LEMAS:

A 150K-Hour Large-scale Extensible Multilingual Audio Dataset with Generative Speech Models

Zhiyuan Zhao[†], **Lijian Lin**, **Ye Zhu**, **Kai Xie**, **Yunfei Liu**, **Yu Li**

International Digital Economy Academy (IDEA)
[†]Project Lead

The development of generative speech models is increasingly constrained by the trade-off between data scale and quality. While recent large-scale web-crawled corpora provide abundant audio, they often lack reliable quality and fine-grained annotations, which are critical for precision-demanding tasks, including text- and prompt-driven speech editing and duration-controllable synthesis. To address this bottleneck, we present the **LEMAS-Dataset**, which, to our knowledge, is currently the largest open-source multilingual speech corpus with word-level timestamps. Covering over 150,000 hours across 10 major languages, LEMAS-Dataset is constructed via a efficient data processing pipeline that ensures high-quality data and annotations. Unlike simple web-crawled collections, our pipeline integrates rigorous forced alignment and a dynamic confidence-based filtering mechanism to exclude samples with unreliable boundaries, thereby guaranteeing temporal alignment accuracy. To validate the effectiveness of LEMAS-Dataset across diverse generative paradigms, we train two benchmark models with distinct architectures and task specializations on this dataset. **LEMAS-TTS**, built upon a non-autoregressive flow-matching framework, leverages the dataset's massive scale and linguistic diversity to achieve robust zero-shot multilingual synthesis. Our proposed accent-adversarial training and CTC loss mitigate cross-lingual accent issues, enhancing synthesis stability. Complementarily, **LEMAS-Edit** employs an autoregressive decoder-only architecture that formulates speech editing as a masked token infilling task. By exploiting precise word-level alignments to construct training masks and adopting adaptive decoding strategies, it achieves seamless, smooth-boundary speech editing with natural transitions. Experimental results demonstrate that models trained on LEMAS-Dataset deliver high-quality synthesis and editing performance, confirming the dataset's quality. We envision that this richly timestamp-annotated, fine-grained multilingual corpus will drive future advances in prompt-based speech generation systems.

Correspondence: Zhiyuan Zhao at zhaozhiyuan@idea.edu.cn
Project page: https://lemas-project.github.io/LEMAS-Project
Dataset: https://huggingface.co/LEMAS-Project
Code: https://github.com/LEMAS-Project

1. Introduction

The field of speech synthesis has witnessed a transformative era driven by the emergence of Generative Foundation Models. Moving beyond specific-task architectures, recent systems utilize large-scale pre-training on massive unlabelled audio to achieve remarkable zero-shot capabilities. Two dominant paradigms have emerged: Discrete modeling, represented by neural codec vocoders like AudioLM [4] and VoiceCraft [26], treats speech synthesis as a token continuation task; Continuous modeling, represented by Flow Matching systems like Voicebox [21] and F5-TTS [9], models the distribution of mel-spectrograms via differential equations. Recent generative speech models increasingly adopt masked infilling and context-conditioned generation paradigms, which have demonstrated strong effectiveness when trained on large-scale, weakly supervised speech data. However, as model capacity continues to grow, the primary bottleneck has shifted from model architecture to data—specifically, the scarcity of high-quality, large-scale, and fine-grained multilingual corpora.

Limitations of Existing Speech Datasets. While recent efforts have substantially scaled up publicly available speech resources, several fundamental limitations remain. Large-scale corpora such as Libri-Light [18], GigaSpeech [7], and WenetSpeech [41] provide abundant audio–text pairs, enabling the training of robust

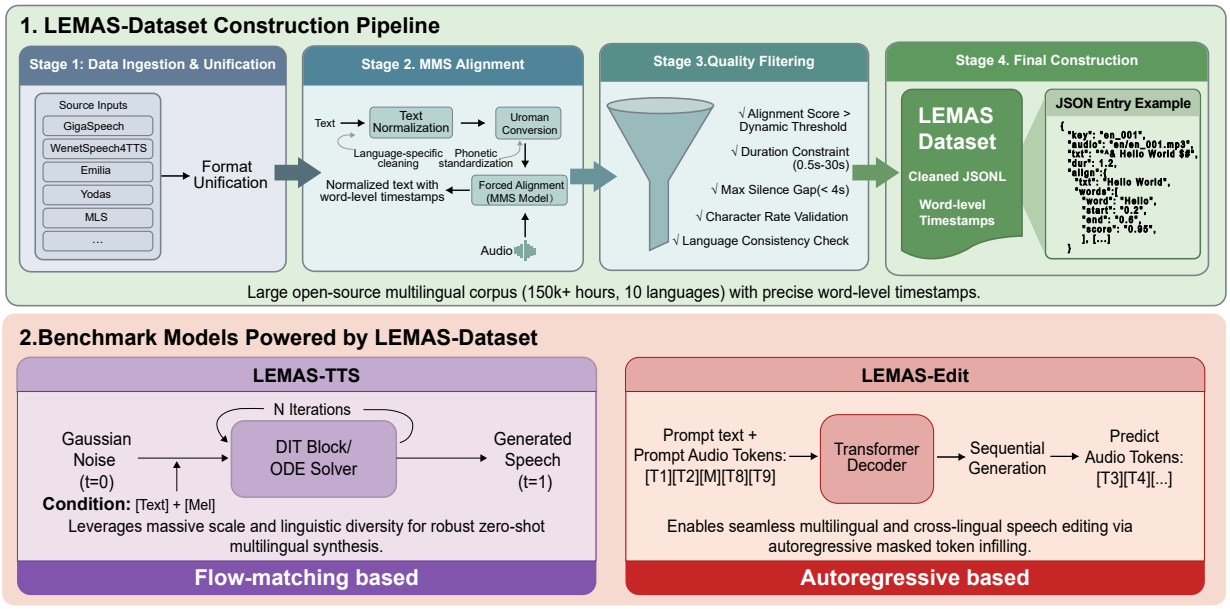

Figure 1 - LEMAS-Dataset contains more than **150k** hours of multi-speaker speech with forced word-level alignments across **10** major languages. Based on LEMAS-Dataset, we train two models. **LEMAS-TTS** performs large-scale, flow-based neural TTS that streams high-fidelity speech from text and a short reference clip, while **LEMAS-Edit** performs codec-based, word-level speech editing.

ASR and TTS models. However, these datasets are predominantly monolingual, focusing on either English or Chinese, and do not address the scarcity of large-scale multilingual resources. Conversely, multilingual datasets such as TEDx [30] and MLS [27] are constrained by limited scale and narrow domain coverage. Specifically, TEDx consists mainly of public speeches, whereas MLS is restricted to the audiobook domain. More recently, large-scale multilingual collections such as YODAS [22] have been introduced, but they rely heavily on minimally filtered web-crawled data, resulting in highly variable quality and a lack of reliable fine-grained annotations.

