CosyVoice 3: Towards In-the-wild Speech Generation via Scaling-up and Post-training

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, Jieping Ye

Speech Team, Tongyi Lab, Alibaba Group

{neo.dzh, funaudiollm}@alibaba-inc.com

Abstract

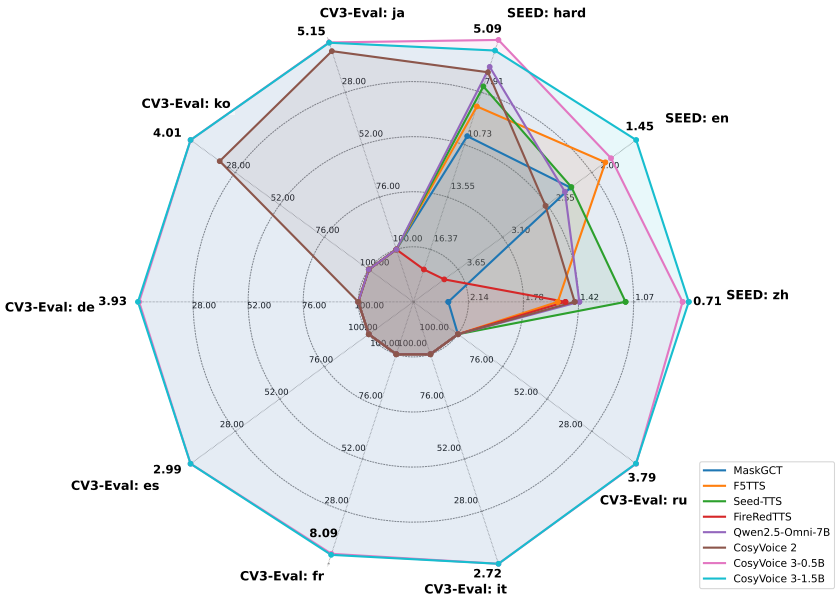
In our prior works, we introduced a scalable streaming speech synthesis model, CosyVoice 2, which integrates a large language model (LLM) and a chunk-aware flow matching (FM) model, and achieves low-latency bi-streaming speech synthesis and human-parity quality. Despite these advancements, CosyVoice 2 exhibits limitations in language coverage, domain diversity, data volume, text formats, and post-training techniques. In this paper, we present **CosyVoice 3**, an improved model designed for **zero-shot multilingual speech synthesis in the wild**, surpassing its predecessor in content consistency, speaker similarity, and prosody naturalness. Key features of CosyVoice 3 include: 1) A **novel speech tokenizer** to improve prosody naturalness, developed via supervised multi-task training, including automatic speech recognition, speech emotion recognition, language identification, audio event detection, and speaker analysis. 2) A **new differentiable reward model for post-training** applicable not only to CosyVoice 3 but also to other LLM-based speech synthesis models. 3) **Dataset Size Scaling**: Training data is expanded from ten thousand hours to one million hours, encompassing 9 languages and 18 Chinese dialects across various domains and text formats. 4) **Model Size Scaling**: Model parameters are increased from 0.5 billion to 1.5 billion, resulting in enhanced performance on our multilingual benchmark due to the larger model capacity. These advancements contribute significantly to the progress of speech synthesis in the wild. We encourage readers to listen to the demo at https://funaudiollm.github.io/cosyvoice3.

1 Introduction

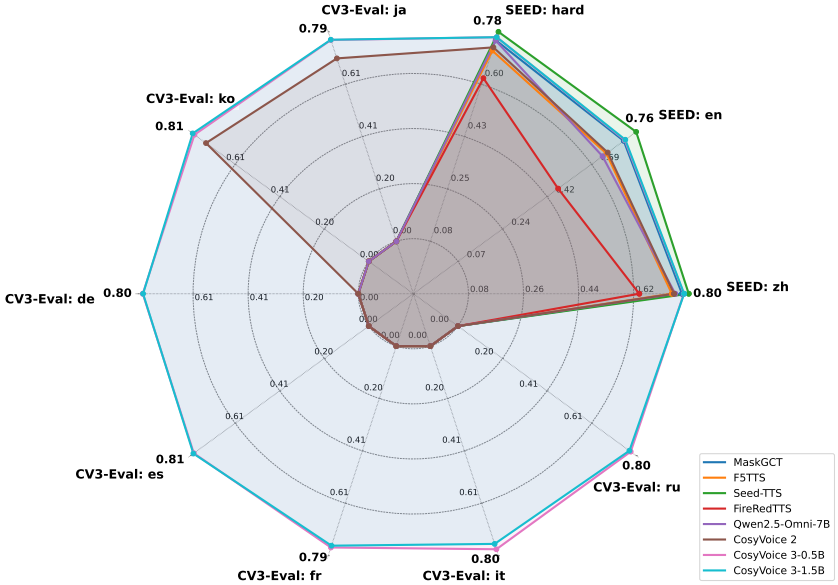
With the rapid development of generative neural networks, text-to-speech (TTS) synthesis has made significant progress, surpassing traditional concatenative and parametric methods in terms of synthetic quality [1–7]. In particular, zero-shot TTS models, which leverage vast multi-speaker datasets, can clone the timbre, prosody, and style of any speaker, and demonstrate performance superior to specific speaker TTS models, achieving human-like prosody naturalness and audio quality [8].

Currently, zero-shot TTS models can be broadly categorized into three types: those using large language models (LLMs) to model discrete acoustic tokens [8–17], those based on diffusion models to automatically learn internal alignments between speech and text [18–26], and coarse-to-fine hybrid systems that use auto-regressive LLMs to model coarse semantics, followed by non-autoregressive models such as diffusion models to render detailed speech features [26–32]. Considering the trade-offs between synthesis quality, streaming compatibility, and flexibility, such two-stage hybrid sys-

Preprint. Work in progress.



(a) Content Consistency



(b) Speaker Similarity

Figure 1: Performance comparison between our CosyVoice 3 and competitive speech generation models in terms of content consistency and speaker similarity on various benchmarks. The numbers in (a) content consistency are CERs or WERs measured by ASR models. The numbers in (b) speaker similarity are cosine similarities of WavLM embeddings between reference and generated utterances. The error rates of 100.00 and the similarities of 0.00 mean that the released models do not support the languages.

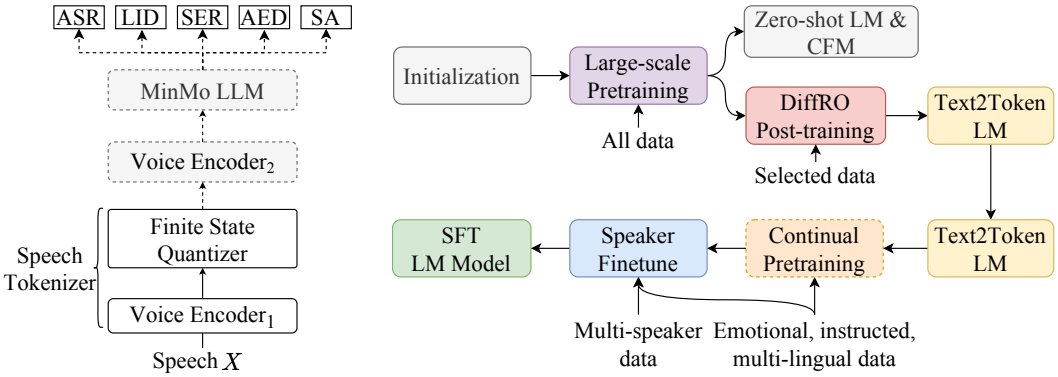
tems have become a mainstream choice in industrial applications. In our previous work, we developed CosyVoice 2 [30]. Through optimizing semantic token utilization, initializing with text-based LLMs, designing a bidirectional streaming scheme, and unifying instruction capability modeling, CosyVoice 2 achieves synthesis quality comparable to human speech, along with ultra-low latency bidirectional streaming synthesis capability that is virtually lossless [30].

Although CosyVoice 2 performs well in general Chinese and English broadcast scenarios, it has noticeable limitations in language coverage, domain diversity, data volume, and text format variety, leaving significant room for improvement towards achieving in-the-wild speech generation. Furthermore, the scaling laws for models and data, as well as post-training techniques suitable for speech generation models, have not been thoroughly explored. To address these issues, we introduce CosyVoice 3, a large zero-shot speech generation model designed for in-the-wild applications, covering more languages and diverse scenarios, and significantly surpassing its predecessor CosyVoice 2 in content consistency, speaker similarity, and prosody naturalness. Our contributions can be summarized as follows:

- We propose a **novel speech tokenizer** derived from a large audio understanding language model. Through supervised multi-task training, such tokenizer enables discrete speech tokens to better capture paralinguistic information such as emotion and pronunciation style.
- We explore post-training strategies suitable for speech generation models and propose a new **differentiable reward optimization (DiffRO)** method, applicable not only to the CosyVoice series but also to other discrete-token-based speech synthesis models.
- We validate **dataset size scaling** in the speech generation domain, expanding the training data from ten thousand hours to one million hours, covering 9 common languages, 18 Chinese accents/dialects, and various text formats, supporting better cross-lingual voice cloning. We also demonstrate the impact of **model size scaling** by increasing the model size from 0.5B to 1.5B, further enhancing the prosody naturalness.
- To address the challenges of diversity and generalizability from unrestrained real-world speech synthesis scenarios, we release the **CV3-Eval** benchmark for zero-shot speech synthesis in the wild, which is built on authentic in-the-wild reference speech from Common Voice, FLUERS, EmoBox, and Web-crawled real-world audio data, and spans a broad range of languages and dialects, domains and environments, emotions and styles.

