system is designed to accommodate multiple annotators working concurrently, even on identical datasets. Each sample is locked to a specific annotator session to prevent overlap, ensuring that annotators do not conflict. Due to the resource-intensive nature of the annotation process, only a single annotation is permitted for each sample. Moreover, administrators can oversee the annotation process via the Admin Dashboard, which features statistical data on the Insight page. This functionality allows administrators to monitor each task's progress and maintain the highest possible annotation quality. As illustrated in Figure 5, administrators can review all datasets and download annotated data for subsequent TTS training.

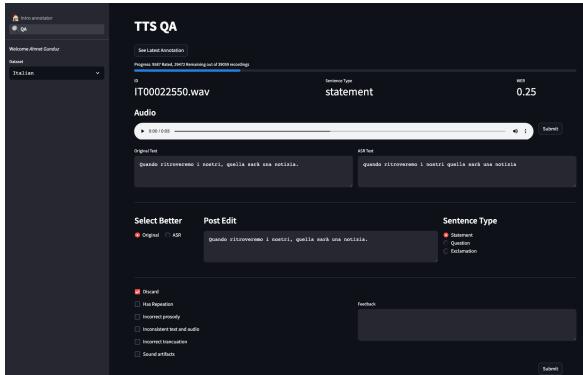


Figure 6: The Screenshot of Annotator App

5. Experimental Setup and Results

In this section, we discuss the experiments conducted to evaluate the efficiency and quality of the trimming process. To evaluate the efficacy of the proposed tool, a series of experiments were conducted using multiple languages: German (DE),

French (FR), Spanish (ES), Italian (IT), and English (EN). Table 4 provides an overview of the datasets before and after annotation, while Table 3 offers a detailed breakdown of file durations, sentence assignments, and assignment percentages by language in our matching algorithm. One of the notable observations from Table 3 is the reduction in total duration after trimming silences from the audio files. For instance, in the German dataset, the duration decreased from 1549.47 seconds to 1330.77 seconds after trimming. This reduction makes the dataset more compact and enhances its quality by eliminating non-informative silences.

5.1. Matching Efficiency

The matching efficiency, as observed from the "Dur(ation) Before Match." and "Dur(ation) After Match." columns in Table 3, is approximately 98% across multiple languages. This high efficiency indicates that most audio data aligns well with the corresponding text, ensuring the dataset is highly reliable for further processing as machine learning.

Moreover, the high percentage of assigned sentences, as indicated in the last column of Table 3, further corroborates the efficiency of the matching algorithm. For example, in the English dataset, 94.9% to 99.8% of the total segments were successfully assigned. As this has been observed for all languages, we conclude that the method's accuracy will persist if the voice actors follow the requirements mentioned in Section 4.2.

5.2. Quality of Annotated Data

The annotation process plays a crucial role in ensuring the dataset's quality. As seen in Table 4, the percentage of post-edited samples is relatively low, such as 1.25% for the Spanish dataset. Additionally, the rate of discarded samples is negligible, standing at 0.00% for the same dataset. These low percentages suggest that most of the data is of high quality and requires minimal intervention during the annotation phase.

5.3. Second Review Pass

To ensure the result of the overall process is not affected by the individual annotators, we included a second review step by another set of annotators native to the respective target languages. We reviewed at least 70% of the datasets, not all of them, due to time and budget restrictions. The instructions given to the second set of annotators were the same as those provided for the first review step. The quality of the edits was carefully screened to verify whether the annotators involved in the first Quality Control stage correctly interpreted the instructions, especially regarding match-

Lang	File	Dur. Before Match.	Dur. After Match.	Dur. Trim.	Total Files	Assigned	Not Assigned	% Assigned
DE	File1	2439.02	1549.47	1330.77	495	480	15	97.0%
DE	File2	2354.78	1493.00	1274.15	494	486	8	98.4%
FR	File1	2271.79	1465.28	1241.03	498	491	7	98.6%
FR	File2	2326.11	1475.08	1253.37	498	488	10	98.0%
ES	File1	2505.61	1499.82	1286.45	498	491	7	98.6%
ES	File2	2216.55	1464.50	1241.68	500	488	12	97.6%
IT	File1	2249.54	1473.51	1247.73	496	489	7	98.6%
EN	File1	1906.00	1285.45	1020.80	530	503	27	94.9%
EN	File2	2692.67	1241.19	1011.56	500	499	1	99.8%

Table 3: Performance Metrics of Sentence Matching Algorithms Across Multiple Languages and Files. The table summarizes the duration before and after matching, the duration after trimming, and the number of sentences assigned and not assigned. The percentage of sentences successfully assigned is also included, with performance generally exceeding 98%, except in one English recording where it is 94.9%.

ing the speech and the punctuation. Proper training of TTS engines requires a close matching between punctuation and speech, considering that punctuation is an essential cue for neural networks to correctly interpret subordinate and coordinate sentences, exclamations, and questions. The results in Table 5 show that the toolset developed for this study effectively generates high-quality data.

