

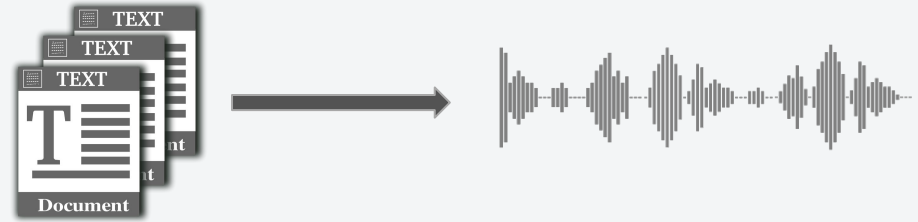
aixplain

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

Motivation

- Early TTS systems were relying on simple, concatenated sound bites, resulting in unnatural speech patterns
- The introduction of deep learning has revolutionized TTS like WaveNet, Tacotron etc.
- Recently Diffusion Probabilistic Models (DPMs) showed human like waveform reconstructions: WaveGrad, DiffWave



Quality Data

Quality Output

Quality of these models highly depends on the dataset used for the training



- 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

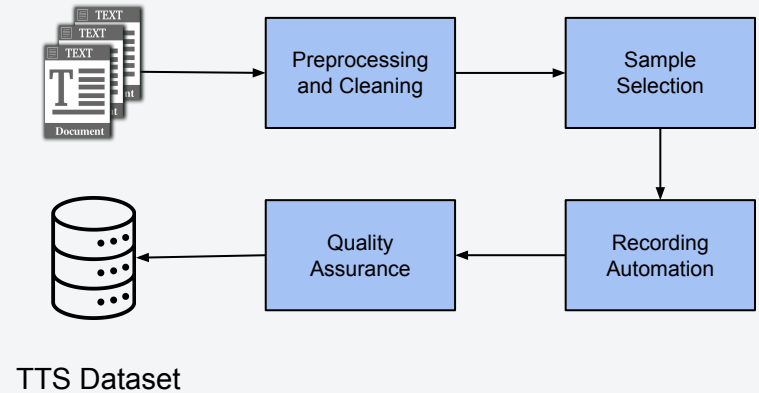
- A novel approach for sample selection that diversifies language specific phoneme distribution, thereby enhancing the linguistic richness of the dataset.
- An automated recording process that minimizes human intervention, increasing efficiency and enabling the speakers to focus on the voice performance.
- 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.

Components:

- Text Preprocessing and Cleaning
- Sample Selection
- Recording Automation
- Quality Assurance

Data:

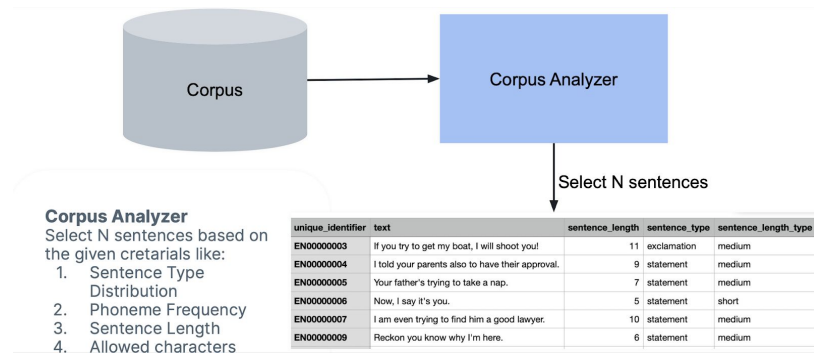
- OPUS corpus (available at <https://opus.nlpl.eu>)
- Languages: German, Spanish, Italian, English, French
- Dataset Collected: 30 hours in each language



Methodology: Sample Selection

A 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
- Monophones, Diphones, and Triphones 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