Through these improvements, CosyVoice 3 achieves state-of-the-art (SOTA) results on multiple benchmarks. We believe that CosyVoice 3 represents a solid step towards in-the-wild speech synthesis.

2 CosyVoice 3



(a) Multi-task speech tokenizer

(b) Training pipeline of CosyVoice 3

Figure 2: Illustrations of (a) Supervised multi-task trained speech tokenizer and (b) The training pipeline in CosyVoice 3. Modules with dashed boxes are only used in the training stage. The speech tokenizer is supervised trained on ASR, language identification (LID), speech emotion recognition (SER), audio event detection (AED), and speaker analysis (SA) tasks. CFM denotes the conditional flow matching model.

Figure 2 illustrates the training procedures for both the supervised multi-task supervised speech tokenizer and the generation models of CosyVoice 3. Different from its predecessor CosyVoice 2, the speech tokenizer in CosyVoice 3 is based on MinMo [33], a pretrained large-scale speech

understanding model demonstrating strong performance across various speech tasks [33]. We also provide an overview of the training pipeline for the zero-shot and speaker fine-tuned (SFT) models, encompassing large-scale pretraining, post-training, continual pretraining, and multi-speaker fine-tuning. The post-training phase is aimed at surpassing the performance limitations of the training data, while the continual pretraining phase focuses on transferring capabilities, such as instruction controllability and multilingual synthesis, from the zero-shot model to the SFT models.

2.1 Speech Tokenizer via Supervised Multi-task Training

As shown in Figure 2a, different from CosyVoice 2 that inserts the Finite Scalar Quantization (FSQ) module [34] into the encoder of the SenseVoice-Large ASR model [35], for CosyVoice 3, we insert the FSQ module into the voice encoder of the MinMo model [33]. Compared to SenseVoice-Large ASR model, MinMo is an advanced multimodal LLM trained on an extensive dataset of over 1.4 million hours of speech, and showcases superior and SOTA performance across diverse benchmarks, including spoken dialogue, multilingual speech recognition, and emotion recognition. To further enhance the ability of capturing semantic information, we leverage a subset of the training data for MinMo to conduct supervised multi-task learning for our speech tokenizer about 530,000 hour, including tasks such as multilingual ASR, language identification (LID), speech emotion recognition (SER), audio event detection (AED), and speaker analysis (SA).

During the training stage, the input speech X goes through the $\mathrm{Voice\ Encoder}_1$ in Figure 2a to obtain the intermediate representations H, where $\mathrm{Voice\ Encoder}_1$ consists of 12 Transformer blocks with rotary positional embedding (RoPE) [36]. The intermediate representations H are then fed into the FSQ module for quantization, and the quantized representations are passed through the rest of MinMo modules, including $\mathrm{Voice\ Encoder}_2$ and $\mathrm{MinMo\ LLM}$, to predict the posterior probabilities of the corresponding text tokens.

In the FSQ module, the intermediate representations H are first projected into a D-dimensional low-rank space, and the values of each dimension are quantized into $[-K, K]$ with the bounded round operation ROUND. Then, the quantized low-rank representations $\bar{H}$ are projected into the original dimension $\tilde{H}$, as follows:

$$\begin{aligned} \bar{H} &= \mathrm{ROUND}(\mathrm{Proj}_{down}(H)) \\ \hat{H} &= \mathrm{Proj}_{up}(\bar{H}) \end{aligned} \tag{1}$$

During the training stage, the straight-through estimation is used to approximate the gradients of the FSQ module and $\mathrm{Voice\ Encoder}_1$. The speech token μ_i is obtained by calculating the index of the quantized low-rank representation $\bar{h}_i$ in the $(2K+1)$-ary system:

$$\mu_i = \sum_{j=0}^{D-1} \bar{h}_{i,j}(2K+1)^j \tag{2}$$

Together the $\mathrm{Voice\ Encoder}_1$, the low-rank projector of the FSQ module, the bounded round operation, and the index calculation form the speech tokenizer of CosyVoice 3. Our speech tokenizer works at a token rate of 25 Hz, i.e., 25 speech tokens per second.

2.2 Reinforcement Learning with Differentiable Reward Optimization

Recent TTS systems [26, 37] have demonstrated that reinforcement learning (RL) is effective in enhancing the quality of generated speech. However, to the best of our knowledge, a generally applicable RL methodology for speech generation has not been established. Unlike LLMs in the NLP task, TTS systems require additional downstream conditional flow matching (CFM) and vocoder models to convert discrete speech tokens into audio waveforms. The computational demands posed by these downstream models are substantial. More seriously, after downstream processing, the resulting voices consistently exhibit high similarity; therefore, it is challenging to differentiate between positive and negative feedback for training the reward model.

In order to address these issues, we introduce the **Differentiable Reward Optimization (DiffRO)** approach to directly optimize the speech tokens rather than the synthesized audio. DiffRO first trains an ASR-like Token2Text model on the ASR training data, then uses the posterior probability as the reward. To further simplify the training strategy, DiffRO uses the Gumbel-Softmax operation

to sample the LLM predicted tokens and then directly optimize the speech tokens to maximize the reward score with back-propagation rather than the RL training loop:

$$\tilde{\mu}_t = \text{GumbelSoftmax} P_{\pi_\theta}(\mu_t | \mu_{1:t-1}; Y) \tag{3}$$

$$R_{ASR}(Y) = \log P_{ASR}(\tilde{Y}_n = Y_n | Y_{1:n-1}; \tilde{\mu}_{1:T}) \tag{4}$$

where μ_t and $\tilde{\mu}_t$ denote the ground-truth speech token and its sampled prediction at timestep t. R_{ASR} is the reward function computed based on the ASR-like Token2Text model. Since $R_{ASR}(Y)$ aims at encouraging $\tilde{\mu}$ to catch all information from the text, it can help the TTS system to comprehend the text clearly and accurately. Therefore, we can directly optimize the LLM to align the output tokens with ASR preference and use the Kullback-Leibler (KL) divergence to prevent the model from deviating too far from the reference model. Different from other RL methods, we compute the KL divergence on the **output token-level logits** rather than on the sequence-level posterior probability.

$$\pi_\theta^* = \max_{\pi_\theta} \mathbb{E}\left[R(Y)\right] - \beta D_{\text{KL}}\left[\pi_\theta(\mu|Y) \| \pi_{\text{ref}}(\mu|Y)\right] \tag{5}$$

$$D_{\text{KL}}\left[\pi_\theta(\mu|Y) \| \pi_{\text{ref}}(\mu|Y)\right] = \sum_{t=1}^{T} \sum_{k=0}^{Q} P_{\pi_\theta}(\mu_t = k) \log \left(\frac{P_{\pi_\theta}(\mu_t = k)}{P_{\pi_{\text{ref}}}(\mu_t = k)} \right) \tag{6}$$

where Q is the codecbook size of the FSQ module and equals to $(2K+1)^{D-1}$.

Besides the Token2Text model, DiffRO also uses other downstream tasks such as SER, MOS score prediction, AED, and other audio understanding tasks for **multi-task reward (MTR)** modeling. The MTR mechanism can help TTS systems to control the voice attributes $\{A_i\}_{i=1}^K$ by following instructions.

$$R_{MTR}(Y, \{A_i\}_{i=1}^K) = \sum_i \log P_{\text{task}_i}(\tilde{A}_i = A_i | \tilde{\mu}) \tag{7}$$

2.3 Pronunciation Inpainting

LLM-based TTS systems predominantly use the BPE text tokenizer, taking raw text as input. Compared to traditional phoneme-based methods, these systems lack controllability in pronunciation. Specifically, when it comes to mispronunciations caused by polyphonic character or rare words that are sparse or do not appear in the training data, there lack robust methods that are based on human intervention.

To achieve an industry-level TTS system that is effectively controllable on pronunciations, we extend CosyVoice 3 to be able to model mixed sequences of words and phonemes with expansion of the vocabulary of tokenizer. To achieve this goal, we construct an auxiliary training set by replacing Chinese **monophonic** characters with pinyin and replacing English **monophonic** words with phonemes using the CMU pronunciation dictionary. This auxiliary dataset is added to the base training set.

2.4 Self-training for Text Normalization

Before text tokenization, TTS systems generally process the raw text by a text normalizaiton (TN) module to convert numbers and special symbols into their verbalization text, which relies on large amounts of hand-crafted rules; however, hand-crafted rules are constantly challenged by coverage on special symbols.