5.4. Review of the Recordings

The annotation process primarily focuses on reviewing the recordings, which are often the most critical in determining the overall quality of the dataset. The high percentages of assigned sentences and the low percentages of post-edited or discarded samples suggest that the end recordings generally meet the quality requirements, yielding a high-quality dataset suitable for various applications. In summary, the experiments demonstrate that the annotation process is highly efficient at creating high-quality datasets. The trimming of silences and the high matching efficiency contribute to the efficiency of the recording process, while the meticulous annotation process ensures its quality. Italian stands as an exception in this study as the speaker could not correctly perform the script, so in both QC phases, many adjustments had to be made.

5.5. Staff for TTS dataset collection

One of the results of this study is that to create a high-quality dataset for TTS the process outlined here has to be operated by a team of professionals, each native in one of the target languages. The speakers should be able to control their prosody according to the punctuation and intonations indicated in the brief and the script. Also, it should be apparent what the expected inflection of the speech should be e.g., a speaker with a regional cadence is not desirable for building a dataset aimed at train-

ing models that have to speak the official language of a country. The annotators should also be native, as they must be able to catch inflections and flaws in prosody, which can confuse the TTS engines at training time with low correlations between text and speech. This is especially true for small training datasets (<100h), where the statistical error cancellation has a lower impact than large datasets.

6. Discussion and Limitations

This paper presents a comprehensive tool for generating high-quality Text-to-Speech (TTS) datasets applicable across various languages. However, there are inherent limitations concerning the tool’s reliance on high-quality Automatic Speech Recognition (ASR) models for quality assurance. Such models may not be universally available for all languages or dialects. Initially, the Whisper ASR model was employed with its default configurations, serving as a language-agnostic solution. Nonetheless, the auto-correction functionality within Whisper posed challenges in accurately segmenting the audio recordings. Specifically, during the batch recording process, voice actors are instructed to repeat a sentence if an error occurs while recording. If insufficient pauses are made between these repetitions, the Voice Activity Detection (VAD) system interprets the repetitions as a single sentence recording. Subsequently, Whisper’s auto-correction feature alters the sentence transcripts, leading to inaccuracies in the dataset. To mitigate this issue, we resorted to utilizing other ASR services on aiXplain platform to cross-verify and correct recordings.

Additionally, for segmented recordings, there were instances where voice actors or actresses incorrectly labeled the recordings with erroneous sentence identifiers. To rectify this issue, we applied our matching algorithm -initially designed for batch recording uploads—to all segmented recordings. Specifically, the algorithm cross-referenced

Language	# of Samples	Bad Prosody	Inconsistent Text-Audio	Truncation	Sound Artifacts	% Edited	% Discarded
German	30000	0	1	0	21	1.90%	0.15%
Spanish	45489	0	0	0	0	1.25%	0.00%
Italian	30001	0	0	0	0	11.38%	0.00%
English	33373	2	0	3	0	1.44%	0.02%
French	30005	0	0	0	0	3.23%	0.00%

Table 4: Performance metrics for the second stage of Quality Control Across Multiple Languages. The table provides the total number of samples, the type of errors found, the % of segments edited, and the % of segments discarded in this second QC.

the original sentence associated with the filename against the ASR-generated transcript of the recording. Re-matching was conducted using the Levenshtein edit distance, employing the same methodology as in the batch recording matching process. As a consequence of these challenges, we intend to facilitate recording samples directly through our proprietary tool in future iterations. This approach aims to minimize the likelihood of labeling errors and streamline the overall data collection process.

