VoxpopuliTTS: a large-scale multilingual TTS corpus for zero-shot speech generation

Wenrui Liu*, **Jionghao Bai***, **Xize Cheng, Jialong Zuo, Ziyue Jiang, Shengpeng Ji, Minghui Fang, Xiaoda Yang, Qian Yang, Zhou Zhao**†
Zhejiang University
{liuwenrui, zhaozhou}@zju.edu.cn

Abstract

In recent years, speech generation fields have achieved significant advancements, primarily due to improvements in large TTS (text-to-speech) systems and scalable TTS datasets. However, there is still a lack of large-scale multilingual TTS datasets, which limits the development of cross-language and multilingual TTS systems. Hence, we refine Voxpopuli dataset and propose VoxpopuliTTS dataset. This dataset comprises 30,000 hours of high-quality speech data, across 3 languages with multiple speakers and styles, suitable for various speech tasks such as TTS and ASR. To enhance the quality of speech data from Voxpopuli, we improve the existing processing pipeline by: 1) filtering out low-quality speech-text pairs based on ASR confidence scores, and 2) concatenating short transcripts by checking semantic information completeness to generate the long transcript. Experimental results demonstrate the effectiveness of the VoxpopuliTTS dataset and the proposed processing pipeline. Demos will be available at https://github.com/voxpopulitts/voxpopulitts.github.io.

1 Introduction

In recent years, speech generation models has made significant progress supported by scale-up TTS datasets. TTS models such as VALL-E (Wang et al., 2023), VoiceBox (Le et al., 2024), and NaturalSpeech 3 (Ju et al., 2024) have achieved impressive results in zero-shot speech generation. Popular TTS datasets have become indispensable resources for researchers.

However, TTS datasets based on audiobooks or studio recordings often exhibit a formal reading style and rarely capture natural elements such as breaths, pauses, and emotion that are common in everyday conversation. LJ Speech provides high-quality English speech samples, but its limited size and uniform style present challenges. Similarly, VCTK Corpus (Veaux et al., 2017), originating from studio recordings, lacks the diversity found in real-life scenarios. LibriTTS (Zen et al., 2019), derived from audiobooks, includes formal reading style speech data, serving as an important part for training and evaluating TTS models. GigaSpeech (Chen et al., 2021) is a ASR dataset, and while primarily used for speech recognition, its large volume of data also supports speech generation research. Although WenetSpeech4TTS (Ma et al., 2024) offers valuable resources for Chinese TTS research and the Emilia (He et al., 2024) dataset expands multilingual data, most of its data is concentrated in English and Chinese, with less than 5,000 hours available for other minor languages, making it insufficient for training TTS models in those languages.

To address these issues, we refine Voxpopuli dataset (Wang et al., 2021) and propose VoxpopuliTTS dataset aimed at providing high-quality, multilingual speech data to support more natural multilingual speech generation model. Our dataset undergoes a series of strict pre-processing steps to ensure purity and diversity. Specifically, our approach entails the following steps: 1) **confidence-based filtering strategy** is employed to exclude unreliable ASR transcription results, thereby avoiding hallucination transcription issues caused by ASR models. 2) **Merging adjacent audio segments**: by examining the semantic information completeness of the transcript, we can determine whether it is necessary to concatenate the preceding and following transcripts to create long transcript with complete semantic information. These features allows our dataset to maintain data quality while offering complete semantic information, contributing to the training of speech generation models that are more natural, fluent, and closer to human speech styles.

*These authors contributed equally to this work.
†Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 10293–10297
January 19–24, 2025.

Finally, we utilize DNSMOS (Reddy et al., 2021) to filter and classify the dataset by quality, dividing it into small, medium, and large subsets, thus allowing for broader application in various TTS tasks with differing quality requirements. By using our dataset, we hope to further advance speech generation, making it more aligned with real-world application needs.

In summary, the key features of VoxpopuliTTS are as follows:

- Scalable and multilingual: VoxpopuliTTS includes 3 languages across English, French, and Spanish, each language with 10,000 hours of audio, totaling 30,000 hours.
- Natural and diverse: VoxpopuliTTS contains speech data at 16kHz from speakers with various accents and backgrounds, which supports the training of more natural and diverse TTS models.
- High-quality transcripts: Through confidence-based filtering strategy and merging adjacent audio segments, the processing pipeline utilizes enhances the speech quality.