We explore LLMs for conducting the TN task, hence building a more unified end-to-end TTS system. Taking raw text as input, we utilize three ways to construct another auxiliary training set: 1) We pass raw text through an internal rule-based TN module, obtain text-normalized text, and synthesize audio by CosyVoice 2. 2) We prompt Qwen-Max [38] to conduct text normalization and then synthesize audio on the normalized text by CosyVoice 2. 3) We prompt Qwen-Max to conduct inverse text normalization on text in existing text-audio pairs and obtain the raw text (that is, unnormalized text).

The raw text and their corresponding audio are considered as a paired sample and directly added to the base traning set. We verify that the new system trained on the extended training set can synthesize raw text directly and exhibits better robustness and coverage on various special symbols.

2.5 Instructed Speech Generation

To enhance controllability and expressiveness of CosyVoice 3, compared to CosyVoice 2, we integrate more expressive speech data into the base training set. The duration of high-quality instruction-following data is expanded from 1,500 hours to 5,000 hours, covering a wider range of types including emotions, speed, voice tones, dialects, accents, and role-playing. The total number of types is increased to over 100, as illustrated in Table 1. Similar to CosyVoice 2, CosyVoice 3 also supports language instructions and fine-grained instructions. For natural language instructions, a natural language description and a special end token, "<|endofprompt|>", is prepended to the input text for speech synthesis. For fine-grained instructions, vocal bursts between text tokens and vocal feature tags are supported for control. For example, markers such as "[laughter]" and "[breath]" in the input text can be used to generate a noticeable laughter and breath, respectively. The tag "XXX" is used to indicate emphasis on specific words.

adventurous	ambitious	ancient	angry
artistic	authoritative	bold	brave
calm	charming	cheerful	clever
commanding	compassionate	confident	conflicted
contempt	courageous	creative	cunning
curious	dark	deceptive	dedicated
defiant	determined	disciplined	disgusted
empathetic	energetic	fearful	fearless
happy	heroic	hopeful	humble
imaginative	indifferent	insightful	intelligent
introspective	joyful	loyal	merciless
mysterious	noble	objective	optimistic
passionate	patient	proud	relaxed
relentless	responsible	sad	selfless
serious	shocked	stealthy	surprised
vengeful	vigilant	wise	fast
loud	slow	soft	adventurer
alchemist	architect	chef	craftsman
detective	doctor	girl	knight
leader	merchant	peppa	poet
robot	ruler	scholar	wanderer
warrior	witch	youth	anhui dialect
cantonese dialect	chongqing dialect	hebei dialect	shandong dialect
shanghai dialect	sichuan dialect	tianjin dialect	xi'an dialect
zhengzhou dialect	chinese english accent	indian english accent	russian english accent

Table 1: The 100 top-appeared speaking styles in pre-training data.

2.6 Capability Transfer in Speaker Fine-tuning

2.6.1 Turning a Monolingual Speaker into a Polyglot

A notable improvement in CosyVoice 3 over its predecessor is the extended language support. To enable a monolingual target speaker to speak multiple languages, we build an auxiliary training dataset, which contains studio-quality monolingual data from randomly-selected speakers covering all supported languages. The speaker ID and the language ID of every utterance are specified in a natural language instruction.

Examples
- 你是说话人小明。请讲法语。
- You are Speaker B. Please speak German.

Table 2: Examples of natural language instructions in the multilingual SFT dataset.

2.6.2 Transferring the Capability of Instructed Generation

Fine-tuning the pre-trained model with speaker-specific data can enhance the quality and expressiveness of generated output for individual speakers. We develop a training dataset that is partially labeled with speaker IDs. It includes high-quality data from the target speaker along with pre-training instruction-following dataset. In the natural language instruction prompt, we specify the speaker prompt and the style prompt. For example, a complete instruction prompt might be, "You are Speaker A. Please talk to me happily." However, some data entries might lack speaker IDs or style labels; in such cases, we leave those fields blank in the prompt. During the fine-tuning process, we also randomly mask the speaker prompt or the style prompt to enhance the model's transfer capability. This method ensures comprehensive instruction coverage across different speakers and helps prevent potential catastrophic forgetting in instructed generation with the pretrained models.

3 The Multilingual Data Pipeline

Compared to Chinese and English, it is more challenging to acquire large-scale high-quality TTS data in other languages. To tackle this challenge, we collect in-the-wild multilingual audio data mainly from Internet audiobooks, videos, and podcasts. Then, we implement a multilingual data processing pipeline to produce model training data with sufficient quality. The pipeline consists of six steps, as follows: 1. Speech detection and segmentation; 2. Noise reduction; 3. ASR transcription; 4. Punctuation adjustment; 5. Volume standardization; and 6. Filtering out data with abnormal audio-text length ratios.

Speech detection and segmentation. Raw data is sequentially processed by speaker diarization, voice activity detection (VAD), and audio event detection modules. As a result, speaker-level speech segments shorter than 30 seconds are obtained. Although we use in-house modules in this step, they can be replaced by open-source alternatives to the same effect.

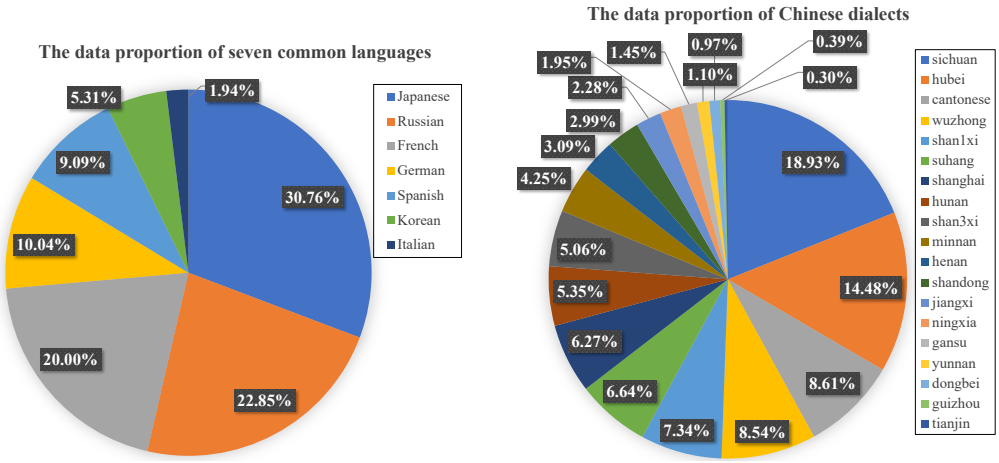
Noise reduction. We employ a MossFormer2 [39] model for noise reduction. Next, based on the energy levels of the leading and trailing frames of the utterances, ones starting or ending with incomplete words due to abnormal truncation are screened out; the remaining utterances, with leading and trailing silence trimmed, are retained for further processing.

ASR transcription. To obtain text transcriptions with adequate reliability, we first use Faster-Whisper Large-V3 [40] for language identification, then employ different open-source ASR models, namely, Faster-Whisper Large-V3, NVIDIA NeMo Canary-1B [41], Meta FAIR seamlessM4T-V2-large [42]), to transcribe the utterances. We then perform cross validation and select transcriptions with an average pair-wise WER lower than 15% among the ASR results from different systems.

Punctuation adjustment. Since the punctuations in the ASR-generated texts may fail to properly represent the actual pauses in the corresponding audio, we use Montreal Forced Aligner [43] to derive the durations between words and clauses or phrases, then add or remove punctuations by preset thresholds ($\geq$300 milliseconds to add a comma while $\leq$50 milliseconds to remove punctuations indicating pauses, i.e. commas, semicolons, colons, full stops, question marks and exclamation marks).

Volume standardization. A simple and straightforward normalization is applied for volume standardization:

$$\text{normalized\_wav} = \frac{\text{raw\_wav}}{\max(\text{raw\_wav})} \times 0.6 \tag{8}$$



(a) Minority languages (b) Chinese dialects

Figure 3: The data percentage of (a) seven minority languages and (b) 19 Chinese accents or dialects.

Filtering out utterances with abnormal audio-text length ratios. After all the above processing steps, speech tokens and text tokens are extracted for every generated utterance-text pair. Then, the utterance-level ratios of the lengths of the speech tokens and text tokens are calculated and sorted. We discard the utterances in the smallest 1% and utterances in the largest 5% in terms of the length ratios, to filter out possible abnormal cases with issues such as a short audio containing no human speech but corresponding to a long text transcription, or a long audio clip containing only a short human speech segment in the target language thus corresponding to a short text transcription.

4 Experimental Settings

4.1 Training Data for Speech Tokenizer

A 530,000-hour supervised multi-task dataset is used to train the speech tokenizer with normalized transcriptions as labels, including automatic speech recognition (ASR), language identification (LID), speech emotion recognition (SER), audio event detection (AED), and speaker analysis (SA). Details of the training data are listed in Table 3. The multilingual ASR training data consists of Chinese, English, Japanese, Korean, Russian, French, and German.

