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

Ahmet Gunduz (ahmet@aixplain.com), **Kamer Ali Yuksel** (kamer@aixplain.com), **Kareem Darwish** (kareem@aixplain.com), **Golara Javadi** (golara@aixplain.com), **Fabio Minazzi** (fabio@translated.com), **Nicola Sobieski** (nicola@translated.com), **Sébastien Bratières** (sebastien@translated.com)

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation
May 20–25, 2024

ABSTRACT

Data availability is crucial for advancing AI applications, especially in voice-based technologies. This study introduces an innovative, open-source tool designed to generate high-quality datasets for text-to-speech models. By integrating language-specific phoneme distribution in sample selection and automating the recording and quality assurance processes, this tool enhances the efficiency and reliability of TTS datasets crucial for the advancement of voice technologies.

INTRODUCTION

Artificial intelligence has significantly impacted various sectors, including voice-based technologies. The quality of TTS systems hinges on the dataset used for training. Our work addresses the challenge of creating high-quality, comprehensive datasets through an automated tool that simplifies and streamlines the data preparation process.

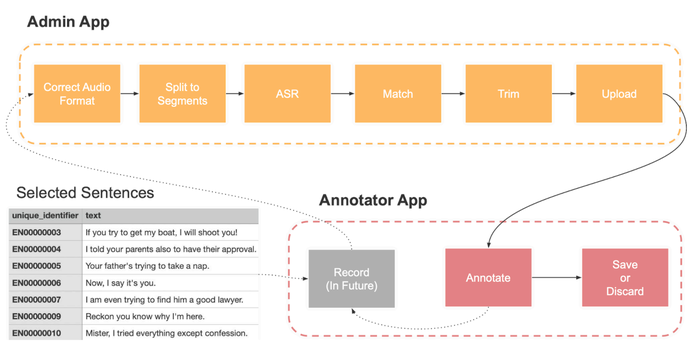
Motivation:
Quality of models highly depends on the dataset used for the training.

Motivation:

- Quality vs quantity: Data-centric approaches proved that quality is as important as quantity of the data in ML training.
- TTS models naturally ask for a nuanced representation of sounds/phonemes for different languages.
- Collecting high quality data is costly and need to be automated.
- Quality assurance with human in the loop.

Solution: An end-to-end approach for dataset generation is needed.

METHODOLOGY



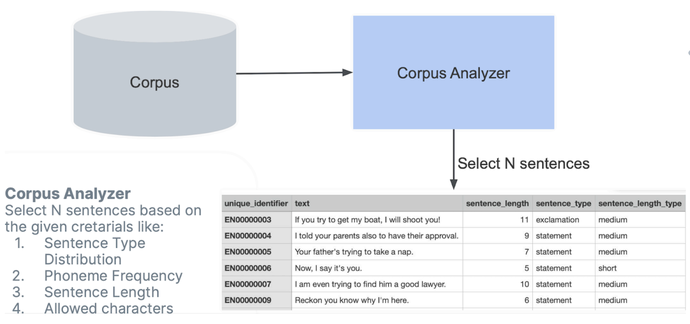
Workflow of the system

Data

- OPUS corpus (available at https://opus.nlpl.eu)
- Languages: German, Spanish, Italian, English, French
- Dataset collected: 30 hours in each language

Sample selection: Comprehensive representation of phonemes across different languages is crucial for producing natural and accurate speech in TTS systems.

- Cleaned and preprocessed Opus dataset into sentences.
- Monophone, diphone, and triphone distributions generated for each language.
- Prioritized sentence to meet language specific distributions.
- Selected N number of sentences based on user constraints like sentence type, sentence length, allowed characters.



Dataset creation process

Preparing recordings:

- Each sentence is assigned to a unique identifier.
- Recording done by voice-actors with their preferred tools.

Procedure:

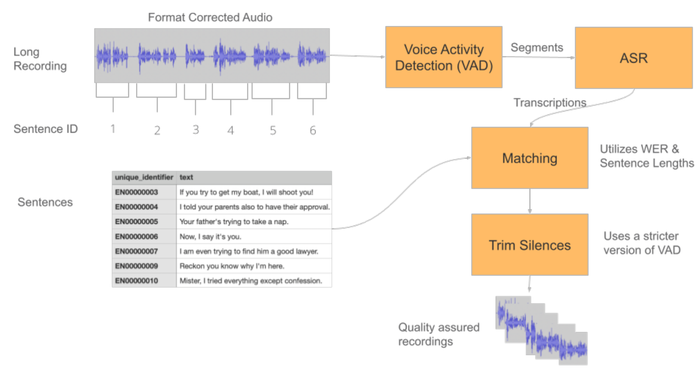
- A single file should have a maximum of 500 sentences.
- File name should include start and end index of sentences.
- A minimum of 2 seconds should be maintained between each sentence.
- In case of errors, the voice actor can re-read a sentence (as many times as (s)he likes), with the condition that the last iteration is correct.

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.

Audio quality requirements

Automated recording process: Provides a user-friendly interface for capturing high-quality audio samples. This process is designed to minimize human intervention, thereby increasing efficiency and allowing speakers to focus solely on their performance.

Integrated quality assurance: Utilizing Automatic Speech Recognition (ASR) models, the tool automatically verifies the accuracy and quality of the recordings. This step flags any recordings that do not meet predefined criteria, ensuring the dataset's reliability.



Automated recording process of an audio with multiple sentences.

RESULTS

Lang	File	Dur. Before Match.	Dur. After Match.	Dur. After Trim.	Total Files	Assigned	Not Assigned	% signed
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%

Performance metrics of sentence matching algorithms across multiple languages and files.

- Over 94% matching accuracy.
- Works for multiple languages.
- Trimming silences matters for high quality data.

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%

Performance metrics for the second stage of quality control across multiple languages.

- The percentage of discarded sentences are at negligible levels.
- Percentage of post-edited samples are in between 1-4%. Only Italian dataset is high because of the actress did not perform the script.

CONCLUSION

We presented a novel tool that automates the generation of high-quality TTS datasets, demonstrating its capability across various languages. This tool significantly contributes to the field by reducing the time and complexity involved in dataset preparation, supporting further advances in voice-based applications.

Innovations and contributions:

- Sample selection: Diversifies phoneme distribution, enhancing linguistic richness.
- Automated recording: Reduces human intervention, optimizing the recording workflow.
- Quality assurance: Utilizes ASR technology to ensure high-quality audio samples.