We also acknowledge promising efforts such as WenetSpeech4TTS [23], which reprocesses WenetSpeech with rigorous cleaning and introduces timestamp-level annotations tailored for speech synthesis. While effective, such datasets remain confined to a single language. In contrast, our goal is to extend this level of data quality and fine-grained temporal supervision to a truly multilingual setting, motivating the construction of LEMAS-Dataset. Besides, built upon LEMAS-Dataset, we design two foundation models: LEMAS-TTS and LEMAS-Edit (Figure 1).

LEMAS-Dataset. To bridge this gap, we introduce LEMAS-Dataset. To the best of our knowledge, LEMAS-Dataset is the largest open-source multilingual speech corpus with rigorous word-level timestamps, comprising over **150,000 hours** of speech across **10 major languages**, including Chinese, English, Russian, Spanish, Portuguese, German, French, Italian, Indonesian and Vietnamese. Table 1 highlights LEMAS-Dataset's advantages in both scale and temporal alignment precision compared to existing corpora.

Constructing such a dataset from in-the-wild audio poses unique challenges, as traditional HMM-based forced alignment tools (e.g., MFA [24]) often struggle with noisy, unsegmented speech. We address this challenge through a robust multi-stage processing pipeline that leverages the MMS aligner [28] to extract word-level timestamps and, crucially, assigns a confidence score to each aligned word. This design enables flexible, reliability-aware data filtering, allowing researchers to dynamically trade off data scale and alignment precision according to downstream requirements, and ensuring high utility for precision-critical applications.

Table 1 - Comparison of LEMAS-Dataset with existing large-scale processed speech datasets. Audiobook data refers to professionally recorded, script-aligned speech with clean acoustic conditions, while in-the-wild data consists of naturally occurring speech collected from real-world scenarios, often exhibiting background noise, spontaneous speaking styles, and imperfect text-audio alignment. LEMAS-Dataset provides extensive total duration and broad language coverage among datasets that include utterance-level timestamps across diverse data sources.

Dataset	Total Duration (h)	Languages	timestamps	Data Source
Libri-light [18]	60k	1	×	Audiobook
GigaSpeech [7]	10k	1	×	Audiobook/In-the-wild
GigaSpeech2 [40]	28k	3	×	Audiobook/In-the-wild
WenetSpeech [41]	22k	1	×	In-the-wild
WenetSpeech4TTS [23]	13k	1	✓	In-the-wild
MLS [27]	51k	8	×	Audiobook
Emilia [17]	102k	6	×	In-the-wild
Voxbox [36]	103k	2	×	Audiobook/In-the-wild
LEMAS-Dataset (ours)	**150k**	**10**	✓	**Audiobook/In-the-wild**

To demonstrate the versatility of LEMAS-Dataset across diverse generative paradigms, we develop two parallel foundation models, utilizing the dataset's scale and precision, respectively:

LEMAS-TTS. Without language-specific alignment regularization, flow-matching models suffer from alignment drift that degrades intelligibility in low-resource languages, as well as accent leakage that replaces target prosody with that of higher-resource languages [9], ultimately undermining zero-shot cross-lingual fidelity. To address these challenges, we extend the F5-TTS [9] architecture with language-aware stabilization mechanisms. Specifically, we unify diverse writing systems into a shared phonetic space and introduce two complementary objectives: a Connectionist Temporal Classification (CTC) loss [16] to improve speech intelligibility, and an accent-adversarial loss [15] to suppress cross-lingual accent interference. These enhancements enable the model to effectively leverage the full 150k-hour corpus, achieving both high speaker similarity and naturalness in zero-shot cross-lingual synthesis.

LEMAS-Edit. To further validate the effectiveness of LEMAS-Dataset for autoregressive speech generation and editing, we develop LEMAS-Edit, a specialized editing model built upon the decoder-only architecture. VoiceCraft [26] formulates speech editing as a masked token-infilling task, whose generation quality is highly sensitive to boundary precision and decoding stability. Misaligned edit boundaries can lead to noticeable artifacts, such as prolonged silences or unnatural acoustic discontinuities [26]. To mitigate repetition artifacts commonly observed during autoregressive decoding, we introduce a repetition penalty at inference time to suppress token looping. We further extend the editing pipeline with system-level components, including denoising, re-alignment using MMS, and consistency verification, enabling precise and robust user-controlled editing. As demonstrated in our demos, LEMAS-Edit achieves seamless, smooth-boundary editing and remains effective on in-the-wild audio with environmental noise or background sounds, highlighting the strong generalization enabled by LEMAS-Dataset.

In summary, our contributions are threefold:

- **LEMAS-Dataset:** We release the largest open-source multilingual speech corpus with rigorous word-level timestamps, comprising over 150,000 hours across 10 languages. The dataset is constructed via a scalable alignment pipeline with word-level confidence estimation, enabling reliable, fine-grained temporal supervision for precision-critical speech generation tasks.
- **LEMAS-TTS:** We develop a non-autoregressive flow-matching TTS model for large-scale multilingual training. By unifying scripts into a shared phonetic space and introducing CTC-based alignment regularization and accent-adversarial objectives, LEMAS-TTS achieves robust zero-shot cross-lingual synthesis with improved intelligibility and accent consistency.
- **LEMAS-Edit:** We develop an autoregressive decoder only speech editing system with stabilized inference through repetition penalty and speech-rate-based anomaly detection, supported by a robust pipeline with denoising, re-alignment, and verification. LEMAS-Edit enables seamless, smooth-boundary multilingual editing and generalizes well to in-the-wild noisy audios.

2. Related Work

2.1 Large-Scale Speech Corpora

The evolution of speech datasets has been characterized by a transition from curated, read speech to massive, multi-domain collections harvested from the web. This progression addresses the growing hunger of end-to-end models for data scale and diversity.

Curated and Crowdsourced Foundations. Early multilingual corpora like GlobalPhone [31] and IARPA BABEL [14] set the standard for high-quality, curated recordings, though they were limited in scale. Mozilla Common Voice [2] later democratized data collection via crowdsourcing, prioritizing speaker diversity. In the domain of audiobooks, LibriSpeech [25] and Multilingual LibriSpeech (MLS) [27] became the de facto benchmarks for ASR, offering clean read speech suitable for reproducible research.

Massive-Scale Multi-Domain Corpora. To overcome the domain limitations of audiobooks, researchers turned to large-scale internet sources. GigaSpeech [7] introduced an evolving, multi-domain English corpus derived from audiobooks, YouTube and podcasts, providing 10,000 hours of high-quality labeled audio. This shift marked a move towards modeling diverse, "in-the-wild" recording conditions. In the landscape of Chinese speech, WenetSpeech [41] (and its recent derivative WenetSpeech4TTS [17]) aggregated tens of thousands of hours of multi-domain Mandarin speech, further validating the efficacy of web-mined data for training industrial-grade ASR and TTS models.

Multilingual and Large-Scale unsupervised Speech Corpora. Recent efforts have greatly expanded multilingual speech datasets in scale and language coverage, focusing on weakly supervised or unlabeled data. VoxPopuli [35] offers 400k hours of unlabeled speech in 23 languages from European Parliament recordings. YODAS [22] collects 500k+ hours from 100+ languages on YouTube, using automatic captions with variable quality. These datasets trade label quality for massive scale, fueling self- and semi-supervised speech research.

2.2 Generative Speech Models

Neural TTS has shifted from task-specific models to unified generative frameworks. It evolved from autoregressive models like Tacotron [38] to non-autoregressive FastSpeech [29], and then to end-to-end variational models like VITS [20]. Today's frontier features large-scale foundation models unifying synthesis, cross-lingual cloning, and editing.

Generative Foundation Models. Speech generation falls into discrete, continuous, and hybrid paradigms, differing in representation and generation methods.

- **Discrete approaches** model speech as sequences of codec tokens autoregressively or via masked infilling, similar to language models. VALL-E [8] pioneered zero-shot TTS using EnCodec tokens; AudioLM [4] adds hierarchical semantic and acoustic token prediction. XTTS [5] extends this to many languages. Recent works focus on efficiency and control, e.g., MaskGCT [37], FishAudio [34], and FireRed-TTS 2 [39].
- **Continuous approaches** generate mel-spectrograms or waveforms directly from noise using diffusion or flow matching. NaturalSpeech 2 [33] demonstrates high-fidelity latent diffusion; VoiceBox [21] simplify pipelines by dropping external encoders. F5-TTS employs a DiT structure with fill-in-the-middle, achieving competitive zero-shot synthesis with a streamlined design.
- **Hybrid approaches** combine discrete semantic modeling with continuous acoustic refinement for diverse and faithful synthesis. CosyVoice 1–3 [10, 11, 12] merge LLM-based semantic tokens with conditional flow matching for multilingual instruction-following TTS; FireRed-TTS [39] cascades discrete and flow-based models for prosody control. IndexTTS 2 [42] further explores semantic compression and style manipulation efficiently.

Multilingual Extensions. With growing training corpora, TTS shifted from mono/bilingual to multilingual. Early open-source models like XTTS-v2 [5] trained on 17 languages jointly. Community models such as OpenAudio-S1-mini [34] support 13 languages with 100k+ hours, enhancing cross-lingual generalization. At a

larger scale, CosyVoice 3 [12] uses a speech tokenizer from a large audio understanding model, training on over a million hours across 9 languages and 18 Chinese dialects. This reflects rapid data scaling and demand for massive multilingual datasets.

Text-Based Speech Editing. Speech editing modifies spoken content while preserving timbre and is now seen as a TTS subtask. VoiceBox [21] shows versatile in-context learning for zero-shot TTS, editing, style conversion, and diverse generation. Non-autoregressive MaskGCT [37] uses mask-and-predict learning; F5-TTS [9] employs a Diffusion Transformer (DiT) supporting TTS and editing jointly. Autoregressive VoiceCraft [26] combines causal masking and delayed stacking for flexible generation and diverse editing within sequences.

3. LEMAS-Dataset

We introduce LEMAS-Dataset, a massive open-source multilingual speech corpus with over 150,000 hours spanning 10 major languages: Chinese (Zh), English (En), Russian (Ru), Spanish (Es), Indonesian (Id), German (De), Portuguese (Pt), Vietnamese (Vi), French (Fr), and Italian (It). LEMAS-Dataset provides word-level alignments along with confidence scores for each word. This fine-grained annotation facilitates a wide range of applications, including data quality filtering based on confidence scores, training duration prediction models, and generating detailed prompts to guide speech synthesis and understanding.

3.1 Dataset Processing Pipeline

Constructing a large-scale multilingual speech corpus for generative modeling requires more than merging existing datasets, which vary widely in annotations, transcription quality, and acoustics. The key insight of the LEMAS-Dataset pipeline is to separate *language diversity* from *annotation fragility*, retaining only samples with reliably grounded temporal structure. To achieve this, we develop a multi-stage pipeline that unifies heterogeneous sources, performs robust multilingual alignment, and applies scalable quality control.

Data Ingestion and Structural Unification. We first aggregate audio–text pairs from a diverse collection of publicly available speech corpora, including GigaSpeech (English) [7], GigaSpeech2 (Indonesian, Vietnamese) [40], WenetSpeech4TTS (Chinese, VoiceBox-based) [23, 36], Emilia (Chinese and English, VoiceBox-based) [17], MLS (English, German, French, Portuguese, Spanish, Italian) [27], multilingual TEDx (German, French, Portuguese, Spanish, Italian) [30], Alcaim (Portuguese) [6], Golos (Russian) [19], and Yodas (German, French, Spanish, Portuguese, Russian, Italian, Indonesian, Vietnamese) [22]. These sources span a wide range of recording scenarios, including audiobooks, podcasts, and in-the-wild conversational speech. Rather than preserving dataset-specific schemas, we normalize all inputs into a unified data representation, enabling language-agnostic downstream processing and large-scale automation.

Multilingual MMS Alignment via Romanized Text. Reliable word-level timestamps are a prerequisite for both TTS training and speech editing tasks, yet traditional forced alignment pipelines often rely on language-specific lexicons and G2P systems, which are brittle under multilingual, noisy, and code-switched conditions. To avoid these limitations, we adopt the Multilingual MMS Forced Aligner [1], a wav2vec-based CTC alignment model [3, 16] provided by torchaudio [2]. This model was trained on over 23,000 hours of speech spanning more than 1,100 languages as part of the "Scaling Speech Technology to 1,000+ Languages" initiative [28].

Prior to alignment, transcripts are language-specifically normalized and then romanized using tools such as Uroman [3], which maps diverse scripts (e.g., Chinese characters, Cyrillic) into a shared Latin representation. This design removes the dependency on language-specific pronunciation dictionaries and allows the alignment model to robustly handle rare words, named entities, and intra-sentential code-switching. The MMS aligner then produces word-level boundaries along with token-level posterior probabilities, which form the basis for subsequent quality control.

Confidence-Driven Quality Filtering. A central principle of our pipeline is that *successful alignment alone is insufficient*; only alignments with high temporal reliability are suitable for generative speech modeling.

[1] https://docs.pytorch.org/audio/2.3/tutorials/forced_alignment_for_multilingual_data_tutorial.html
[2] https://github.com/pytorch/audio
[3] https://github.com/isi-nlp/uroman

We therefore retain only samples for which word-level alignments can be extracted, and apply a series of rule-based filters grounded in alignment confidence and temporal consistency:

- **Alignment Confidence Filtering.** For each utterance, we compute an average alignment confidence score based on the posterior probabilities of aligned tokens. To accommodate variability across languages and source corpora, we apply dataset-specific thresholds $x \in [0.2, 0.5]$, retaining only samples whose confidence exceeds the corresponding threshold.
- **Duration and Pause Constraints.** Utterances shorter than 0.5 seconds or longer than 30 seconds are removed to ensure compatibility with TTS training. Additionally, samples containing prolonged unaligned regions or silence intervals exceeding 4 seconds are discarded, as such patterns often indicate segmentation or transcription errors.
- **Speech Rate Normalization.** To exclude abnormally fast or slow speech, we enforce a language-specific constraint on the ratio between textual length and audio duration. For each language L, the character-based ratio len(text)/duration must fall within a predefined interval $[\text{min_ratio}_L, \text{max_ratio}_L]$, empirically determined from corpus statistics.
- **Language and Character Validation.** We restrict the dataset to ten target languages: Chinese (zh), English (en), Russian (ru), Spanish (es), Indonesian (id), German (de), Portuguese (pt), Vietnamese (vi), French (fr), and Italian (it). All transcripts are verified to belong to one of these languages, and samples containing unsupported scripts, emojis, or excessive non-linguistic symbols are removed according to language-specific character sets.

All thresholds are selected empirically to balance data scale and alignment precision. Notably, we do not apply explicit speech enhancement or background noise removal during preprocessing. While many TTS systems remain sensitive to acoustic interference, we intentionally preserve in-the-wild characteristics, with the expectation that more advanced generative models will be better equipped to leverage such variability. Consistent with this design choice, examples on our demo page show that the proposed editing model remains effective on noisy recordings and can, to some extent, reproduce acoustic environments consistent with the reference audio, highlighting the practical benefits of retaining realistic background conditions.

Final Representation and Storage. The resulting dataset is stored in a unified JSON-based format, containing a unique sample key, audio file paths, original and normalized transcripts, and word-level timestamps with associated confidence scores. For efficient storage and distribution, all audio files are converted to MP3 format while preserving their original sampling rates.

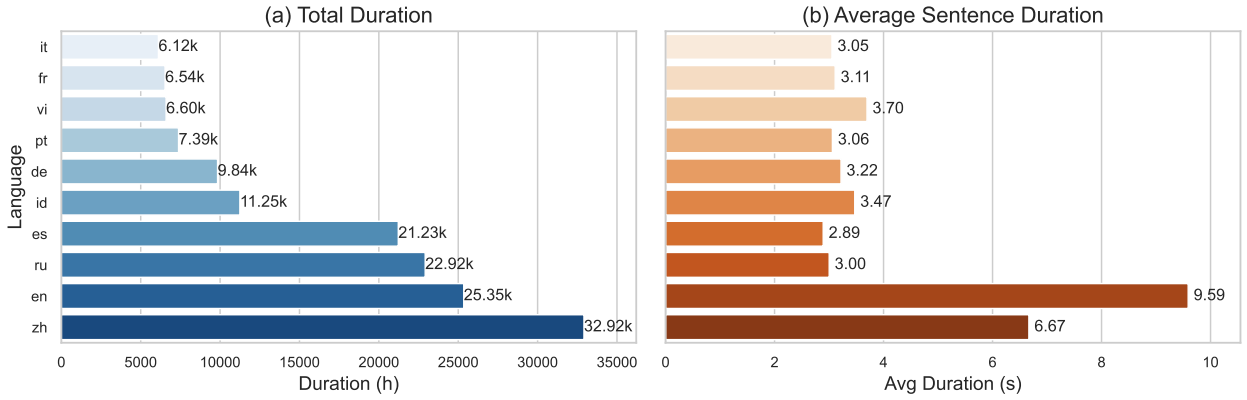

Figure 2 - Dataset Statistics of LEMAS-Dataset. (a) Language-wise duration, (b) the average sentence duration in seconds. The dataset shows substantial variation in both data volume and sentence length across languages.

3.2 Dataset Statistics

The LEMAS-Dataset consists of approximately 150,144 hours of audio data spanning 10 major languages. As illustrated in Figure 2, we explicitly curate the dataset to balance speech duration across languages, resulting in

a relatively uniform distribution despite differences in source data availability. While high-resource languages such as Chinese (Zh) and English (En) remain slightly larger, all languages are represented at a substantial scale. In particular, even the lower-resource subsets (e.g., Italian and Vietnamese) exceed 6,000 hours, which enables effective monolingual training as well as joint multilingual modeling. To facilitate reproducible research and consistent benchmarking, the dataset is partitioned into training (train) and evaluation (eval) subsets.

Training Set. Detailed statistics for the training partition are provided in Table 2. The training set contains over 150,000 hours of audio in total, designed to support the training of large-scale acoustic models. We note that sentence-level statistics vary across languages due to differences in source corpora. Specifically, Chinese and English data are primarily derived from GigaSpeech, Emilia, and WenetSpeech4TTS, which contain longer-form utterances, whereas data for other languages are mainly sourced from YODAS, which consists of shorter, pre-segmented utterances without additional merging or restructuring. As a result, the average utterance length in Chinese and English is noticeably longer than in other languages.

Table 2 - Detailed statistics of the **training** set from LEMAS-Dataset. The units **h**, **s**, and **M** denote hours, seconds, and millions (10^6), respectively. Avg. Duration and Avg. Words indicate average duration and average words per utterance.

Language	Total Duration (h)	Avg. Duration (s)	Utterances (M)	Total Words (M)	Avg. Words
it	6.12k	3.05	7.21	48.86	6.78
fr	6.54k	3.11	7.56	65.84	8.71
vi	6.60k	3.70	6.43	89.90	13.99
pt	7.38k	3.06	8.68	60.32	6.95
de	9.84k	3.22	11.00	80.12	7.28
id	11.25k	3.47	11.66	85.97	7.37
es	21.22k	2.89	26.41	183.67	6.96
ru	22.92k	3.00	27.47	163.02	5.93
en	25.35k	9.59	9.52	268.68	28.24
zh	32.92k	6.67	17.78	496.96	27.96
total	150.14k	4.18	133.71	1,543.34	12.02

Table 3 - Detailed statistics of **evaluation** set from LEMAS-Dataset. The units **min** and **s** denote minutes and seconds, respectively. Avg. Duration and Avg. Words indicate average duration and average words per utterance.

Language	Total Duration (min)	Avg. Duration (s)	Utterances	Total Words	Avg. Words
it	44.22	5.31	500	6,599	13.20
fr	38.17	4.58	500	6,546	13.09
vi	36.74	4.41	500	6,727	13.45
pt	41.69	5.00	500	5,812	11.62
de	38.65	4.64	500	5,599	11.20
id	47.20	5.67	500	6,133	12.27
es	40.52	4.86	500	6,216	12.43
ru	40.24	4.82	500	5,138	10.28
en	67.46	8.10	500	11,325	22.65
zh	75.84	9.10	500	18,627	37.25
total	470.73	5.65	5,000	78,722	15.74

Evaluation Set. To establish a consistent benchmark, we curate a separate evaluation set, as detailed in Table 3. Unlike the training set, the evaluation set is balanced by strictly selecting 500 utterances per language based on rigorous quality criteria. Specifically, we filter for samples with an average word-level alignment score greater than 0.9, a word count exceeding 5, and a duration between 3 and 15 seconds. Additionally, sentence-end silence is trimmed to a maximum of 0.2 s to standardize acoustic characteristics. From the pool of samples that satisfy the above conditions, we further compute language-specific statistics of the character-to-duration ratio and select the 500 utterances closest to the per-language mean. This additional step helps exclude atypical or anomalous cases, such as singing or exaggerated speaking styles, while preserving representative speaking patterns for each language. Finally, to further ensure data quality, approximately 20% of the evaluation samples were manually reviewed to verify that the test set consists exclusively of natural speech. This curation ensures that evaluation metrics (e.g., WER or MOS) are comparable across different languages without being biased by data volume differences or quality variations.

4. LEMAS-TTS

We design a TTS method, LEMAS-TTS, builds upon the F5-TTS architecture [9], utilizing its non-autoregressive Flow Matching framework and Diffusion Transformer (DiT) backbone. While F5-TTS excels at zero-shot tasks, scaling it to a robust multilingual model presents challenges such as alignment stability and effective prosody transfer across languages. To overcome these, we introduce architectural improvements that unify linguistic representations and stabilize the training process.

4.1 Multilingual Training Framework

To facilitate high-fidelity multilingual synthesis, we revise the training paradigm to include a unified phonetic input space, auxiliary supervision objectives, and fine-grained prosody modeling.

Unified Phonetic Representation. To bridge the linguistic gap across diverse languages, we depart from the mixed character/word-level tokenization of the original system. Instead, we adopt a unified phonetic front-end. For Chinese, we employ an initial-final Pinyin representation with tones to capture syllabic structure; for other languages, we utilize the International Phonetic Alphabet (IPA) sequences derived via eSpeak-NG [13]. We choose IPA for these languages because eSpeak-NG's frontend can automatically handle language switching, and using IPA also facilitates verification of correct pronunciation. All phonetic tokens are augmented with explicit language identifiers (e.g., `<zh>`, `<en>`) to project distinct languages into a shared latent space while preserving language-specific phonotactics. Additionally, we leverage word-level timestamps to insert explicit pause tags (`#1`–`#4`) within the phoneme sequences to denote short, medium, long and abnormal pauses, enabling simple yet effective pause control.

Auxiliary Stabilization Objectives. Standard flow-matching objectives can sometimes face convergence difficulties when dealing with complex multilingual scenarios. To address this, we introduce two auxiliary loss functions to improve alignment accuracy and speech intelligibility:

- **CTC Alignment Loss.** To enforce strict monotonicity between the acoustic trajectory and linguistic content, we apply a Connectionist Temporal Classification (CTC) loss on the decoder outputs. This is implemented via a lightweight projection head [16] that maps the predicted mel-spectrograms to phone sequences, explicitly encouraging accurate duration modeling.
- **Accent-Adversarial Disentanglement.** Inspired by IndexTTS2 [42], we aim to disentangle speaker timbre from accent information. We attach an accent classifier to the conditioning pathway and optimize it via a Gradient Reversal Layer (GRL). By supervising this classifier with pseudo-labels from an off-the-shelf language identification model, the encoder is forced to learn accent-invariant representations, thereby reducing cross-lingual accent leakage.

Cross-Lingual Prosody Modeling. To capture fine-grained prosodic variations, we integrate a Prosody Encoder adapted from the Seamless Expressive architecture [32]. Specifically, we extract 80-bin filterbank features from segmented sub-utterances and encode them into a 512-dimensional prosody embedding using pretrained ECAPA-TDNN backbone. These embeddings are aligned with both mel-frames and text tokens and injected into the DiT via linear projection layers. This mechanism provides explicit prosodic supervision, significantly improving the naturalness of prosody transfer across languages.

Design Analysis and Exclusions. During the development of LEMAS-TTS, we also explored explicit speaker contrastive learning (InfoNCE) and clip-and-shuffle data augmentation. However, empirical observations indicated that enforcing strict speaker invariance via InfoNCE conflicted with the flow-matching objective, leading to reduced stability. Similarly, shuffling segmented mel-chunks disrupted the temporal coherence required for the DiT to learn consistent continuations. Consequently, these components were excluded from the final architecture in favor of the alignment and prosody objectives described above.

4.2 Inference Pipeline and Sampling Dynamics

To ensure robust multilingual generation, our inference framework strictly aligns with the training paradigm to minimize distribution shift, while incorporating an adjustable sampling strategy that enables flexible control over sampling dynamics to enhance stability across varying acoustic conditions and speech styles.

Unified Multilingual Front-End. We deploy a standardized text processing pipeline that mirrors the training stage to ensure input consistency. Firstly, The system leverages automatic language identification to apply language-specific normalization rules (e.g., number expansion and punctuation filtering). Then, text is converted into the shared phonetic space defined in our training strategy: Chinese is mapped to tonal Pinyin with initial–final decomposition, while other languages are converted into IPA sequences. Crucially, the system supports explicit insertion of spaces and pause tokens (`#1-#3`) within the text, enabling precise control over pause durations.

Dynamic Guidance and Sway Sampling. The original Sway Sampling strategy in F5-TTS allocates more steps in the early generation phase to improve coarse structural accuracy. However, it performs suboptimally in multilingual and cross-lingual scenarios. To address this, we redesign the time-warping function and classifier-free guidance (CFG) schedule as one-parameter families, as illustrated in Figure 3, enabling smooth interpolation between linear and non-linear behaviors. By increasing early-stage steps and CFG scale, we improve pronunciation accuracy, while reducing CFG strength later allows greater model flexibility for enhanced speaker similarity and audio quality.