2 Speech data processing pipeline

This section provides a detailed description of the speech data processing pipeline for VoxpopuliTTS, which primarily involves data pre-processing and filtering low-quality speech-text pairs. The data pre-processing stage includes voice activity detection (VAD), automatic speech recognition (ASR), word alignment, speaker diarization, and merging adjacent audio segments. The process of filtering low-quality speech-text pairs involves assessing speech quality and ASR confidence.

2.1 Data pre-processing

VAD. Due to the long duration of audio in Voxpopuli dataset, which can range from a few minutes to several hours, the first step in speech data processing is to use a VAD model [1] to segment long-term speech into shorter speech segments, while filtering out unvoiced segments.

ASR. The multilingual FasterWhisper [2] is adopted to process speech segments and predicts the language type and transcripts. The Voxpopuli dataset provides the language type for each audio file. However, Whisper may, in rare instances, produce incorrect language types, resulting in erroneous transcripts. We will filter out these speech-text pairs that contain language recognition errors. Following WhisperX (Bain et al., 2023), Wav2vec 2.0 (Baevski et al., 2020) for the specific language is applied to align words with audio frames through the CTC path, allowing us to obtain the start and end timestamps for each word.

Speaker Diarization. After generating the transcript by the FasterWhisper model, the speaker diarization model (Plaquet and Bredin, 2023; Bredin, 2023) [3] is applied to identify the speaker for each word and speech segment. This allows for grouping the speech segments that belong to the same speaker together.

Merging adjacent audio segments. We observe an interesting phenomenon within FasterWhisper: when converting speech segments into transcripts, it tends not to append end punctuation (e.g., '.') at the end of a transcript if the current speech segment doesn't constitute a complete utterance. Only after processing the entire utterance does FasterWhisper output a transcript that ends with '.', '?' or other end punctuation. Therefore, the punctuation at the end of the transcript can be used to determine whether the semantic information in the transcript is complete.

This insight suggests that we can merge multiple transcripts using just end punctuation, resulting in a final transcript that maintains complete semantic information, and avoids potential disruptions caused by issues with the VAD model or unvoiced frames.

As illustrated in Figure 1, when the VAD model divides speech into two segments, popular speech data processing pipelines might treat these as separate speech segments, generating two transcripts "*I have supported this report,*" and "*and effective consumers enforcement policy is central to the functioning of the single market.*". These transcripts lack complete semantic information, which can confuse downstream TTS models and degrade their performance in speech generation. In our proposed solution, the two speech segments are formed to an utterance, and the final transcription is comprised of these 2 transcripts.