Language	**Duration (hours)**
Automatic Speech Recognition (ASR) - Multilingual	365K
Language Identification (LID)	85K
Speech Emotion Recognition (SER)	48K
Audio Event Detection (AED)	21K
Speaker analysis	11K

Table 3: Details of the training data for speech tokenizer.

4.2 Scaling up Dataset Size and Model Size for CosyVoice 3

In CosyVoice 3, we scale up the data volume from multiple aspects. For widely used Chinese and English data, we employ a combination of low-cost data production pipelines and self-training data construction to enhance diversity in domains, styles, text formats, and rare cases. Regarding domain diversity, we collect voice data from various fields such as e-commerce, navigation, finance, and education. In terms of style diversity, we add conversations, speeches, singing, and more. For text diversity, we construct different text formats for the same speech through text normalization (TN)

and inverse text normalization (ITN), enhancing the model's robustness to varied text formats. Additionally, we use self-training to strategically create numerous rare cases with an early version of CosyVoice 3 to improve synthesis stability. In terms of language coverage, we augment the Chinese and English dataset with seven common languages, including Japanese, Russian, French, German, Spanish, Korean, and Italian, with the data percentage shown in Figure 3a. Our previous work [27] shows that the supervised multi-task speech tokenizer could performs well on some unseen languages (that is, Spanish and Italian in the case of CosyVoice 3). In addition to standard common tongue pronunciations, we increase the coverage of Chinese accents and dialects, supporting 19 common accents or dialects, with the data percentage shown in Figure 3b. Through these data scaling efforts, the training data of CosyVoice 3 reaches one million hours, covering the majority of user cases in daily life and advancing towards in-the-wild zero-shot speech generation.

In addition to scaling dataset size, scaling up model size is crucial for current large-scale models. Therefore, we increase the model sizes of both the text-to-speech language model (LM) and the Conditional Flow Matching (CFM) model in CosyVoice 3. Specifically, the text-to-speech LM is increased from 0.5B to 1.5B parameters. For the CFM, we adopt the recent diffusion transformer (DiT) [25, 44] as the backbone, increasing the number of parameters from 100M to 300M. Preliminary experiments demonstrate the strong performance of the DiT architecture; hence, the complicated text encoder and the length regularization module are no longer needed and removed from CosyVoice 3. We solve the frame rate mismatch issue between speech tokens and Mel features by a simple interpolation operation.

4.3 Evaluation Settings for Zero-shot Capability

For evaluating CosyVoice 3's zero-shot speech generation capabilities, we focus on three key aspects: content consistency, speaker similarity, and audio quality. For content consistency, we measure the Character Error Rate (CER) or Word Error Rate (WER) of the ASR transcription against the given text, using Whisper-large V3 [45] for English ASR and Paraformer [46] for Chinese ASR. To assess speaker similarity, we extract speaker embeddings from the generated speech using the ERes2Net speaker verification model [47] and calculate the cosine similarity with the embedding of the reference speech. For audio quality, we score the generated speech using the DNSMOS network [48], the scores of which show high correlations with human auditory perception.

We conduct evaluations on two test sets. The first is the widely used SEED-TTS-Eval test set [26], where test cases are categorized into Mandarin, English, and hard Chinese subsets. To facilitate fair comparison with other models, we also use a WavLM-based speaker recognition model to calculate the speaker similarity [49]. Notably, recent advances in speech generation models have left little room for improvements, with models achieving quite similar scores; hence, we introduce a new multilingual benchmark **CV3-Eval** for evaluation, detailed in Section 4.4.

To perform a comprehensive comparison with CosyVoice 3, we employ 10 commonly-used speech generation models as the baselines, which achieve state-of-the-art (SOTA) or competitive performance in some aspects. Specifically, non-autoregressive (NAR) models include MaskGCT [15], E2 TTS [24], F5-TTS [25], and F5R-TTS [37], while autoregressive (AR) baselines are Seed-TTS [26], FireRedTTS [29], Qwen2.5-Omni [50], CosyVoice [27], CosyVoice 2 [30], and Spark TTS [17].

4.4 CV3-Eval: a Multilingual Benchmark

With the rapid development of speech generation models, existing evaluation benchmarks no longer meet the model assessment requirements, especially for zero-shot voice cloning. Firstly, most evaluation benchmarks such as Librispeech [51] are sampled from audio books, where the speaker's pronunciations are clean and standard. As a result, some systems can effortlessly synthesize high-quality audio that even beats the ground truth audio. However, source audio is often noisy in real-world application scenarios, presenting challenges that these benchmarks fail to address. Secondly, most benchmarks are designed for Chinese and English, while multilingual evaluation benchmarks are absent. Finally, traditional benchmarks only focus on the pronunciation accuracy, speaker similarity, and the MOS scores for audio quality. These evaluation metrics cannot accurately measure the comprehensive capability of a TTS system, including aspects such as emotion expression, rhythmic richness, voice controllability, and cross-lingual voice cloning.

To better evaluate CosyVoice 3, we establish a multilingual benchmark, CV3-Eval, including subsets for both objective and subjective evaluation.

Objective Evaluation. The objective evaluation subset is further split into three subsets, including multilingual voice cloning, cross-lingual voice cloning, and emotion cloning, as follows:

- **Multilingual Voice Cloning**: The multilingual voice cloning subset contains 9 languages with 500 samples for each language, including Chinese (zh), English (en), Japanese (ja), Korean (ko), German (de), France (fr), Russian (ru), Italian (it), and Spanish (es). The source audio and target text are sampled from CommonVoice [52] and FLUERS [53] datasets. To simulate real-world application scenarios, we do not filter out audio with noisy background or long silence, which poses challenges to the robustness of the TTS system. In addition, we construct two hard-case test sets for Chinese and English, where the target text includes rare words, tongue twisters, domain-specific terms, etc.
- **Cross-lingual Voice Cloning**. For the cross-lingual voice cloning subsets, the source audio and target text are from different languages, including zh, en, ja, and ko. This subset can evaluate the language transfer capability of the TTS system.
- **Emotion Cloning**. The audio prompts in the emotion cloning subset are sourced from EmoBox [54] and SeCap [55], including both Chinese and English samples. Due to the insufficient expressiveness of some emotion labels, we only include samples labeled as happy, sad, or angry, with 100 samples for each language. We further categorize these samples into text-related and text-unrelated parts, depending on whether the target text is semantically consistent with the target emotion. This helps us determine whether the synthetic emotional features are primarily derived from the text content or the prompt audio.

Subjective Evaluation. Besides the objective evaluation subset, we also prepare three subjective subsets for expressive voice cloning, expressive voice continuation, and Chinese accent voice cloning.

- **Expressive Voice Cloning**. To explore the model's capacity for generating expressive speech, the Expressive Voice Cloning benchmark is designed to include audio prompts with distinctive features, such as highly emotional intonation, whisper and shout, and extreme slow or fast speaking rate. Audio prompts are selected from different challenging application scenarios such as news, podcasts, TV drama, academic reports, poetry recitation, etc. Voices of some public figures are also sampled for evaluation.
- **Expressive Voice Continuation**. Due to the high variability in human perception, achieving a fair subjective evaluation of expressive voice cloning is challenging. To mitigate this issue, we design a voice continuation task. Specifically, we select 120 audio samples with different emotions, rhythms, speeds, and volumes from the website and cut the first 3 seconds of the audio clip as prompt speech. Therefore, we can evaluate the synthesized remaining speech based on its similarity with the ground truth speech.
- **Chinese Accent Voice Cloning**. Since there is currently no reliable objective method to evaluate the authenticity of accents, we construct a subjective evaluation dataset for Chinese dialects. The dataset includes 18 different Chinese dialects, such as Cantonese, Dongbei, Minnan, Shanghai dialects, etc. All prompt speech samples are sourced from in-house industrial data.

5 Experimental Results

5.1 Objective TTS Results on SEED-TTS-Eval

Table 4 presents the TTS performance of CosyVoice 3 and several recent models across the SEED test sets, which include the Chinese *test-zh*, English *test-en*, and the challenging *test-hard* sets. The evaluation focuses on content consistency (WER/CER) and speaker similarity (SS).

For **content consistency**, CosyVoice 3 achieves significant improvements over CosyVoice 2, with relative gains of 44% on *test-zh* and 51% on *test-en*. In the *test-hard* set, CosyVoice 3 reduces the CER from 6.83% to 5.09% (26% relative improvement). Compared to other baselines, CosyVoice 3 consistently excels across all metrics. Notably, CosyVoice 3-1.5B$_{RL}$ records the lowest CER

Model	*test-zh*		*test-en*		*test-hard*	
	CER (%) ↓	SS ↑	WER (%) ↓	SS ↑	CER (%) ↓	SS ↑
Human	1.26	0.755 (0.775)	2.14	0.734 (0.742)	-	-
Vocoder Resyn.	1.27	0.720	2.17	0.700	-	-
Non-autoregressive Models						
MaskGCT [15]	2.27	0.774 (0.752)	2.62	0.714 (0.730)	10.27	0.748 (0.720)
E2 TTS (32 NFE) [24]	1.97	0.730	2.19	0.710	-	-
F5-TTS (32 NFE) [25]	1.56	0.741 (0.794)	1.83	0.647 (0.742)	8.67	0.713 (0.762)
F5R-TTS [37]	1.37	0.754	-	-	8.79	0.718
Autoregressive Models						
Seed-TTS [26]	1.12	**0.796**	2.25	**0.762**	7.59	**0.776**
FireRedTTS [29]	1.51	0.635 (0.653)	3.82	0.460 (0.526)	17.45	0.621 (0.639)
Qwen2.5-Omni-7B [50]	1.70	0.752	2.72	0.632	7.97	0.747
Qwen2.5-Omni-7B$_{RL}$ [50]	1.42	0.754	2.33	0.641	6.54	0.752
CosyVoice [27]	3.63	0.723 (0.775)	4.29	0.609 (0.699)	11.75	0.709 (0.755)
CosyVoice 2 [30]	1.45	0.748 (0.806)	2.57	0.652 (0.736)	6.83	0.724 (0.776)
Spark TTS [17]	1.20	0.672	1.98	0.584	-	-
CosyVoice 3-0.5B	1.16	0.780 (**0.840**)	2.02	0.718 (**0.790**)	6.08	0.758 (0.815)
CosyVoice 3-0.5B$_{RL}$	0.75	0.774 (0.836)	1.76	0.695 (0.783)	**5.09**	0.750 (0.809)
CosyVoice 3-1.5B	1.12	0.781 (0.837)	2.21	0.720 (0.789)	5.83	0.758 (**0.816**)
CosyVoice 3-1.5B$_{RL}$	**0.71**	0.775 (0.836)	**1.45**	0.695 (0.784)	5.66	0.750 (0.810)

Table 4: Zero-shot TTS performance comparison between CosyVoice 3 and the baselines on the SEED test sets in terms of content consistency (WER/CER) and speaker similarity (SS). For speaker similarity, the results outside parentheses are measured by WavLM-based models while the results inside parentheses are measured by ERes2Net. While the **boldface** denotes the best result, the underline denotes the second best.

of 0.71% in *test-zh* and the lowest WER of 1.45% in *test-en*, showcasing its superior synthesis accuracy. In the challenging *test-hard* scenario, CosyVoice 3-0.5B$_{RL}$ achieves the lowest CER of 5.09%, while the 1.5B variant follows closely with 5.66%. The larger model's underperformance compared to the smaller one is due to the limited dataset available for pretraining and post-training, particularly in challenging scenarios. We plan to expand our dataset to tens of millions of hours to support more effective training of larger models in the future.

Regarding **speaker similarity**, CosyVoice 3 demonstrates a strong ability to replicate speaker characteristics accurately. It outperforms CosyVoice 2 and other baselines, except Seed-TTS, as shown through both WavLM-based and ERes2Net measurements. The similarity gap between CosyVoice 3 and Seed-TTS is primarily due to differences in speaker diversity and pretraining data volume. Enhancing speaker similarity in CosyVoice 3 can be achieved by scaling up pretraining data, a direction we intend to pursue in future work. Additionally, Table 4 shows that RL post-training contributes to 12% to 35% relative improvements in content consistency, enhancing robustness and adaptability in multilingual and complex synthesis tasks. With RL post-training, CosyVoice 3 establishes a new state of the art in TTS performance, demonstrating substantial advancements over previous models.

5.2 Objective Evaluation on Multilingual Benchmark CV3-Eval

5.2.1 Results of Multilingual Voice Cloning

We evaluate CosyVoice 3 against competitive open-source TTS systems, including F5-TTS, Spark-TTS, and GPT-SoVits[1], using the Multilingual Voice Cloning subset of CV3-Eval benchmark. Table 5 provides CERs for Chinese, Japanese, and Korean, and WERs for English, German, Spanish, French, Italian, and Russian. The Multi-lingual Voice Cloning subset proved to be significantly challenging, as CosyVoice 3 is the only system capable of covering all languages in this subset. For most languages, the performance difference between CosyVoice3-0.5B and CosyVoice3-1.5B is minimal. Furthermore, as shown in Table 6, generating rare words, tongue twisters, and domain-specific terms remains difficult for CosyVoice 3, highlighting areas for future improvement.

[1]https://github.com/RVC-Boss/GPT-SoVITS

Model	zh	en	ja	ko	de	es	fr	it	ru
F5-TTS	5.47	8.90	–	–	–	–	–	–	–
Spark-TTS	5.15	11.0	–	–	–	–	–	–	–
GPT-SoVits	7.34	12.5	–	–	–	–	–	–	–
CosyVoice2	4.08	6.32	9.13	19.7	–	–	–	–	–
+ DiffRO	3.00	4.72	6.36	5.14	–	–	–	–	–
CosyVoice3-0.5B	3.89	5.24	10.4	12.8	7.41	4.25	12.9	6.68	6.77
+ DiffRO	2.89	3.68	5.15	4.02	4.51	2.99	8.56	2.94	3.79
CosyVoice3-1.5B	3.91	4.99	7.57	5.69	6.43	4.47	11.8	10.5	6.64
+ DiffRO	3.01	3.71	5.27	4.01	3.93	3.26	8.09	2.72	4.11

Table 5: CER(%) and WER(%) on the CV3-Eval Multilingual Voice Cloning subset. – means the language is unsupported.

Model	hard-zh			hard-en		
	WER	SS	DNSMOS	WER	SS	DNSMOS
CosyVoice2	12.58	72.6	3.81	11.96	66.7	3.95
+ DiffRO	10.66	71.7	3.81	10.25	62.4	3.97
CosyVoice3-0.5B	14.15	78.6	3.75	9.04	75.9	3.92
+ DiffRO	8.26	77.8	3.80	7.60	73.9	3.95
CosyVoice3-1.5B	9.77	78.5	3.79	10.55	76.1	3.95
+ DiffRO	9.06	78.2	3.81	7.56	74.6	3.95

Table 6: WER(%), Speaker Similarity (SS), and MOS scores on the **hard samples** in the CV3-Eval Multilingual Voice Cloning subset.

5.2.2 Results of Cross-lingual Voice Cloning

Table 7 illustrates the significant improvements CosyVoice 3 offers over CosyVoice 2 in cross-lingual voice cloning. Notably, CosyVoice 2 struggles with transferring voice from Japanese to Chinese due to the character overlap of two languages, a problem resolved in CosyVoice 3 by converting all Japanese characters into kana. Additionally, scaling the model size proves beneficial: CosyVoice3-1.5B exhibits better WERs across all conditions compared to CosyVoice3-0.5B, while maintaining similar speaker similarity. This indicates that larger models can enhance performance on challenging tasks due to increased capacity.

Since most open-source TTS systems only support Chinese and English, we further evaluate CosyVoice 3 against baselines for the zh2en and en2zh cross-lingual voice cloning tasks, as shown in Table 8. Compared to CosyVoice 3, F5-TTS and Spark-TTS show inferior performance on WER, with Spark-TTS also lagging significantly in SS compared to F5-TTS and CosyVoice 3. Regarding MOS scores, CosyVoice 3 demonstrates better results for en2zh and comparable results for zh2en. Overall, CosyVoice3-1.5B remains the leading model for zh2en and en2zh cross-lingual transfer tasks.

5.2.3 Results of Emotional Voice Cloning

In the CV3-Eval Emotional Voice Cloning subset, we employ the emo2vec-large-plus model[2] as a classifier to assess the emotion expression capabilities of TTS systems. The results, displayed in Table 9, reveal that most TTS systems perform well on text-related subsets, with CosyVoice 3 achieving the highest performance. Each system excels in expressing specific emotions, with "happy" being the easiest emotion to convey across all models. However, in text-unrelated tasks, emotion accuracy drops significantly, particularly for "sad" and "angry" emotions. This indicates that TTS systems primarily infer the emotional tone of output audio from text sentiment. This observation provides valuable insights into the less satisfactory performances and highlights areas for future improvement.

[2]https://www.modelscope.cn/models/iic/emotion2vec_plus_large/summary

Model	to-zh			to-en			to-ja			to-ko		
	en	ja	ko	zh	ja	ko	zh	en	ko	zh	en	ja
CosyVoice2	13.5	48.1	7.70	6.47	17.1	11.2	13.1	14.9	5.86	24.8	21.9	21.5
CosyVoice3-0.5B	8.48	6.86	5.24	4.99	6.83	5.86	18.3	16.8	4.99	41.0	20.4	12.8
+ DiffRO	5.16	3.22	1.03	3.40	4.41	4.78	7.91	7.25	3.29	16.9	11.6	8.2
CosyVoice3-1.5B	8.01	6.78	3.30	4.32	5.39	5.94	13.7	13.4	4.19	31.6	14.0	10.5
+ DiffRO	5.09	3.05	1.06	2.98	4.20	4.19	7.08	6.80	3.93	14.4	5.87	7.92

Table 7: WER(%) results on the CV3-Eval Cross-lingual Voice Cloning subset.

Model	en2zh			zh2en		
	WER	SS	MOS	WER	SS	MOS
F5-TTS	11.6	64.2	3.77	5.57	64.7	3.77
Spark-TTS	12.4	48.4	3.65	7.36	56.7	3.61
CosyVoice2	13.5	63.3	3.87	6.47	64.3	3.75
CosyVoice3-0.5B	8.48	67.4	3.82	4.99	67.8	3.75
CosyVoice3-1.5B	8.01	66.9	3.83	4.32	66.4	3.77

Table 8: WER(%), Speaker Similarity (SS), and MOS scores from CosyVoice 3 and the baselines on the zh2en and en2zh voice cloning tasks.

Model	Text-Related			Text-Unrelated		
	happy	sad	angry	happy	sad	angry
F5-TTS	0.92	0.52	0.72	0.80	0.28	0.64
Sparks-TTS	0.80	0.56	0.50	0.50	0.60	0.36
GPT-SoVits	0.88	0.54	0.50	0.48	0.40	0.30
CosyVoice2	0.84	0.72	0.58	0.56	0.44	0.38
CosyVoice3-0.5B	0.92	0.70	0.72	0.64	0.42	0.58
CosyVoice3-1.5B	0.86	0.64	0.72	0.64	0.44	0.48
+ DiffRO-EMO	0.98	0.68	0.84	0.98	0.50	0.68

Table 9: Emotion Accuracy on the Text-Related and Text-Unrelated subsets of the CV3-Eval Emotional Voice Cloning subset.

5.2.4 Subjective Evaluation Results

In addition to objective metrics, we performe a subjective evaluation using Mean Opinion Scores (MOS). The test samples comprise 200 Chinese and English sentences, each assessed by 10 native speakers (5 male and 5 female). Scores ranged from 1 to 5, in 0.5-point increments. Figure 4 shows the MOS scores for the CosyVoice 2, CosyVoice 3-0.5B, and CosyVoice 3-1.5B models across both languages, along with their average scores.

For Chinese, all three models perform similarly but still lag behind human speech. In English, CosyVoice 2 scores lower than human benchmarks, CosyVoice 3-0.5B matches human scores, and CosyVoice 3-1.5B scores notably higher. Overall, CosyVoice 3-1.5B outperforms CosyVoice 3-0.5B, with both surpassing CosyVoice 2, illustrating the advantages of data and model scaling.

Despite some differences from human speech in Chinese, CosyVoice 3 models still score above 4.45. This gap is primarily due to a few low-scoring cases in the synthetic output compared to real speech, indicating the need for improved synthesis stability in future work.

5.3 Ablation of Speech Tokenizer

5.3.1 Up-streaming Recognition Tasks

Our supervised multi-task learning-based speech tokenizer exhibits strong performance across various speech and sound tasks. Specifically, as shown in Table 10, the FSQ-MinMo-based speech

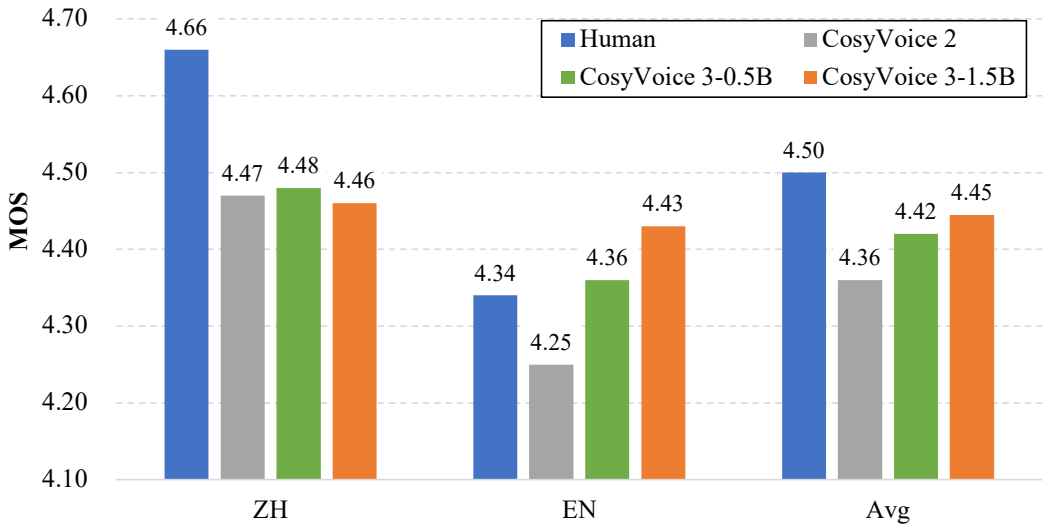


Figure 4: The Mean Opinion Scores (MOS) of zero-shot cloning models on Chinese, English and their average.

Method	C.V. EN	C.V. CN	C.V. JA	C.V. KO	Fluers EN	Fluers CN
SenseVoice	7.70	8.67	-	-	4.57	6.98
MinMo	7.36	8.56	-	-	4.43	6.71
VQ-SenseVoice	18.26	11.56	-	-	7.65	5.03
FSQ-SenseVoice	10.67	7.29	-	-	6.58	4.43
FSQ-MinMo	11.36	9.21	13.90	9.78	4.46	3.35

Table 10: Comparison between VQ and FSQ inside the Sensevoice-large and MinMo encoders in terms of ASR WERs and CERs (%) across language-specific subsets of the CommonVoice (**C.V.**) and the Fluers benchmarks. **FSQ-MinMo** is the tokenizer used in CosyVoice 3.

tokenizer in CosyVoice 3 effectively maintains multilingual ASR capabilities. By focusing exclusively on speech-related tasks in FSQ-MinMo and excluding others from the training set, we achieve superior recognition performance compared to MinMo on the Fluers CN test set. Additionally, Table 11 illustrates that the FSQ-MinMo model performs comparably to the MinMo model on the AIR-Bench benchmark, which includes tasks such as LID, Gender, Age, Emotion, Vocal Sound, and Sound Question classification.

5.3.2 Down-streaming TTS Tasks

Beyond upstream recognition tasks, we also evaluate the tokenizer in downstream TTS tasks to directly assess synthesis performance by replacing the CosyVoice 3 tokens with others and maintaining the model architectures of LM and CFM unchanged. Table 12 presents the results of various models trained on datasets of two different scales: 3,000 hours and 170,000 hours. Alongside our

Method	Language ID	Gender	Age	Emotion	Vocal Sound	Sound Question
MinMo	99.2	84.8	70.1	62.4	90.7	59.1
FSQ-MinMo	99.2	72.8	41.8	68.4	61.3	57.7

Table 11: Performance comparison between MinMo and FSQ-MinMo models in terms of Accuracy on the AIR-Bench benchmark, including Language ID, Gender, Age, Emotion, Vocal Sound, and Sound Question classification tasks.

Model	*test-zh*		*test-en*		*test-hard*	
	CER (%) ↓	**SS** ↑	**WER (%)** ↓	**SS** ↑	**CER (%)** ↓	**SS** ↑
3000-hour Dataset						
SoundStream(1^{st} **VQ)** [56]	14.19	0.457	25.34	0.301	27.05	0.455
HuBERT [57]	18.68	0.716	6.50	0.609	33.83	0.699
W2v-BERT 2.0 [58]	2.62	0.381	6.72	0.261	23.89	0.374
CosyVoice 2.0 [30]	1.92	0.668	7.21	0.535	15.99	0.645
CosyVoice 3.0-0.5B	1.68	0.710	6.60	0.614	27.60	0.679
170,000-hour Dataset						
CosyVoice 2.0 [30]	1.45	0.806	2.57	0.736	**6.83**	0.776
CosyVoice 3.0-0.5B	**1.27**	**0.815**	**2.46**	**0.747**	6.96	**0.787**

Table 12: Performance comparison of down-streaming zero-shot TTS modeling using different tokenizers on the SEED test sets in terms of content consistency (WER/CER) and speaker similarity (SS). While the **boldface** denotes the best result.

supervised semantic tokenizers, CosyVoice 2.0[3] and CosyVoice 3.0, we evaluate the self-supervised tokenizers, HuBERT[4] and W2v-BERT 2.0[5] which are widely used in other TTS models. In addition, we also involve the unsupervised tokenizer, SoundStream[6], which quantizes the acoustic waveform into groups of discrete tokens by the residual vector quantization based variational autoencoder (RVQ-VAE). Since other tokens have only single codebook, only the first VQ group of SoundStream is employed for comparison.

On the 3,000-hour dataset, supervised semantic tokenizers exhibit similar speaker similarity to HuBERT while significantly outperforming W2v-BERT 2.0. This is because both HuBERT and supervised semantic tokenizers focus on semantic information, minimizing acoustic interference, whereas W2v-BERT 2.0 retains all contextual information, both semantic and acoustic, due to its training approach. This allows the conditional flow matching model to better emphasize the acoustic characteristics of reference speech while disregarding acoustic interference in speech tokens. Regarding content consistency, supervised tokenizers achieve the lowest CER on the test-zh set and deliver comparable performance on the test-en and test-hard sets. The notably high CER of HuBERT on the test-zh set underscores its language-specific limitations. As expected, acoustic tokens of SoundStream achieve notably high error rates on all evaluated test sets, indicating poor content consistency to the synthesis text. This is because these acoustic tokens neither attempt to model contextual information like self-supervised tokens nor align with the text like supervised tokens, resulting in a lack of sufficient semantic information.

Increasing the training data volume from 3,000 to 170,000 hours leads to significant improvements in content consistency and speaker similarity, especially for English and challenging scenarios, with relative WER/CER improvements ranging from 63% to 75%. As indicated in Table 4, further scaling the dataset to one million hours enhances performance, but the rate of improvement begins to plateau. This suggests that our multi-task supervised tokenizer is scalable and benefits from larger datasets up to a point of diminishing returns.

5.4 Ablation of Reinforcement Learning

Our experiments demonstrate that DiffRO significantly enhances the performance of TTS systems, including both CosyVoice 2 and CosyVoice 3. As indicated in Tables 4, 5, and 7, DiffRO achieves relative improvements ranging from 20% to 50% in terms of WER. The enhancements are particularly notable in low-resource languages and cross-lingual scenarios, with over 50% relative WER improvement in half of the conditions; notably, CosyVoice 3-0.5B shows a 68.7% relative improve-

[3]https://github.com/FunAudioLLM/CosyVoice

[4]https://github.com/facebookresearch/fairseq/tree/main/examples/hubert

[5]https://huggingface.co/amphion/MaskGCT/tree/main/semantic_codec

[6]https://huggingface.co/amphion/MaskGCT/tree/main/acoustic_codec

Method	zh			en		
	Errors	Corrections	Rate(%)	Errors	Corrections	Rate(%)
RepAll + MixPhn	13	9	69.2	11	8	72.7
RepMono + MixPhn	15	15	100	9	9	100
RepMono + CatPhn	15	13	86.7	8	8	100

Table 13: Corretion Rate of procunciation inpainting.

ment in Korean. Furthermore, RL training reduces the performance gap between the 0.5B and 1.5B models.

Regarding speaker similarity, RL slightly reduces speaker similarity across most datasets, although the change is minimal. This suggests the persistence of the "hacking" problem in DiffRO, where the model focuses more on target rewards, potentially neglecting other metrics. Introducing a speaker similarity module as a reward task may mitigate this issue but could increase WER. Additionally, incorporating the SER task as a reward model aims to enhance CosyVoice 3's emotion expression. Table 9 shows that DiffRO-EMO allows CosyVoice3-1.5B to achieve top emotion accuracy across most emotions in both text-related and text-unrelated tasks. However, this improvement in emotion expression can adversely affect pronunciation, highlighting the challenge of balancing rewards in DiffRO, which will be addressed in future work.

Moreover, as seen in Table 6, DiffRO's improvements in WER, SS, and DNSMOS on hard sample test sets are less pronounced than those on overall test sets. This is likely due to the hard samples comprising rare words, tongue twisters, and repeated words, which present significant challenges for reward models.

5.5 Pronunciation Inpainting

We construct an evaluation set to compare different pronunciation inpainting methods, focusing on challenging cases of Chinese polyphonic characters and English polyphonic words. Correction rate serves as the metric for assessing inpainting capability. As shown in Table 10, the best method achieves a 100% correction rate.

The "RepAll" approach involves considering all Chinese characters and English words as potential replacements, using internal G2P models for phoneme prediction during training data augmentation. While this method offers extensive coverage of character-phoneme combinations, it introduces mismatches due to G2P predictions. Conversely, "RepMono" only replaces monophonic characters or words, ensuring accuracy in the training set.

The key distinction between "CatPhn" and "MixPhn" lies in whether the Chinese character is retained and concatenated with its phoneme representation or replaced solely by the phoneme. "CatPhn" preserves semantic completeness but requires the model to prioritize phoneme representation over the character, which is exacerbated when only monophonic characters are considered. To mitigate this, we introduce some noisy data, such as replacing a character with a different-sounding one while retaining the correct phoneme representation. However, achieving a competitive correction rate with "MixPhn" remains challenging.

5.6 Instructed Generation

We evaluate the effectiveness of instructed generation capabilities using the Expresso [59] dataset alongside an internal expressive dataset. The Expresso dataset is a multi-speaker expressive speech collection featuring eight distinct speaking styles, evaluated on a subset of 3,000 samples. Our internal dataset includes 3,600 samples, matching the domains of the instruction-following training dataset and encompassing over 50 different emotions, speeds, dialects, accents, and role-playing speaking styles.

The evaluation results are presented in Table 14. CosyVoice 3 shows notable improvements in style similarity, with an approximate 11% relative increase over its predecessor. In terms of content consistency, CosyVoice 3 demonstrates a higher WER on the Expresso test set but a lower WER on our internal test set. This discrepancy is largely due to the ASR model's bias towards standard

pronunciations over emotional ones, as indicated by the higher WER for ground-truth utterances compared to CosyVoice 2. Objectively evaluating content consistency in emotional speech remains a challenging issue.

While we have explored various styles through instructed generation, singing has not been included and will be addressed in future work. Currently, CosyVoice 3's instructed generation focuses on emotion, speech, and style, primarily related to the language model (LM). Timbre, more closely associated with conditional flow matching (CFM), has not yet been considered. Editing timbre using natural language or other modalities is a promising and underexplored area [60].

Model	Expresso			Internal Dataset		
	WER	SIM	MOS	WER	SIM	MOS
GroundTruth	10.0	100	3.65	8.98	100	3.47
CosyVoice 2	9.42	60.98	3.54	7.75	72.99	3.53
CosyVoice 3-0.5B	13.72	67.82	3.56	7.30	80.45	3.51
CosyVoice 3-1.5B	13.43	68.25	3.56	7.31	81.06	3.51

Table 14: Comparison of WER (%), Style Similarity (SIM), and MOS scores across different models for instructed TTS tasks.

5.7 Results on Speaker Fine-tuned Models

To ensure timbre consistency in the SFT models, we utilize an unsupervised clustering method to identify the timbre centers for each speaker. These clustering centers are then used as speaker embeddings in the conditional flow matching model. As illustrated in Figure 5, increasing the volume and diversity of training data, along with upgrading speech tokens, leads to a reduction in error rates for the fine-tuned models, particularly noticeable in the test-en and test-hard sets. This indicates that improving the base model can also benefit the speaker fine-tuned models.

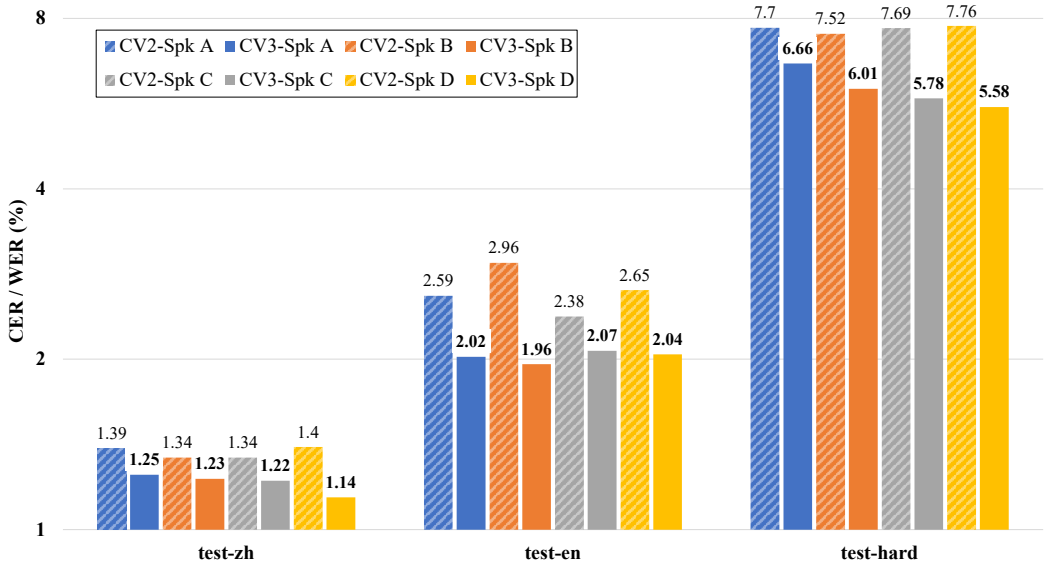


Figure 5: Content consistency results of CosyVoice 3 and CosyVoice 2 SFT models under the SEED-TTS-Eval settings. Word error rate (WER) is used for test-en set, while character error rate (CER) is used for the others.

5.8 Results on Turning a Monolingual Speaker into a Polyglot

In our experiments, we aim to transform a monolingual speaker into a polyglot using the training process described in Section 2.6.1. As shown in Figure 6, the CERs/WERs for languages such as

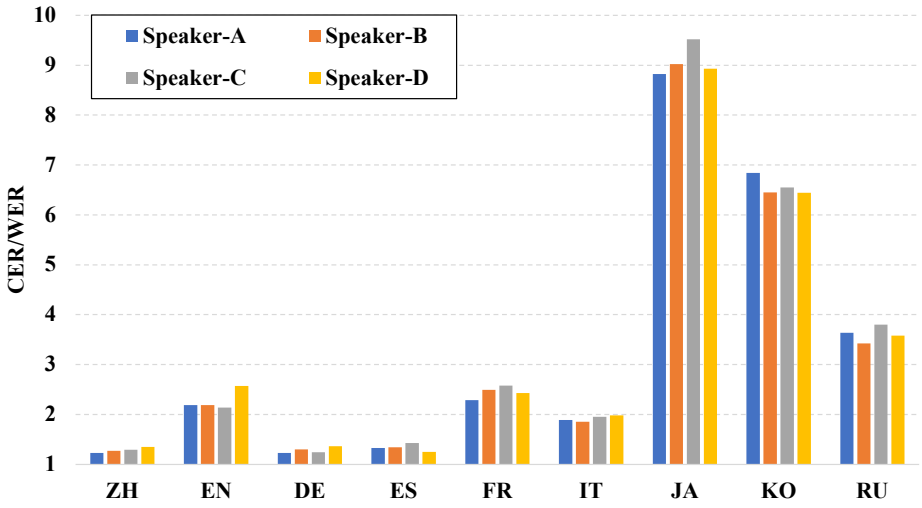


Figure 6: Content consistency results of CosyVoice 3 for turning a monolingual speaker into a polyglot. Character error rate (CER) is used for ZH, KO, and JA, while word error rate (WER) is used for the others.

Chinese, English, German, Spanish, French, Italian, and Russian are all below 4%, demonstrating the effectiveness of our continual training approach.

However, Japanese poses a challenge with a higher character error rate of 9%, which can be attributed to two main factors: the conversion of kanji into kana before speech synthesis introduces additional errors, and the multiple pronunciations of Japanese characters add complexity. For Korean, the CER is approximately 6%, mainly due to the limited volume and quality of available data. We will extend the Korean data in the future work.

6 Conclusion

To conclude, this report introduces CosyVoice 3, an advanced zero-shot speech synthesis model tailored for in-the-wild applications. By scaling up both data and model parameters, CosyVoice 3 overcomes previous limitations in language coverage and synthesis quality, delivering superior content consistency, speaker similarity, and prosody naturalness. Our innovations, including a novel speech tokenizer and post-training strategies, enhance the model's ability to capture intricate paralinguistic details. Achieving state-of-the-art results across multiple benchmarks, CosyVoice 3 represents a significant step forward in speech synthesis, paving the way for more versatile and high-quality voice generation in diverse real-world scenarios.

7 Limitations

CosyVoice 3 has several limitations that need to be addressed in future work. CosyVoice 3 cannot control acoustic characteristics, such as timbre, through textual instructions, which could be an interesting and valuable area of exploration for role-playing applications. Furthermore, CosyVoice 3 does not perform quite well for generating singing voice. This could be improved by adding singing data into the training stages of both tokenizer and LM model.

References

[1] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In *INTERSPEECH*, pages 4006–4010. ISCA, 2017.

[2] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *ICASSP*, pages 4779–4783. IEEE, 2018.

[3] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. *CoRR*, abs/1710.07654, 2017.

[4] Wei Ping, Kainan Peng, and Jitong Chen. Clarinet: Parallel wave generation in end-to-end text-to-speech. In *ICLR (Poster)*. OpenReview.net, 2019.

[5] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In *NeurIPS*, pages 3165–3174, 2019.

[6] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *AAAI*, pages 6706–6713. AAAI Press, 2019.

[7] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *ICLR*. OpenReview.net, 2021.

[8] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111, 2023.

[9] Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Trans. Assoc. Comput. Linguistics*, 11:1703–1718, 2023.

[10] Yakun Song, Zhuo Chen, Xiaofei Wang, Ziyang Ma, and Xie Chen. ELLA-V: stable neural codec language modeling with alignment-guided sequence reordering. *CoRR*, abs/2401.07333, 2024.

[11] Chenpeng Du, Yiwei Guo, Hankun Wang, Yifan Yang, Zhikang Niu, Shuai Wang, Hui Zhang, Xie Chen, and Kai Yu. VALL-T: decoder-only generative transducer for robust and decoding-controllable text-to-speech. *CoRR*, abs/2401.14321, 2024.

[12] Detai Xin, Xu Tan, Kai Shen, Zeqian Ju, Dongchao Yang, Yuancheng Wang, Shinnosuke Takamichi, Hiroshi Saruwatari, Shujie Liu, Jinyu Li, and Sheng Zhao. RALL-E: robust codec language modeling with chain-of-thought prompting for text-to-speech synthesis. *CoRR*, abs/2404.03204, 2024.

[13] Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *CoRR*, abs/2406.05370, 2024.

[14] Bing Han, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Yanming Qian, Yanqing Liu, Sheng Zhao, Jinyu Li, and Furu Wei. VALL-E R: robust and efficient zero-shot text-to-speech synthesis via monotonic alignment. *CoRR*, abs/2406.07855, 2024.

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

[16] Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, Helen Meng, and Furu Wei. Autoregressive speech synthesis without vector quantization. *CoRR*, abs/2407.08551, 2024.

[17] 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.

[18] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In *NeurIPS*, 2023.

[19] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun

Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *ICML*. OpenReview.net, 2024.

[20] Yiwei Guo, Chenpeng Du, Ziyang Ma, Xie Chen, and Kai Yu. Voiceflow: Efficient text-to-speech with rectified flow matching. In *ICASSP*, pages 11121–11125. IEEE, 2024.

[21] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast TTS architecture with conditional flow matching. In *ICASSP*, pages 11341–11345. IEEE, 2024.

[22] Yuan Gao, Nobuyuki Morioka, Yu Zhang, and Nanxin Chen. E3 TTS: easy end-to-end diffusion-based text to speech. In *ASRU*, pages 1–8. IEEE, 2023.

[23] Keon Lee, Dong Won Kim, Jaehyeon Kim, and Jaewoong Cho. Ditto-tts: Efficient and scalable zero-shot text-to-speech with diffusion transformer. *CoRR*, abs/2406.11427, 2024.

[24] Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, Yanqing Liu, Sheng Zhao, and Naoyuki Kanda. E2 TTS: embarrassingly easy fully non-autoregressive zero-shot TTS. *CoRR*, abs/2406.18009, 2024.

[25] 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. *CoRR*, abs/2410.06885, 2024.

[26] 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, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao 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. *CoRR*, abs/2406.02430, 2024.

[27] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *CoRR*, abs/2407.05407, 2024.

[28] Xingchen Song, Mengtao Xing, Changwei Ma, Shengqiang Li, Di Wu, Binbin Zhang, Fuping Pan, Dinghao Zhou, Yuekai Zhang, Shun Lei, et al. Touchtts: An embarrassingly simple tts framework that everyone can touch. *arXiv preprint arXiv:2412.08237*, 2024.

[29] Haohan Guo, Kun Liu, Feiyu Shen, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kaituo Xu. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *CoRR*, abs/2409.03283, 2024.

[30] 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.

[31] Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, et al. Ditar: Diffusion transformer autoregressive modeling for speech generation. *arXiv preprint arXiv:2502.03930*, 2025.

[32] Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang. Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system. *arXiv preprint arXiv:2502.05512*, 2025.

[33] Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, et al. Minmo: A multimodal large language model for seamless voice interaction. *arXiv preprint arXiv:2501.06282*, 2025.

[34] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *ICLR*. OpenReview.net, 2024.

[35] Tongyi Speech Team. Funaudiollm: Voice understanding and generation foundation models for natural interaction
between humans and llms. *arxiv*, 2024.

[36] Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[37] Xiaohui Sun, Ruitong Xiao, Jianye Mo, Bowen Wu, Qun Yu, and Baoxun Wang. F5r-tts: Improving flow matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*, 2025.

[38] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[39] Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jiaqi Yip, Dianwen Ng, and Bin Ma. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation, 2024.

[40] Systran. Faster whisper large v3, 2023.

[41] NVIDIA. Canary 1b, 2024.

[42] AI at Meta. Seamlessm4t v2, 2023.

[43] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech 2017*, pages 498–502, 2017.

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[45] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023.

[46] Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.

[47] Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. An enhanced res2net with local and global feature fusion for speaker verification. *arXiv preprint arXiv:2305.12838*, 2023.

[48] Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. Dnsmos P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP*, pages 886–890. IEEE, 2022.

[49] Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151. IEEE, 2022.

[50] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[51] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[52] 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. *arXiv preprint arXiv:1912.06670*, 2019.

[53] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

[54] Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. In *Proc. INTERSPEECH*, 2024.

[55] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shixiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. Secap: Speech emotion captioning with large language model. *arXiv preprint arXiv:2312.10381*, 2023.

[56] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:495–507, 2022.

[57] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[58] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE, 2021.

[59] Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*, 2023.

[60] Zhengyan Sheng, Zhihao Du, Heng Lu, Shiliang Zhang, and Zhen-Hua Ling. Unispeaker: A unified approach for multimodality-driven speaker generation. *arXiv preprint arXiv:2501.06394*, 2025.