Preparing Recordings

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

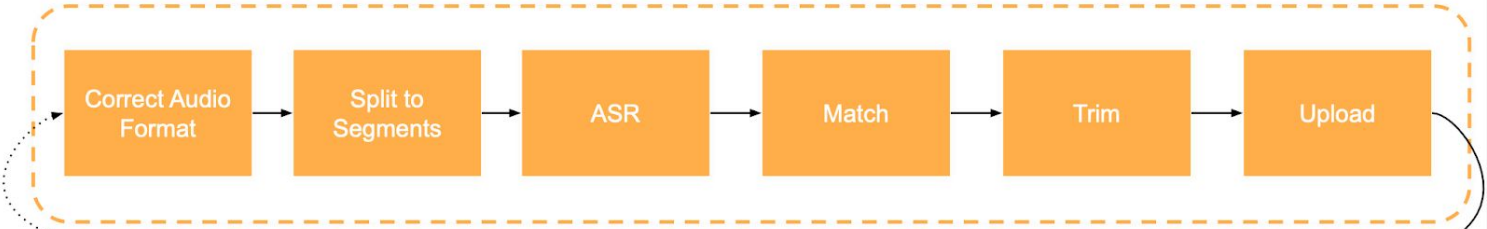
Procedure:

- 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.

Audio Requirements:

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.