7. Future Work

Future work could focus on several avenues to enhance the tool’s capabilities. The first is developing an efficient recording tool to minimize the speakers’ distraction from their performance. In light of the challenges encountered with erroneous labeling of segmented recordings, we intend to facilitate recording samples directly through our proprietary tool in future iterations. This approach aims to minimize the likelihood of labeling errors and streamline the overall data collection process. Another potential direction is the integration of more advanced ASR models to improve the quality assurance process, especially for under-represented languages. Another avenue could be incorporating machine learning algorithms to automate the annotation process further, thereby reducing the need for human intervention. Additionally, the tool could be extended to support more complex data types and formats, making it more versatile and applicable to a broader range of TTS applications.

To further enhance the quality and accuracy of the generated TTS dataset, an innovative approach can be employed by leveraging an unreference speech dataset like movies or video recordings using a reference-less metric, such as NoRefER (Yuksel et al., 2023a,b), which enables the assessment of transcription accuracy without the need for a reference transcription. This capability is particularly beneficial in creating high-quality TTS datasets from speech data that has not been previously transcribed or for which no reliable reference exists. Selecting these high-fidelity transcriptions for the

TTS dataset ensures a foundation of exceptional quality. Incorporating suggestions from analyzing the NoRefER attentions can streamline curating a high-quality dataset by improving the annotation efficiency and effectiveness (Javadi et al., 2024).

8. Conclusion

This paper has presented an end-to-end tool engineered to automate and streamline the creation of high-quality datasets for Text-to-Speech (TTS) models. To our knowledge, no other tool with similar capabilities currently exists in the literature or the market. The tool incorporates several innovative features, including a sample selection algorithm for language-specific phoneme distribution, an automated recording process, and quality assurance mechanisms powered by Automatic Speech Recognition (ASR) models. Experimental results across multiple languages demonstrate the tool’s efficacy in producing datasets that are both comprehensive and of high quality. The annotation process, facilitated by a user-friendly interface, further ensures the reliability of the generated datasets.

The proposed tool represents a significant step forward in TTS dataset generation, offering a scalable and efficient solution for creating high-quality datasets. Its modular design allows for easy customization and adaptation, making it a valuable resource for researchers and practitioners alike in the rapidly evolving landscape of voice-based technologies. One of the key advantages is the tool’s ability to automate and streamline the dataset creation process, thereby reducing the time and effort required. However, it is essential to acknowledge certain limitations. The overall dataset quality depends on the professional level of the team operating it. The tool’s quality assurance mechanisms rely on the availability and accuracy of ASR models, which may not be universally applicable across all languages or dialects. Additionally, while the tool significantly reduces the manual effort required in dataset creation, human intervention is necessary.

9. References

- Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2020. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Golara Javadi, Kamer Ali Yuksel, Yunsu Kim, Thiago Castro Ferreira, and Mohamed Al-Badrashiny. 2024. [Word-level asr quality estimation for efficient corpus sampling and post-editing through analyzing attentions of a reference-free metric](#). In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2024, Seoul, Korea, April 14-19, 2024*.
- Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. 2021. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409*.
- Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.
- Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. 2022. Dataperf: Benchmarks for data-centric ai development. *arXiv preprint arXiv:2207.10062*.
- Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning*. PMLR.
- Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.
- Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR.
- Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE.
- Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *International conference on machine learning*. PMLR.
- Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.
- Kamer Ali Yuksel, Thiago Castro Ferreira, Ahmet Gunduz, Mohamed Al-Badrashiny, and Golara Javadi. 2023a. [A reference-less quality metric for automatic speech recognition via contrastive-learning of a multi-language model with self-supervision](#). In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.
- Kamer Ali Yuksel, Thiago Castro Ferreira, Golara Javadi, Mohamed Al-Badrashiny, and Ahmet Gunduz. 2023b. [Norefer: a referenceless quality metric for automatic speech recognition via semi-supervised language model fine-tuning with contrastive learning](#). In *Proc. INTERSPEECH 2023*, pages 466–470.
- Kamer Ali Yuksel, Ahmet Gunduz, Shreyas Sharma, and Hassan Sawaf. 2022. [Efficient machine translation corpus generation](#). In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas*. AMTA.
- Chaoning Zhang, Chenshuang Zhang, Sheng Zheng, Yu Qiao, Chenghao Li, Mengchun Zhang, Sumit Kumar Dam, Chu Myaet Thwal, Ye Lin Tun, Le Luang Huy, Donguk kim, Sung-Ho Bae, Lik-Hang Lee, Yang Yang, Heng Tao Shen, In So Kweon, and Choong Seon Hong. 2023. A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need? *arXiv preprint arXiv:2303.11717*.