[1] https://huggingface.co/pyannote/voice-activity-detection

[2] https://huggingface.co/Systran/faster-whisper-large-v3

[3] https://huggingface.co/pyannote/speaker-diarization-3.1

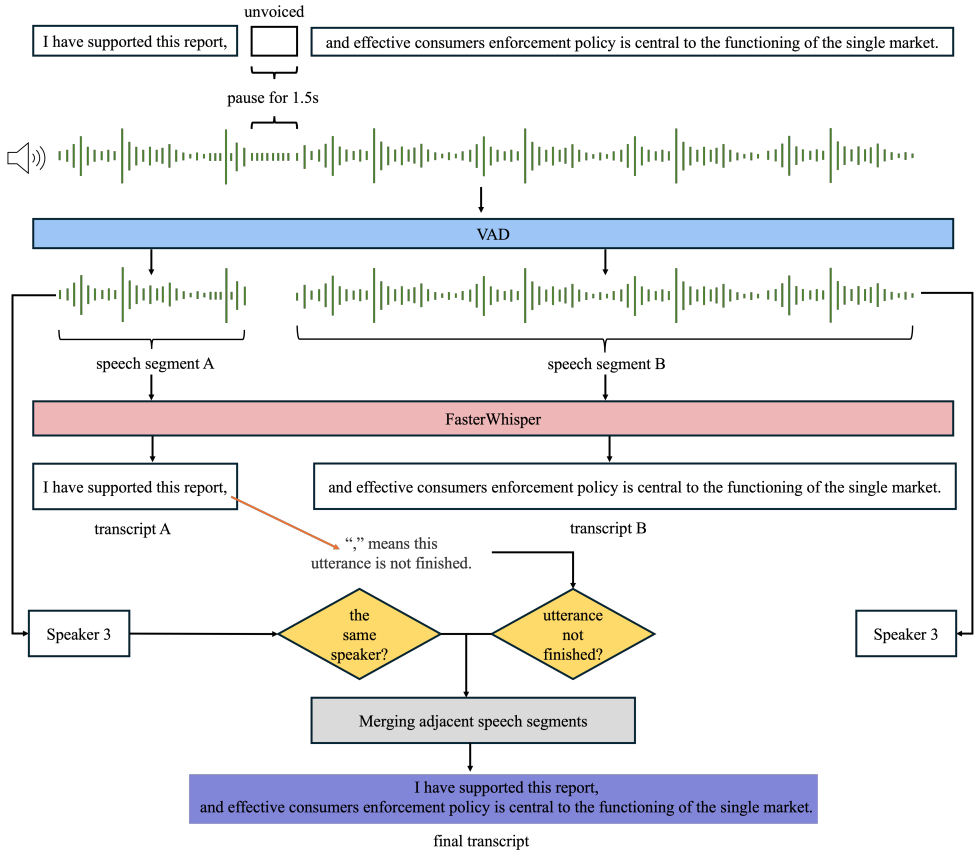

Figure 1: Illustration of merging adjacent speech segments. The input audio signals contains a 1.5s pause, so that the VAD model divides the speech into two speech segments. It is found that speech segments A and B belong to the same speaker by the speaker diarization model. Moreover, transcript A ends with ",", indicating that its semantic information is incomplete and it can be merged with the next transcript until the new transcript ends with end punctuation (e.g., '.'). At last, the final transcript is generated, and the final utterance is comprised of the two speech segments. This approach prevents the erroneous separation of a speech with complete semantic information into two separate utterances.

2.2 Filtering low-quality speech-text pairs

Filtering by DNSMOS metric. Voxpopuli dataset mainly records speeches from speakers of various backgrounds and countries. However, the audio quality in these recordings can be lower than desired, which affects both the ASR model's recognition and the TTS model's training process. To filter out low-quality audio, we employed DNSMOS (Reddy et al., 2021) to evaluate the naturalness of the speech recordings. Any audio with a naturalness score below 3.0 is discarded. This filtering strategy results in the removal of 30% of the audio in English, French, and Spanish.

Filtering by ASR confidence. We find that when FasterWhisper converts speech to transcript, lower logits often correlate with inaccurate transcripts. In other words, the logits output by FasterWhisper during transcription can be regarded as the confidence level of the transcript; the lower the confidence value, the less reliable the transcript is. We defined three confidence metrics, including the average logits of the first three words, the average logits of the last three words, and the average logits of all words, respectively. During our data processing, we find that these confidence metric effectively filter out hallucinated transcripts. We set the confidence level below 0.7.

3 VoxpopuliTTS dataset

We refine Voxpopuli dataset and then propose VoxpopuliTTS dataset, which includes English, French, and Spanish, totaling approximately 30,000 hours. As shown in Table 1, We have divided VoxpopuliTTS into three subsets of 1000 hours, 5000 hours, and 10000 hours as small subset, medium subset and large subset. The 1000-hour subset shows higher quality, which can be used for fine-tuning the TTS model. And the 10000-hour subset