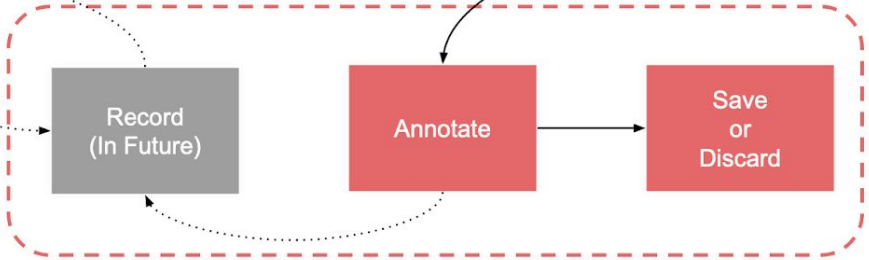
Admin App



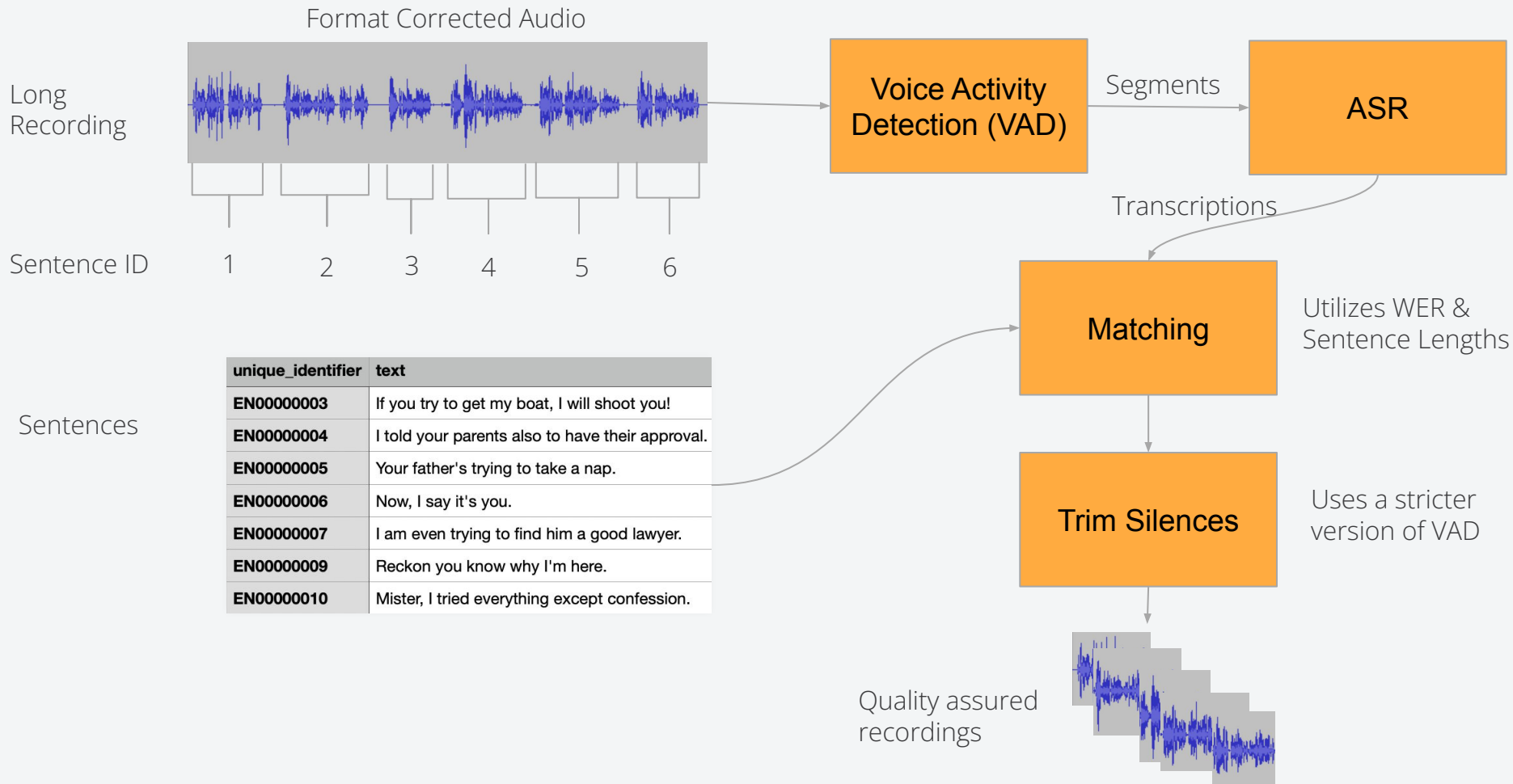
Selected Sentences

unique_identifier	text
EN00000003	If you try to get my boat, I will shoot you!
EN00000004	I told your parents also to have their approval.
EN00000005	Your father's trying to take a nap.
EN00000006	Now, I say it's you.
EN00000007	I am even trying to find him a good lawyer.
EN00000009	Reckon you know why I'm here.
EN00000010	Mister, I tried everything except confession.

Annotator App



Recording Automation



Recording Automation

Admin App

Uploading recordings, dataset creation, monitoring annotations process, annotation task assignment

TTS Datasets

Create a new TTS dataset or select an existing one

Dataset

French

Actions

Upload Recordings

Deliverable Name

[Download example csv file](#) [Download example xiv file](#)

Upload CSV File

Drag and drop file here
Limit 8GB per file • CSV

Browse files

Upload WAVs as zip

Drag and drop file here
Limit 8GB per file • ZIP

Browse files

Check if recordings are already segmented

Define a regex term to extract start and end id from the file names

Start ID Regex

From (id+) -

End ID Regex

- (id+)

Test Regex

Annotator App

Annotate the recordings: verifying the audio matches sentence, post-editing the text, discarding recordings with specific problems

ID IT00022550.wav

Sentence Type statement

WER 0.25

Audio

0:00 / 0:03

Submit

Original Text

Quando ritroveremo i nostri, quella sarà una notizia.

ASR Text

quando ritroveremo i nostri quella sarà una notizia

Select Better

Original ASR

Post Edit

Quando ritroveremo i nostri, quella sarà una notizia.

Sentence Type

Statement Question Exclamation

Discard

Has Repeation

Incorrect prosody

Inconsistent text and audio

Incorrect truncation

Sound artifacts

Feedback

Submit

Matching Efficiency for Different Files and Languages

Lang	File	Dur. Before Match.	Dur. After Match.	Dur. After 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%

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

Performance Details After Quality Control

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%

- 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

Future Directions

- Ability to Record in the tool
- Automation of Annotation Process through Active Learning
- Apply approach for more complex data types and formats

Leveraging unreferenced datasets like movies and videos with a referenceless ASR metric like NoRefER

Conclusion

- Recording processes for voice actors is automated successfully by utilizing ASR, VAD and WER matching mechanism
- Demonstrated success in producing comprehensive and high-quality datasets across **multiple languages.**
- Modular design allow to meet diverse research needs
- Proposed method relies on high quality ASR models but with whisper like models that is a negligible dependency



Thank you

Code



Demo Video