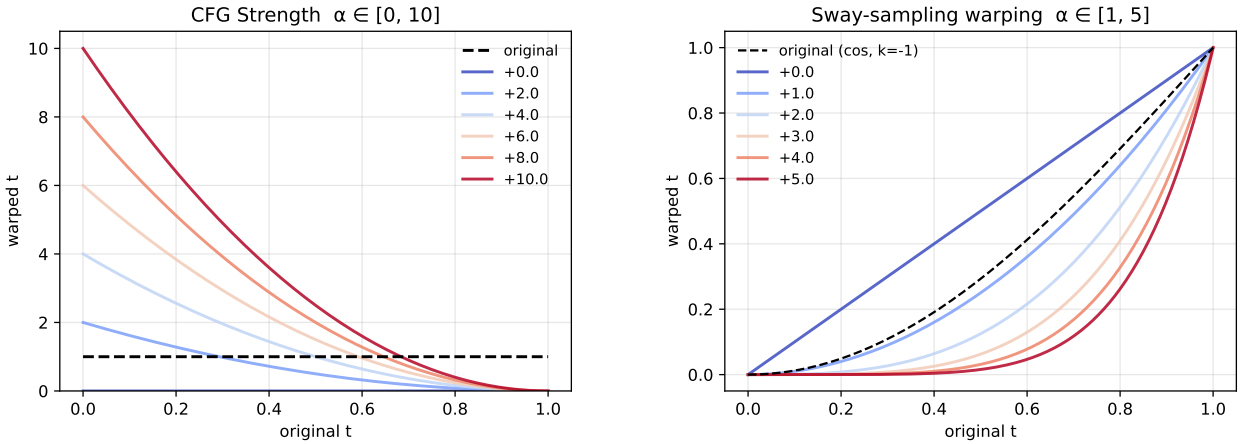

Figure 3 - Time-dependent control schedules used in our sampling strategy. Left: CFG strength schedules with different maximum guidance levels, emphasizing early timesteps and gradually decaying over sampling. Right: Sway-sampling time warping with varying strengths, compared to a cosine-based baseline (dashed), where stronger warping allocates more steps to later timesteps. Together, these schedules demonstrate how guidance strength and sampling allocation can be shaped over time to influence generation behavior.

nificantly enhancing perceptual stability for long, multilingual utterances.

First, we refine the Classifier-Free Guidance (CFG) mechanism. Let $f_\theta^{\text{cond}}(x,t)$ and $f_\theta^{\text{uncond}}(x,t)$ denote the conditional and unconditional flows, respectively. We define the guided flow as:

$$f_\theta^{\text{cfg}}(x,t) \;=\; f_\theta^{\text{cond}}(x,t) \;+\; g(t)\,\big(f_\theta^{\text{cond}}(x,t) - f_\theta^{\text{uncond}}(x,t)\big), \tag{1}$$

where the time-dependent schedule $g(t)$ is defined to decay quadratically as:

$$g(t) \;=\; \lambda\,(1-t)^2, \tag{2}$$

with λ being a scaling factor (e.g., $\lambda = 5$). This schedule applies stronger structural correction during the initial high-noise regime ($t \approx 0$) and gradually relaxes guidance near $t = 1$ to preserve acoustic naturalness, as shown in Equations (1) and (2).

Second, to further optimize the discretization of the flow trajectory, we update the Sway Sampling strategy. Instead of a linear or cosine-based grid, we utilize a power-law reparameterization as Equation (3) represents. Given a uniform grid $s_k \in [0,1]$ for $k = 0,\ldots,K$, the effective time steps are defined as:

$$t_k \;=\; s_k^{1+\gamma}, \tag{3}$$

where $\gamma \geq 0$ is a tunable curvature parameter. Increasing γ effectively stretches the early sampling steps and compresses the later ones. Empirically, this redistribution of computational effort improves the recovery of signal from noise, significantly enhancing perceptual stability for long, multilingual utterances.

Versatile Conditioning Modes. The inference interface is extended to support flexible control over prosody and style. By enabling the prosody encoder, users can extract a global prosody embedding from a reference waveform. This embedding is injected into both the mel and text conditioning streams to facilitate precise style transfer. Additionally, we introduce a reference-free mode that discards conditioning on reference audio, allowing the model to generate speech with randomly sampled speaker characteristics and more native-like pronunciation. This encourages the model to rely on internal priors for timbre and energy, enabling pure text-driven synthesis.

5. LEMAS-Edit

We develop LEMAS-Edit by significantly extending the architecture of VoiceCraft [26], transforming it from a monolingual English model into a multilingual speech editing framework. Our contributions encompass an expansion of training data, a re-engineered distributed training framework, and the introduction of adaptive decoding strategies to ensure generation stability.

5.1 Multilingual Adaptation and Training

Crucially, when the language context switches within a sentence, new language tags are inserted dynamically. This structure allows the autoregressive model to switch its internal linguistic context mid-utterance, naturally supporting mixed-lingual editing tasks.

We have re-engineered the training infrastructure to support large-scale distributed learning. The framework now handles data sharding across multiple heterogeneous datasets and GPU ranks, ensuring balanced data consumption. A significant optimization is the simplification of the optimization step. In the original implementation, gradient accumulation was used to simulate larger batches, which added complexity to the training dynamics. We remove this mechanism entirely in favor of single-step updates on full batches. This modification not only simplifies the codebase but also makes the relationship between batch size and effective gradient updates transparent, enhancing training stability and robustness against memory fragmentation when processing long multilingual sequences on large GPU clusters.

5.2 Inference and Decoding Strategy

History-Aware Repetition Control. Autoregressive codec models often suffer from repetition loops, especially in long-form generation. To mitigate this, we introduce a dynamic, history-aware penalty mechanism within the `top-k/top-p` sampling process. Unlike static penalties, our approach scales with the generation length. We modify the logits of previously generated tokens based on their frequency using the following formulation:

$$\text{penalty} = \frac{\text{repetition_penalty}}{100} \cdot \text{num_gen} + 1. \tag{4}$$

Here, `num_gen` represents the current number of generated tokens. For any token present in the history, its positive logits are divided by this penalty, while negative logits are multiplied by it. This progressively increases the cost of repeating tokens as the sequence grows, effectively suppressing both silence loops and content stuttering without degrading the initial prompt adherence.

Adaptive Re-Generation Mechanism. To ensure that the generated audio duration aligns with natural speech patterns, we implement a multi-round adaptive inference scheme. First, we establish a baseline speaking rate by analyzing the reference audio:

$$\texttt{avg_speed} = \frac{\#\text{encodec frames}}{\#\text{raw text tokens}}. \tag{5}$$

This estimated rate allows us to impose a dynamic `max_rate` constraint on the decoder, preventing run-away generation.

However, simple constraints are insufficient for complex edits. Therefore, we introduce a feedback loop. During generation, if the model triggers an internal anomaly flag (`RE_GEN`) or if the output length is pathologically short (specifically, $< 0.5 \times$ `avg_speed`), the system automatically triggers a re-generation pass. In each subsequent round, we iteratively relax the constraints by expanding the editing mask boundaries (providing more acoustic context) and incrementing the `repetition_penalty`. This iterative process continues until the output matches the expected prosodic structure or a maximum retry limit is reached.

5.3 Multilingual Editing Front-End

We have upgraded the inference front-end to support a professional-grade multilingual workflow, focusing on accuracy and flexibility:

Robust Recognition and Alignment. Precise text-to-audio alignment is essential for accurate speech editing. To enhance transcription robustness, we replace the standard English ASR with the multilingual Whisper `large` model. Additionally, we employ the MMS Forced Aligner to obtain more accurate word-level timestamps. This fine-grained temporal information is crucial for constructing precise editing masks, ensuring modifications are strictly confined to the intended regions without affecting adjacent words.

Signal Enhancement. To handle real-world recordings, the pipeline integrates two distinct denoising backends. Users can select UVR5[4] for mild denoising that preserves background ambience (crucial for maintaining naturalness in edited speech), or DeepFilterNet[5] for aggressive suppression of noise in low-quality inputs.

Long-Form Processing. Processing extended recordings in a single pass often degrades attention performance. Our system automatically segments long audio files into manageable chunks, processes the edits independently, and stitches the results back together using smooth cross-fades at zero-crossing points. This ensures perceptual continuity across the entire recording.

Unified Interface. The system exposes two distinct operational modes within a single interface: *TTS Mode*, which generates speech autoregressively starting from the end of the sentence backwards using reference text and audio, and *Editing Mode*, which modifies existing audio. The latter leverages the LEMAS-Dataset's precise alignment capabilities to allow granular, human-in-the-loop adjustments of start and end times, bridging the gap between automated models and professional editing requirements.

6. Experiment

6.1 Evaluation of LEMAS-TTS

We evaluate LEMAS-TTS on 10 languages and compare it with the contemporaneous open-source baseline OpenAudio-S1-mini[6], which supports the same language set. Word error rate (WER) is evaluated using FunASR Paraformer-zh[7] for Chinese and Whisper-large-v3[8] for all other languages, while speaker similarity (SIM) is computed with WavLM-large[9], following the Seed-TTS evaluation protocol [1]. The evaluation data are constructed by randomly sampling 10 utterances per language from the LEMAS-Dataset eval set and synthesizing speech using target texts drawn from all utterances, resulting in a total of 10,000 synthesized samples that cover both intra-language and cross-language synthesis, enabling an assessment of pronunciation stability under text variation. Due to the lack of native evaluators for languages other than Chinese and English, no subjective evaluation was conducted.

As shown in Table 4, LEMAS-TTS consistently outperforms OpenAudio-S1-mini across all languages, achieving lower WER and higher speaker similarity. We further compare two variants of LEMAS-TTS, with and without a prosody encoder. The prosody-aware model consistently yields lower WER, indicating improved pronunciation stability, while slightly reducing speaker similarity. Subjective listening suggests that prosodic

[4] https://ultimatevocalremover.com
[5] https://github.com/Rikorose/DeepFilterNet
[6] https://huggingface.co/spaces/fishaudio/openaudio-s1-mini
[7] https://huggingface.co/funasr/paraformer-zh
[8] https://huggingface.co/openai/whisper-large-v3
[9] https://huggingface.co/microsoft/wavlm-large

Table 4 - Objective evaluation of TTS models on multilingual test sets. WER (Word Error Rate) measures transcription accuracy (lower is better), while SIM (Similarity) evaluates perceptual closeness to the target speaker (higher is better). Results compare OpenAudio-S1-Mini and LEMAS-TTS models, with and without prosody control, across 10 languages.

Language	Model	WER↓(%)	SIM↑
de	OpenAudio-S1-Mini	3.95	0.448
	LEMAS-TTS w/o prosody	1.20	0.533
	LEMAS-TTS w/ prosody	0.44	0.535
en	OpenAudio-S1-Mini	4.98	0.451
	LEMAS-TTS w/o prosody	2.21	0.557
	LEMAS-TTS w/ prosody	1.07	0.556
es	OpenAudio-S1-Mini	7.66	0.488
	LEMAS-TTS w/o prosody	4.93	0.560
	LEMAS-TTS w/ prosody	4.28	0.552
fr	OpenAudio-S1-Mini	13.41	0.458
	LEMAS-TTS w/o prosody	10.45	0.517
	LEMAS-TTS w/ prosody	9.87	0.518
id	OpenAudio-S1-Mini	32.77	0.518
	LEMAS-TTS w/o prosody	7.05	0.550
	LEMAS-TTS w/ prosody	6.38	0.534
it	OpenAudio-S1-Mini	9.27	0.489
	LEMAS-TTS w/o prosody	5.57	0.544
	LEMAS-TTS w/ prosody	4.43	0.537
pt	OpenAudio-S1-Mini	8.05	0.476
	LEMAS-TTS w/o prosody	3.12	0.555
	LEMAS-TTS w/ prosody	2.09	0.546
ru	OpenAudio-S1-Mini	20.75	0.475
	LEMAS-TTS w/o prosody	12.29	0.545
	LEMAS-TTS w/ prosody	11.84	0.538
vi	OpenAudio-S1-Mini	-	0.451
	LEMAS-TTS w/o prosody	22.14	0.541
	LEMAS-TTS w/ prosody	15.38	0.531
zh	OpenAudio-S1-Mini	9.59	0.550
	LEMAS-TTS w/o prosody	4.64	0.570
	LEMAS-TTS w/ prosody	3.00	0.546
Avg	OpenAudio-S1-Mini	12.27	0.480
	LEMAS-TTS w/o prosody	8.06	**0.547**
	LEMAS-TTS w/ prosody	**6.39**	0.539

conditioning trades expressive richness for increased articulation stability. We therefore release both variants to accommodate different application preferences. For Vietnamese, the WER of OpenAudio-S1-mini remains unusually high across repeated measurements; we therefore exclude this result from the average WER calculation to avoid disproportionate influence on the aggregated results.

6.2 Evaluation of LEMAS-Edit

The original VoiceCraft [26] was trained exclusively on the English GigaSpeech corpus. LEMAS-Edit is trained on selected subsets extracted from the LEMAS-Dataset, WenetSpeech4TTS, GigaSpeech, and MLS. This joint training set covers seven major languages (Zh, En, De, Fr, Pt, Es, It), enabling the model to learn cross-lingual phonotactics and prosody transfer. To efficiently adapt to this multilingual setting, we employ a warm-start initialization strategy. We utilize the pre-trained 330M VoiceCraft checkpoint as a foundation.

We further evaluate LEMAS-TTS and LEMAS-Edit on text-based speech editing. We randomly select

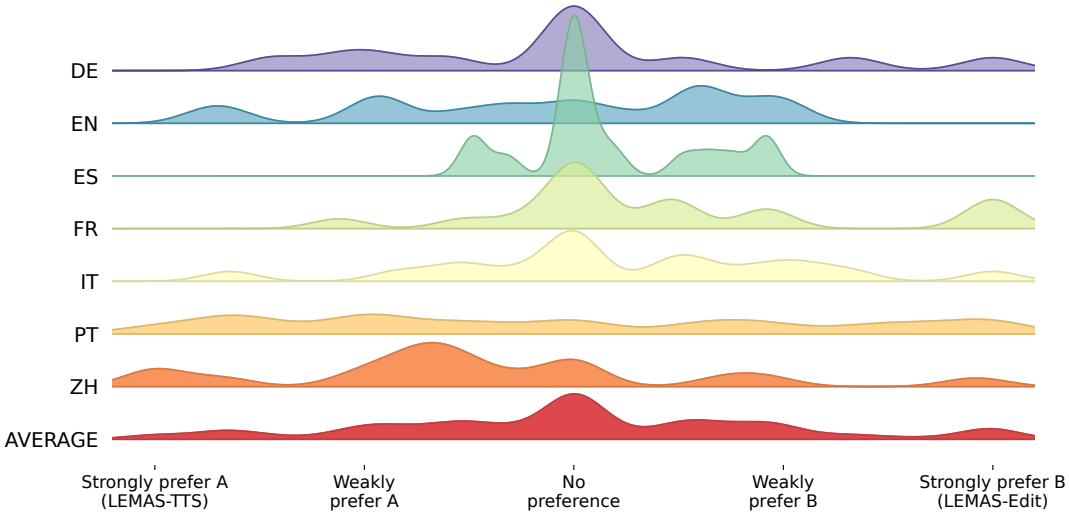

Figure 4 - Preference distribution of audio naturalness across different languages. The ridgeline plot illustrates the subjective results of the A/B preference test comparing *LEMAS-TTS* (Model A) and *LEMAS-Edit* (Model B). Individual scores are normalized to a scale of 0–100, where 0 represents a strong preference for Model A and 100 indicates a strong preference for Model B. The vertical distribution (Kernel Density Estimation) for each language reveals the consensus among users, with the "AVERAGE" row representing the aggregated performance across all tested languages.

20 utterances from the LEMAS-Dataset evaluation set across 7 languages (2–3 per language). For each sample, word-level alignments are provided, and ChatGPT randomly replaces one or more consecutive words to form editing cases. Six listeners conduct an AB test on editing naturalness (Figure 4). Despite minor language-specific biases, preferences are overall balanced, indicating that LEMAS supports robust and architecture-agnostic editing evaluation. All examples are available on the demo page.

7. Conclusion

In this work, we introduced **LEMAS-Dataset**, a large-scale multilingual speech dataset designed to bridge the gap between data quantity and quality in generative speech research. With over 150,000 hours of audio across 10 languages, accompanied by high-quality word-level timestamps and confidence scores, LEMAS-Dataset provides a foundation for training and evaluating next-generation speech generative models. We demonstrated the value of **LEMAS-Dataset** from two complementary perspectives. Through **LEMAS-TTS**, we showed that carefully designed alignment and disentanglement objectives are essential for stabilizing large-scale multilingual flow-based models. Through **LEMAS-Edit**, we validated that precise temporal supervision is a key enabler for robust zero-shot speech editing. Together, these results highlight that progress in generative speech modeling is increasingly driven not only by model architecture, but by the scale, diversity, and granularity of the underlying data. We hope that **LEMAS** will serve as a shared benchmark and a practical resource for the community, catalyzing further research in multilingual speech synthesis, editing, and controllable audio generation. Future work includes extending the dataset to additional languages and speaking styles, as well as exploring tighter integration between alignment-aware data curation and generative model training.

Acknowledgments

We would like to sincerely thank the following individuals for their invaluable help and insightful discussions. The names are listed in alphabetical order by last name: Xie Chen, Chen Cheng, Zhen Guo, Jiaxin Huang, Rui Li, Fei Peng, Daniel van Strien, Tiezhen Wang, Baogang Yao, Fei Yu, Xiaoling Zhang, Changyin Zhou.

References

[1] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhengnan Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuan-Qiu-Qiang Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, WenJie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, pages 12449–12460, 2020.

[4] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533, 2023.

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

[6] CETUC. CETUC Alcaim: European Portuguese speech dataset. URL http://www02.smt.ufrj.br/~igor.quintanilha/alcaim.tar.gz. Accessed: 2025-12-28.

[7] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

[8] Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[9] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

[10] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

[11] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[12] Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shi-Min Zhang, Wen Wang, and Jieping Ye. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

[13] eSpeak NG. espeak ng (version 1.51). https://github.com/espeak-ng/espeak-ng/tree/1.51, 2022. Accessed: 2025-12-25.

[14] Mark JF Gales, Shakti P Rath, et al. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[16] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[17] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE, 2024.

[18] Jacob Kahn, Morgane Riviere, Weiyi Zheng, Eugene Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Likhomanenko, Vineel Pratap, Awni Kahn, et al. Libri-light: A benchmark for asr with limited or no supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673, 2020.

[19] Nikolay Karpov, Alexander Denisenko, and Fedor Minkin. Golos: Russian dataset for speech research. *arXiv preprint arXiv:2106.10161*, 2021.

[20] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

[21] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034, 2023.

[22] Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. Yodas: Youtube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

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

[24] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proc. Interspeech 2017*, pages 498–502, 2017.

[25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[26] Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973*, 2024.

[27] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. In *Interspeech*, 2020.

[28] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sravya Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 24(368), 2023.

[29] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. volume 32, 2019.

[30] Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. The multilingual tedx corpus for speech recognition and translation. *arXiv preprint arXiv:2102.01757*, 2021.

[31] Tanja Schultz and Tim Schlippe. Globalphone: A multilingual text & speech database in 20 languages. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[32] Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. Seamlessm4t-massively multilingual & multimodal machine translation, 2023.

[33] Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *The Twelfth International Conference on Learning Representations*, 2024.

[34] Fish Audio Team. Fish speech: Openaudio s1 mini. https://huggingface.co/fishaudio/openaudio-s1-mini, 2024.

[35] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talwalkar, Elizabeth Xiao, Yatharth Saraf, Piyush Puvvada, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Association for Computational Linguistics (ACL)*, 2021.

[36] Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025.

[37] Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*, 2024.

[38] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. 2017.

[39] Kun Xie, Feiyu Shen, Junjie Li, Fenglong Xie, Xu Tang, and Yao Hu. Fireredtts-2: Towards long conversational speech generation for podcast and chatbot. *arXiv preprint arXiv:2509.02020*, 2025.

[40] Yifan Yang, Zheshu Song, Jianheng Zhuo, Mingyu Cui, Jinpeng Li, Bo Yang, Yexing Du, Ziyang Ma, Xunying Liu, Ziyuan Wang, et al. Gigaspeech 2: An evolving, large-scale and multi-domain asr corpus for low-resource languages with automated crawling, transcription and refinement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2673–2686, 2025.

[41] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE, 2022.

[42] Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*, 2025.