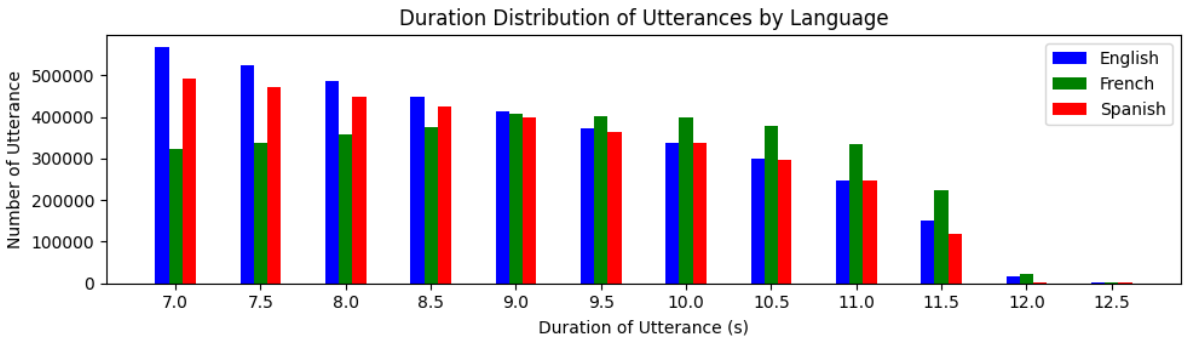

Figure 2: The distribution of durations in VoxpopuliTTS large subset.

Table 1: VoxpopuliTTS dataset. The training data for each language includes three subsets: small, medium, and large. Each language has a validation set and a test set that each retains 100 hours.

Language	Subset	Duration	DNSMOS
en	small	1000h	3.86
	medium	4902h	3.69
	large	10003h	3.30
fr	small	997h	3.88
	medium	5026h	3.72
	large	9982h	3.43
es	small	983h	3.89
	medium	4782h	3.73
	large	8902h	3.44

can be used for training of the large-scale TTS model. Figure 2 shows the duration distribution of the large subset in VoxpopuliTTS dataset. There are a significant number of longer speech segments, which is beneficial for training a better TTS model.

4 Experiments

Evaluation metrics. The Whisper model is utilized to transcribe the generated speech and calculate the Word Error Rate (WER). To evaluate speaker similarity, 3D-speaker toolkit is used to extract speaker embeddings from the generated speech and reference speech, and then we compute the cosine similarity between the normalized embeddings. UTMOS (Saeki et al., 2022) is employed as an automatic MOS system to assess the naturalness of the speech.

TTS baseline model. XTTS_v2(Casanova et al., 2024) takes the mel-spectrogram as input, and adopts VQ-VAE to discretize the mel-spectrogram with a codebook consisting of 8192 codes. We train the autoregressive GPT model in the XTTS model on the large subset of VoxpopuliTTS, to evaluate the effectiveness of VoxpopuliTTS. As shown in Table 2, after training on VoxpopuliTTS, the XTTS model has demonstrated cross-lingual TTS capabilities. The high Word Error Rate (WER) reflects the complexity of VoxpopuliTTS, which increases the difficulty of training for speech generation.

Table 2: Zero-shot evaluation of the XTTS model in VoxpopuliTTS testset.

language	WER	SIM	UTMOS	MOS
en	8.25	67.02%	3.43	3.37 ± 0.09
fr	8.98	67.45%	3.62	3.52 ± 0.05
es	9.20	71.69%	3.54	3.31 ± 0.12

5 Conclusion

In this paper, we present VoxpopuliTTS, a multilingual, multi-speaker, and multi-style TTS dataset based on the Voxpopuli dataset. This dataset is designed to facilitate various TTS downstream tasks, including zero-shot learning. We utilized open-source tools to refine Voxpopuli dataset and introduce the confidence-based filtering strategy, along with merging adjacent audio segments, to eliminate low-quality speech-text pairs and to integrate transcripts based on semantic information. This ensures the high quality and semantic richness of the audio in the dataset. Experimental results demonstrate the effectiveness of our dataset.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech

transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, and Junbo Zhang. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv e-prints*.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. *arXiv preprint arXiv:2407.05361*.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2024. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36.

Linhan Ma, Dake Guo, Kun Song, Yuepeng Jiang, Shuai Wang, Liumeng Xue, Weiming Xu, Huan Zhao, Binbin Zhang, and Lei Xie. 2024. Wenetspeech4tts: A 12,800-hour mandarin tts corpus for large speech generation model benchmark. *arXiv preprint arXiv:2406.05763*.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multiclass cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Christophe Veaux, Junichi Yamagishi, and Kirsten Macdonald. 2017. Cstr vctk corpus: English multispeaker corpus for cstr voice cloning toolkit.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech.