MiniMax-Speech

Intrinsic Zero-Shot Text-to-Speech with a Learnable Speaker Encoder

MiniMax[1]

We introduce MiniMax-Speech, an autoregressive Transformer-based Text-to-Speech (TTS) model that generates high-quality speech. A key innovation is our learnable speaker encoder, which extracts timbre features from a reference audio without requiring its transcription. This enables MiniMax-Speech to produce highly expressive speech with timbre consistent with the reference in a zero-shot manner, while also supporting one-shot voice cloning with exceptionally high similarity to the reference voice. In addition, the overall quality of the synthesized audio is enhanced through the proposed Flow-VAE. Our model supports 32 languages and demonstrates excellent performance across multiple objective and subjective evaluations metrics. Notably, it achieves state-of-the-art (SOTA) results on objective voice cloning metrics (Word Error Rate and Speaker Similarity) and has secured the top position on the public TTS Arena leaderboard. Another key strength of MiniMax-Speech, granted by the robust and disentangled representations from the speaker encoder, is its extensibility without modifying the base model, enabling various applications such as: arbitrary voice emotion control via LoRA; text to voice (T2V) by synthesizing timbre features directly from text description; and professional voice cloning (PVC) by fine-tuning timbre features with additional data. We encourage readers to visit https://minimax-ai.github.io/tts_tech_report for more examples.

1. Introduction

The advent of codec-based models has catalyzed significant advancements in TTS. These models, when trained on large-scale datasets, demonstrate the remarkable capability to generate high-quality speech from only a few seconds of reference audio. This proficiency has fostered their widespread adoption across diverse applications, including conversational AI, audio content creation for blogs, interactive voice assistants, and immersive e-book narration.

Two primary modeling methodologies are prevalent in large TTS models: autoregressive (AR) language models and non-autoregressive (NAR) diffusion models. NAR diffusion models have gained attention for their rapid inference capabilities. However, such NAR models often employ duration modeling techniques. One approach is explicit phoneme-duration alignment (1–4), which can constrain naturalness and diversity. An alternative, global duration modeling (5–11), involves the model learning implicit alignments; this, however, can increase modeling complexity and reduce robustness in challenging cases. Conversely, AR models are renowned for their potent modeling capacity, an inherent strength that allows them to generate speech exhibiting superior prosody, intonation, and overall naturalness (12–16).

Most previous AR TTS models (13; 17) have required both speech and transcription as prompts

---

[1]Please send correspondence to model@minimax.io.

in voice cloning, a methodology categorized as one-shot learning. However, semantic or linguistic mismatches between prompt and target speech, compounded by decoding length limitations, often result in suboptimal generation quality. MiniMax-Speech distinguishes itself by integrating a speaker encoder into its AR model. This key feature, based on concepts explored in works like (18), enables zero-shot voice cloning using only a speech prompt. This approach eliminates the dependency on reference transcription, thus inherently supporting cross-lingual and multilingual synthesis and avoiding issues from text-speech mismatches. By conditioning solely on vocal characteristics, it allows for a wider, more flexible decoding space for prosody and style during generation, leading to richer and more varied outputs. Notably, even in one-shot scenarios where a reference transcription is available, this integrated speaker encoder contributes to superior speaker similarity in the synthesized speech. The strategy employed by MiniMax-Speech thus effectively mitigates these common issues and offers greater flexibility for various extensions, such as T2V and PVC.

Another technical aspect warranting discussion is the inherent learnability of the speaker encoder we adopted (14; 18). Some models often utilize modules pre-trained on Speaker Verification (SV) tasks as their fixed speaker encoders (19). However, the training data and optimization objectives inherent to SV tasks can differ from the specific requirements of the TTS task itself. In contrast, our learnable approach involves jointly training the speaker encoder with the AR model. This end-to-end optimization strategy allows the speaker encoder to be better tailored to the demands of the TTS task. Subsequent experimental results compellingly demonstrate that employing a learnable speaker encoder yields superior performance in terms of both speaker similarity and intelligibility of the synthesized speech.

Flow matching models (20; 21), when utilized as decoders, are capable of producing high-fidelity speech outputs (17; 19). A prevalent paradigm involves flow matching models first predicting mel-spectrogram, which is subsequently converted to an audio waveform by a vocoder. However, the mel-spectrogram can act as an information bottleneck, inherently limiting the ultimate achievable speech quality. In contrast, VAE, benefiting from end-to-end training, demonstrates stronger representation learning capability. Employing VAE-derived representations as the modeling objective for flow matching systems subsequently improves speech quality (22–24). Building upon this VAE foundation, MiniMax-Speech innovatively introduces Flow-VAE. This hybrid approach integrates VAE and flow models to enhance the information representation power of the VAE encoder, thereby further improving both audio quality and speaker similarity.

We summarize our contributions as follows:

1. We present MiniMax-Speech, a TTS model that supports 32 languages and generates high-fidelity speech with near-indistinguishable human resemblance, achieving SOTA results on multiple objective and subjective evaluation metrics.
2. Based on autoregressive Transformer architecture and equipped with a learnable speaker encoder module, our model demonstrates strong expressiveness in zero-shot voice cloning. Furthermore, it enhances speaker similarity in one-shot scenarios when provided with a reference audio prompt.
3. We employ a flow matching model based on our novel Flow-VAE, which further improves the audio quality and speaker similarity of the generated speech.
4. We also detail several downstream applications of our model, including fine-grained control over emotional expression in synthesized speech, extensive voice library via T2V, and improving synthesis similarity for target speakers by PVC.
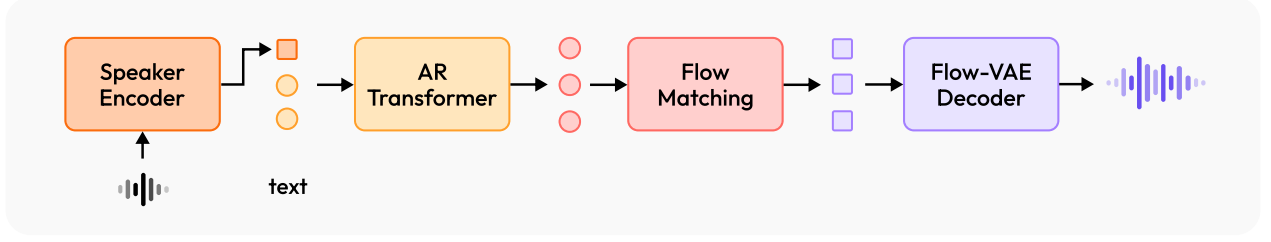
Figure 1 | **An overview of the architecture of MiniMax-Speech.**

2. Method

MiniMax-Speech is an innovative TTS system designed for high-fidelity voice cloning, with a particular emphasis on its robust zero-shot capabilities. As shown in Figure 1, it primarily comprises three components: a tokenizer, an autoregressive Transformer, and a latent flow matching model, which consists of flow matching module and Flow-VAE module. The text tokenizer utilizes Byte Pair Encoding (BPE), while the audio tokenizer employs Encoder-VQ-Decoder architecture (18; 25) quantization on mel-spectrograms with a rate of 25 tokens per second and connectionist temporal classification (CTC) supervision. This speech tokenizer achieves a high compression rate, while effectively preserving ample acoustic details and semantic information. The details of the autoregressive Transformer and latent flow matching model are as follows.

2.1. Zero-Shot Autoregressive Transformer

MiniMax-Speech employs an autoregressive Transformer (26) architecture to generate discrete audio tokens from textual input. The system excels at high-fidelity speaker cloning, particularly zero-shot voice cloning, where it synthesizes speech emulating a target speaker's distinct timbre and style from only a single, untranscribed audio segment.

To enable this powerful zero-shot capability, MiniMax-Speech incorporates a learnable speaker encoder, inspired by (18). In contrast to other speech synthesis models that use pre-trained speaker encoders (19; 27), the encoder in MiniMax-Speech is trained jointly with the autoregressive Transformer. This joint optimization allows the speaker encoder to be specifically tailored to the speech synthesis task, enhancing the model's synthesis quality by providing richer and more relevant speaker-specific information. Additionally, since the speaker encoder is learnable, it can be trained on all languages within the training dataset. Compared to pre-trained speaker encoders that might not have been exposed to the same diversity of languages, our learnable speaker encoder ensures broader linguistic coverage and potentially enhances generalization.

The speaker encoder extracts salient speaker-specific characteristics, such as vocal timbre and prosodic style, from the reference audio (which differs from the target speech to be generated). Variable-length audio segments serving as voice prompts are transformed by this encoder into a fixed-size conditional vector. This vector subsequently guides the autoregressive model in generating the target speech with the desired speaker identity.

The voice cloning capabilities of Minimax-Speech are best understood through the paradigms of zero-shot and one-shot learning, concepts adapted from the capabilities observed in Large Language Models (LLMs) like GPT-3 (28). In LLMs, zero-shot refers to performing a task based solely on an instruction without any prior examples, while one-shot (or few-shot) involves providing one (or a few) in-context examples to guide the model. We adapt these concepts to TTS as follows:

- **Zero-Shot Voice Cloning**: The Core Strength of MiniMax-Speech. In this primary mode, Minimax-Speech synthesizes speech in a target speaker's voice using only a reference audio segment to define the voice characteristics (shown in Figure 2b). Crucially, no explicit examples of that speaker's voice paired with text are provided at inference time as prompts, and no speaker-specific fine-tuning is performed. The reference audio itself acts as the primary "instruction" for the desired vocal timbre and style.
- **One-Shot Voice Cloning**: An Optional Enhancement. Building upon the zero-shot foundation, this mode enhances cloning fidelity by providing an additional explicit example. Specifically, a paired text-audio sample from the target speaker is supplied as an "in-context" prompt alongside the standard speaker embedding derived from the reference audio (shown in Figure 2c). This approach mirrors the one-shot prompting strategy in LLMs and is similar to techniques in prior works like VALL-E (13), CosyVoice 2 (17) and Seed-TTS (6) (their prompting method, which requires a paired text-audio sample, is illustrated in Figure 2a). While these aforementioned models are often described as "zero-shot" in their respective publications, their reliance on a paired text-audio prompt for speaker conditioning categorizes them as "one-shot" methods according to our stricter definition. Our "instrinsic zero-shot" approach (Figure 2b), in contrast, exclusively utilizes an untranscribed reference audio segment to derive speaker characteristics, without any accompanying text prompt.

While the optional one-shot prompting can offer finer-grained stylistic cues in specific scenarios, the system's architecture is fundamentally designed for powerful and flexible zero-shot synthesis. The conditioning encoder in MiniMax-Speech seamlessly supports both methods, but its true innovation lies in enabling high-quality voice cloning without reliance on paired data or fine-tuning. The advantages of this zero-shot-centric design, facilitated by the learnable speaker encoder, are manifold:

- **Text-Free Reference**: By operating solely on the reference audio waveform, it eliminates the need for textual transcriptions of the target speaker's audio. This ensures that the speaker identity is learned purely from vocal characteristics, disentangled from the semantic content of any specific reference utterance.
- **Rich Prosodic Variation and Flexible Decoding**: The zero-shot approach, conditioned only on the speaker condition extracted by the encoder, allows for the generation of speech with diverse prosodic variations. The model is not constrained by the prosody of a specific text-audio prompt (as in one-shot methods), leading to a wider decoding space and outputs that maintain high fidelity to the target speaker's unique vocal identity while exhibiting a natural range of expression.
- **Robust Cross-Lingual Synthesis**: The speaker encoder captures language-agnostic vocal characteristics, enhancing cross-lingual synthesis. This outperforms prompt-based cloning methods that rely on text-speech reference pairs, which struggle when the reference language differs from the target language or when semantic content mismatches.
- **Foundation for Extensibility**: The robust and disentangled speaker representation provided by the encoder serves as a flexible foundation for various downstream applications, as detailed in Section 4. Tasks like emotion control, T2V, and PVC can leverage this core speaker identity representation without fundamentally altering the base model.
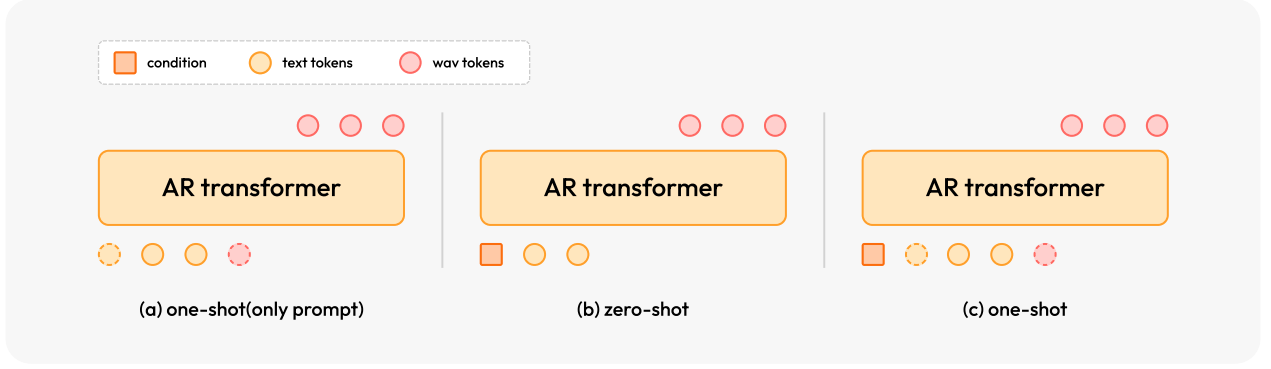
Figure 2 | **Different Voice Cloning Approaches in AR Transformer.** The dotted line represents a provided example of a text-to-speech pair.

2.2. Latent Flow Matching

2.2.1. Overview

In MiniMax-Speech, the flow matching model utilizes a transformer (26) architecture, which has powerful context modeling capabilities. Our flow matching model is designed to model the distribution of continuous speech features (latent), which are extracted from an encoder-decoder module trained on audio, rather than mel-spectrograms. When training this encoder-decoder module (where its encoder extracts these continuous speech features and its decoder is typically a neural vocoder (29–33)), the KL divergence is employed as a constraint. This renders the latent distribution easier to predict and more compact. Furthermore, due to the joint training of the latent extraction module (encoder) and the neural vocoder (decoder), the waveform reconstruction error from latent features is smaller compared to that from mel-spectrograms, which elevates the ceiling of latent feature modeling.

Figure 3(a) illustrates the proposed Flow-VAE model, which we use to optimize the encoder-decoder module. Traditional Variational Autoencoders (VAEs) typically assume a standard normal distribution for their latent space. In contrast, Flow-VAE introduces a flow model (34–37), which can flexibly transform the latent space using a series of reversible mappings to learn a more expressive posterior distribution to more accurately capture complex patterns in the data. This fusion solution can make full use of VAE's initial modeling ability of data and the flow model's accurate fitting ability of complex distribution, so as to better capture the complex structure and distribution characteristics in the data, improve the accuracy of data modeling, and thus significantly outperforming the traditional VAE model.

To enhance the audio quality and timbre similarity of the flow matching model, inspired by CosyVoice 2 (17), we incorporate both global timbre information and prompt information as mentioned in Figure 3(b). Specifically, the global timbre information is extracted from mel-spectrogram features using the speaker encoder. During the training process, information from the beginning of the current sentence is utilized as a prompt with a certain probability. Consequently, at inference stage, our model supports both zero-shot and one-shot synthesis modalities.

2.2.2. KL-Divergence for Flow-VAE

In the Flow-VAE model, our goal is to provide enough information for the posterior encoder, which is the encoder of the Flow-VAE model, so we use the waveform of the target speech x as input instead of the mel spectrogram, then we apply a flow model f_θ to reversibly transform a normal distribution
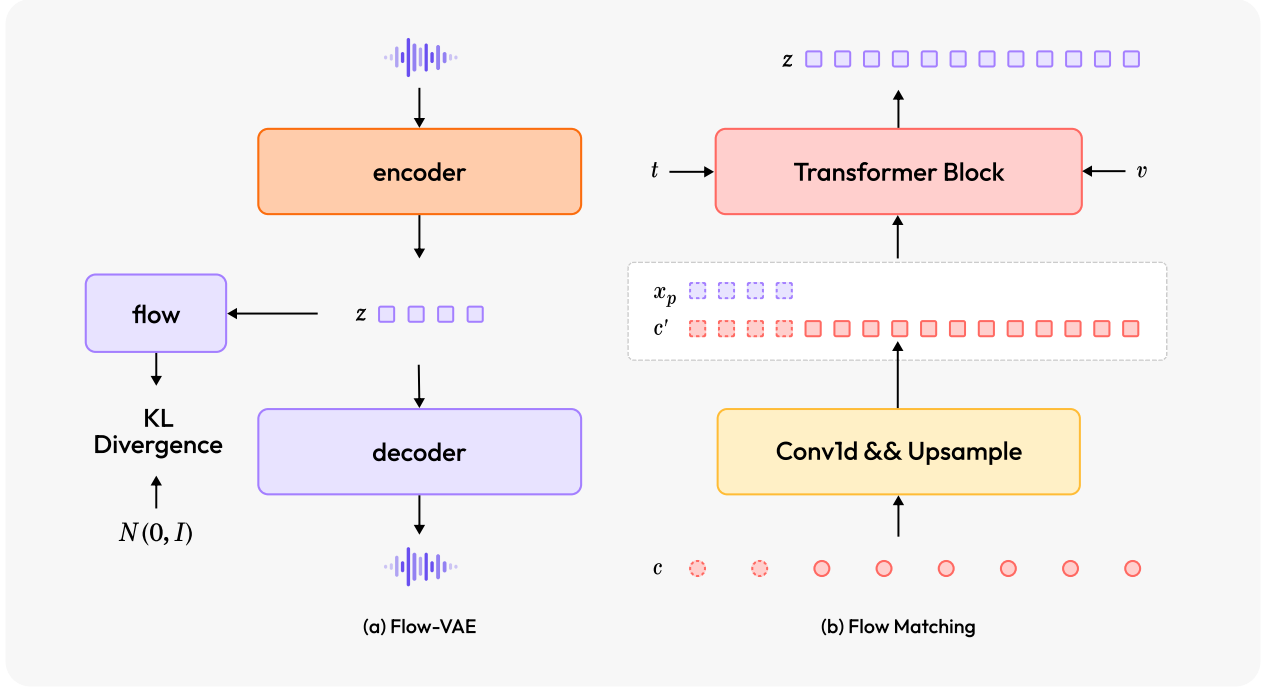
Figure 3 | **An overview of the proposed latent flow matching Architecture.** (a) The Flow-VAE model consists of an encoder, which is used to extract continuous speech features z, and a decoder, which is used to restore continuous speech features back to waveforms, and a flow model, which converts the distribution of continuous speech features to a standard normal distribution. z represents continuous speech features, which is the target for the flow matching model to generate. (b) The flow matching model, which conditions on AR transformer output c, a speaker embedding v, provided continuous speech features x_p and intermediate state x_t at timestep t on the probabilistic density path. c' represents the output of AR transformer after upsampling. The dotted box indicates the provided prompt information.

into a standard normal distribution. The KL divergence is:

$$L_{kl} = D_{KL}(q_\phi(\tilde{z}|x)||p(\tilde{z})) = \log q_\phi(\tilde{z}|x) - \log p(\tilde{z}) \tag{1}$$

$$q_\phi(\tilde{z}|x) = N(f_\theta(\tilde{z}); \mu_\phi(x), \sigma_\phi(x)) \left| det \frac{\partial f_\theta(\tilde{z})}{\partial \tilde{z}} \right| \tag{2}$$

$$z \sim N(\mu_\phi(x), \sigma_\phi(x)) \tag{3}$$

$$p(\tilde{z}) = N(\tilde{z}, 0, I) \tag{4}$$

In our experiment, as show in Figure 3(a), the flow model transforms the normal distribution output by the encoder through a series of reversible transformations. Finally, we calculate the KL loss between the distribution output by the flow model and the standard normal distribution. In this way, the output of the encoder can be constrained to a normal distribution instead of a standard normal distribution, enhancing the information expression ability of the encoder.

3. Experiments

This section presents a comprehensive evaluation of MiniMax-Speech, systematically assessing its performance across multiple dimensions. We begin by describing the datasets employed for training and evaluation. Our main analysis focuses on three key aspects: (1) voice cloning fidelity, objectively measured for both zero-shot and one-shot approaches; (2) perceptual naturalness, evaluated through extensive human preference tests; and (3) multilingual and cross-lingual synthesis capabilities, rigorously tested across diverse languages. Additionally, we conduct ablation studies to investigate the impact of key architectural decisions, including speaker conditioning methodology and the Flow-VAE framework.

3.1. Datasets

Minimax-Speech is trained on a multilingual speech dataset spanning 32 languages. Throughout the collection process, recognizing the paramount importance of transcription accuracy, we implemented a rigorous dual Automatic Speech Recognition (ASR) verification process. Text punctuation was further refined through the comprehensive consideration of VAD and ASR-generated timestamps. Notably, the original steady-state noise inherent in the recordings was preserved. Furthermore, consistent vocal timbre was maintained within each audio file by a multi-speaker verification model.

3.2. Voice Clone Evaluation

The fidelity of voice cloning was quantitatively assessed using WER and SIM metrics on the Seed-TTS-eval (6) test set. This dataset comprises two distinct subsets: test-zh (approximately 2,000 Chinese samples) and test-en (approximately 1,000 English samples). Each sample in these subsets includes a reference audio and a corresponding ground-truth audio from the identical speaker. For WER computation, synthesized English and Chinese audio were transcribed using Whisper-large-v3 (38) and Paraformer-zh (39), respectively. SIM was determined by calculating the cosine similarity between speaker embeddings, which were extracted using a speaker verification model fine-tuned on WavLM-large. These choices of ASR and speaker verification models align with the established methodology of the Seed-TTS-eval test set.

Table 1 | **Objective Evaluation Metrics on the Seed-TTS Test Set.** The **bolded value** indicates the best indicator for each column, and the underlined value indicates the second best. Note that the reference audio is utilized as input to the speaker encoder in both cloning methods of MiniMax-Speech, and it additionally serves as prompt exemplar for the one-shot paradigm.

Model	Cloning Method	test-zh		test-en	
		WER ↓	SIM ↑	WER ↓	SIM ↑
Ground Truth	-	1.25	0.750	2.14	0.730
Seed-TTS	one-shot	1.12	0.796	2.25	**0.762**
CosyVoice 2	one-shot	1.45	0.748	2.57	0.652
MiniMax-Speech	zero-shot	**0.83**	0.783	**1.65**	0.692
MiniMax-Speech	one-shot	0.99	**0.799**	1.90	0.738

As presented in Table 1, the MiniMax-Speech model achieved markedly lower WER in both zero-shot and one-shot cloning scenarios compared to Seed-TTS (6), CosyVoice 2 (17), and ground truth audio. This demonstrates that speech synthesized by MiniMax-Speech during cloning is characterized by clear, stable pronunciation and a reduced incidence of articulation errors. Notably, the WER for MiniMax-Speech in zero-shot cloning was superior to its one-shot counterpart. Furthermore,

subjective listener feedback indicated that speech synthesized via zero-shot cloning was perceived as more natural and realistic. The zero-shot approach, empowered by our speaker encoder, directly leverages the reference audio's acoustic properties without the additional influence of a language model prompt exemplar. This leads to superior intelligibility (lower WER) and enhanced naturalness, as the model has greater freedom in generating prosody faithful to the text being synthesized, rather than being biased by the prompt's prosody. The speaker encoder effectively captures the core vocal identity, allowing the autoregressive model to generate diverse and natural speech. While one-shot prompting improves SIM, the zero-shot method demonstrates a compelling balance of clarity and naturalness.

Regarding SIM, the MiniMax-Speech model achieved a SIM score in zero-shot cloning comparable to that of the ground truth audio. This underscores the speaker encoder's efficacy in extracting and preserving speaker identity even without textual or prosodic cues from a prompt. When an exemplar audio was introduced as a prompt in the one-shot cloning setting, the SIM score surpassed that of the ground truth audio, outperformed CosyVoice2, and was on par with Seed-TTS. This finding suggests that the incorporation of a prompt exemplar, building upon the zero-shot approach, can further augment the similarity of the cloned voice, potentially by providing more explicit cues for fine-grained vocal characteristics.

3.3. Subjective Evaluation

To comprehensively evaluate MiniMax-Speech in real-world scenarios, we submitted our model to Artificial Arena[2], a public TTS model leaderboard. Artificial Arena ranks models using ELO scores, derived from human preference judgments as users listen to and compare speech samples from various models. For this demanding evaluation, all speech samples from MiniMax-Speech were generated using its advanced zero-shot speaker cloning capability. This approach, while offering immense flexibility, presents a significant challenge in achieving SOTA quality.

As presented in Figure 4, our model, MiniMax-Speech (referred to as Speech-02-HD[3] on the leaderboard), secured the leading position. This top ranking not only places MiniMax-Speech ahead of a strong field of competitors but also underscores its distinct advantages. Specifically, when compared to other leading models such as those from OpenAI and ElevenLabs, the ELO scores reflect a clear user preference for MiniMax-Speech's superior naturalness and heightened expressiveness.

Perhaps even more strikingly, MiniMax-Speech demonstrated a considerable ELO score advantage over models from major technology providers like Google, Microsoft, and Amazon. This significant gap suggests that MiniMax-Speech's underlying architecture represents a more advanced, next-generation approach. Critically, our zero-shot generated speech achieved a level of quality and user preference that surpassed systems which often rely on extensive data to individually train models for specific speakers (e.g., requiring tens of hours of audio per voice) to reach their peak performance.

Crucially, achieving this high degree of naturalness and expressiveness—sufficient to outperform leading industry models, including those potentially built on extensive speaker-specific data—while relying exclusively on zero-shot cloned speaker timbres, robustly underscores the advanced capabilities and generalization power of our model. This outstanding performance in a public, preference-based benchmark highlights MiniMax-Speech's ability to deliver a highly compelling and human-like listening experience in real-world applications, even when generating novel voices on the fly.

---

[2]Artificial Arena: https://artificialanalysis.ai/text-to-speech/arena?tab=leaderboard

[3]The MiniMax-Speech model described in this paper corresponds to the Speech-02-HD model in Artificial Arena. The Speech-02-Turbo model, which also participated in the evaluation, employs a different model architecture primarily to enhance inference speed and reduce operational costs. The T2A-01-HD model is an older version of our model.

Artificial Analysis Speech Arena Leaderboard

Creator	Model	Arena ELO	95% CI	# Appearances
MiniMax	Speech-02-HD	1153	−24/+24	1187
OpenAI	TTS-1 HD	1151	−14/+12	6576
MiniMax	Speech-02-Turbo	1136	−23/+23	1163
OpenAI	TTS-1	1135	−14/+14	6456
ElevenLabs	Multilingual v2	1116	−13/+13	7824
ElevenLabs	Turbo v2.5	1112	−14/+13	7633
ElevenLabs	Flash v2.5	1110	−17/+15	2886
Cartesia	Sonic English (Oct '24)	1107	−13/+12	8109
OpenAI	GPT-4o Realtime Preview	1101	−29/+31	644
Kokoro	Kokoro 82M v1.0	1090	−17/+16	2873
MiniMax	T2A-01-HD	1081	−16/+16	2946
OpenAI	GPT-4o mini TTS	1066	−33/+33	485
Amazon	Polly Generative	1063	−16/+15	3042
Microsoft Azure	Azure Neural	1058	−13/+12	7899
Amazon	Polly Long-Form	1058	−15/+15	4541
Speechify	Simba	1047	−15/+16	3148
Google	Studio	1039	−14/+15	4797

Figure 4 | **Artificial Arena Evaluation Results (May 12, 2025).**

3.4. Multilingual Evaluation

MiniMax-Speech supports synthesis in 32 languages. To assess its multilingual performance, we constructed a dedicated test set comprising 24 languages[4]. For each language, this set includes 100 distinct test sentences. Synthesized speech was generated using the cloned voices of two speakers per language—one male and one female—sourced from the Mozilla Common Voice dataset ([40]). Each speaker rendered 50 unique sentences from the 100 available for that language. We assessed the performance of MiniMax-Speech against the ElevenLabs Multilingual v2 model in multilingual synthesis. Speech from both models was synthesized using zero-shot voice cloning. The evaluation metrics and methodology are consistent with those described in Section 3.2. For all languages except Chinese, the Whisper-large-v3 model was used for text recognition.

As indicated in Table 2, regarding WER, the performance of MiniMax-Speech is comparable to that

---

[4]https://huggingface.co/datasets/MiniMaxAI/TTS-Multilingual-Test-Set

Table 2 | **Objective Evaluation Metrics on the Multilingual Test Set.**

Language	WER ↓		SIM ↑	
	MiniMax	ElevenLabs	MiniMax	ElevenLabs
Chinese	**2.252**	16.026	**0.780**	0.677
English	**2.164**	2.339	**0.756**	0.613
Cantonese	**34.111**	51.513	**0.778**	0.670
Japanese	**3.519**	10.646	**0.776**	0.738
Korean	**1.747**	1.865	**0.776**	0.700
Arabic	**1.665**	1.666	**0.736**	0.706
Spanish	**1.029**	1.084	**0.762**	0.615
French	**4.099**	5.216	**0.628**	0.535
Italian	**1.543**	1.743	**0.699**	0.579
Vietnamese	**0.880**	73.415	**0.743**	0.369
Thai	**2.701**	73.936	**0.800**	0.588
Turkish	1.520	**0.699**	**0.779**	0.596
Indonesian	1.237	**1.059**	**0.729**	0.660
Portuguese	1.877	**1.331**	**0.805**	0.711
Dutch	1.143	**0.803**	**0.738**	0.680
German	1.906	**0.572**	**0.733**	0.614
Russian	4.281	**3.878**	**0.761**	0.676
Ukrainian	1.082	**0.997**	**0.730**	0.647
Polish	1.415	**0.766**	**0.802**	0.729
Romanian	2.878	**1.347**	**0.809**	0.699
Greek	2.016	**0.991**	**0.826**	0.733
Czech	3.875	**2.108**	**0.796**	0.685
Finnish	4.666	**2.964**	**0.835**	0.759
Hindi	6.962	**5.827**	**0.818**	0.730

of Multilingual v2. For languages such as Chinese, Cantonese, Thai, Vietnamese, and Japanese, where Multilingual v2 exhibited a WER exceeding 10%, MiniMax-Speech consistently outperformed it. This robust performance, particularly in languages with complex tonal structures or diverse phonetic inventories (e.g., Chinese, Cantonese, Thai, Vietnamese), suggests MiniMax-Speech's architecture is adept at capturing and reproducing nuanced acoustic features critical for intelligibility in these languages, an area where Multilingual v2 appears to face greater challenges. Concerning SIM, MiniMax-Speech demonstrated markedly superior SIM scores across all tested languages compared to the Multilingual v2 model. This consistent superiority in SIM across a diverse linguistic landscape underscores the effectiveness of MiniMax-Speech's speaker encoder and synthesis pipeline in preserving speaker identity, irrespective of the target language's phonetic characteristics, a key benefit of its text-agnostic reference processing. This suggests that MiniMax-Speech produces cloned speech that is closer to the ground truth human voice across the 24 evaluated languages.

3.5. Cross-lingual Evaluation

A significant advantage of MiniMax-Speech, stemming from its speaker encoder architecture, is its inherent support for cross-lingual speech synthesis. This enables the synthesis of speech for any given speaker in all languages supported by the model. Two key aspects contribute to this capability:

Firstly, for zero-shot speaker cloning, MiniMax-Speech requires only a short audio segment from the target speaker, without any corresponding transcription. This minimalist data requirement significantly lowers the barrier to entry and operational complexity for cloning new voices. This contrasts with some one-shot cloning models (6; 17) that necessitate transcribed reference audio. Such a dependency on transcriptions not only complicates the cloning process but also introduces the risk of transcription errors negatively impacting the quality of the cloned voice. MiniMax-Speech's approach, by eliminating the need for transcriptions in its zero-shot cloning, simplifies the workflow and mitigates potential issues arising from inaccurate transcriptions.

Secondly, the speaker encoder module extracts a conditional vector that primarily captures voice timbre while being largely devoid of textual semantic information. This characteristic facilitates the model's ability to decouple voice timbre from linguistic content and subsequently recombine them, thereby enabling each distinct voice timbre to articulate speech across all supported languages.

To validate the cross-lingual synthesis capabilities enabled by the speaker encoder, we conducted evaluations using Chinese speakers from our multilingual test set. This involved synthesizing speech from these Chinese speakers uttering phrases in various other target languages.

Table 3 | **Cross-lingual Speech Synthesis Performance of MiniMax-Speech (Zero-shot vs. One-shot).**

Target Language	WER ↓		SIM ↑	
	Zero-Shot	One-Shot	Zero-Shot	One-Shot
Czech	**2.823**	5.096	0.605	**0.648**
Romanian	**3.081**	5.353	0.625	**0.69**
Finnish	**4.527**	8.112	0.554	**0.655**
Thai	**2.826**	4.107	0.729	**0.748**
Arabic	**1.446**	2.649	0.619	**0.632**
French	**4.497**	5.489	0.586	**0.645**
Vietnamese	**0.659**	1.788	0.692	**0.725**

As indicated in Table 3, MiniMax-Speech, when utilizing its zero-shot cloning method, achieves significantly lower WER across all tested languages compared to its one-shot method. Furthermore, the achieved WERs indicate a high level of intelligibility, approaching that of high-quality native synthesis in the target languages. These results demonstrate that the speaker encoder architecture provides MiniMax-Speech with excellent cross-lingual synthesis capabilities.

In contrast, while MiniMax-Speech's one-shot cloning approach yields higher SIM, its pronunciation accuracy in cross-lingual synthesis, as indicated by its notably higher WER, is considerably poorer. Consequently, these findings underscore the advantages of MiniMax-Speech's speaker encoder architecture, highlighting its flexibility in supporting both zero-shot and one-shot cloning paradigms, and particularly its superior performance in zero-shot cross-lingual synthesis, as evidenced by high pronunciation accuracy.

3.6. Speaker Condition Evaluation

To evaluate the effectiveness of different speaker conditioning approaches, we conducted an ablation study using three distinct models trained on a substantial subset of our Chinese speech data. The first model implemented our learnable speaker encoder architecture, the second utilized speaker embeddings (SpkEmbed) extracted from a pre-trained speaker verification model (41), and the third employed a one-shot learning strategy with only an example audio prompt. We assessed these

configurations using WER and SIM metrics.

Table 4 | **Ablation Study on Speaker Conditioning Methods.**

Method	Cloning Mode	WER ↓	SIM ↑
Speaker Encoder	Zero-shot	1.252	0.730
Speaker Encoder	One-shot	1.243	0.746
SpkEmbed	Zero-shot	1.400	0.746
SpkEmbed	One-shot	1.704	0.744
OnlyPrompt	One-shot	1.207	0.726

Analysis of Table 4 demonstrates that the speaker encoder method provides the most robust performance, achieving strong results for both WER and SIM. In comparison, utilizing speaker embeddings from a pre-trained speaker model (SpkEmbed), while maintaining reasonable SIM, adversely affects the WER (e.g., 1.400 for SpkEmbed vs. 1.252 for Speaker Encoder in zero-shot), indicating a potential loss of speech clarity. This suggests an advantage of our learnable speaker encoder, which can be optimized jointly with the synthesis model, potentially adapting more effectively to the nuances of the target speech synthesis task compared to a fixed, pre-trained speaker verification model. Conversely, relying solely on the prompt (OnlyPrompt) in a one-shot setting, while achieving the best WER in this specific ablation study (1.207), significantly compromises SIM (0.726).

Our learnable speaker encoder, especially in one-shot mode (WER 1.243, SIM 0.746), strikes an optimal balance, surpassing SpkEmbed in WER and OnlyPrompt in SIM. These results confirm its effectiveness in preserving both speech intelligibility and voice characteristics. It thus offers a more balanced speaker conditioning solution than the alternatives. Its ability to maintain strong speaker identity (SIM 0.730) and good intelligibility (WER 1.252) in zero-shot synthesis further underscores its advantages. However, the reference audio for the speaker encoder must differ from the target audio for AR Transformer synthesis. Using identical audio during training can cause semantic leakage and degrade performance.

3.7. Flow-VAE Evaluation

To evaluate the performance of VAE and Flow-VAE, we conducted comparisons in two primary aspects: vocoder resynthesis and TTS synthesis. We randomly selected a portion from the open-source Chinese and English test sets of Seed-TTS (6) as our test set.

Vocoder Resynthesis: To compare the waveform reconstruction capabilities of VAE and Flow-VAE, we employed both models to perform resynthesis. Metrics were computed by comparing the synthesized audio with the original audio across multiple dimensions. The results, as presented in Table 5, indicate that the Flow-VAE model demonstrates significant advantages over the VAE model across all evaluated metrics.

TTS Synthesis: To assess the performance of latent features derived from VAE and Flow-VAE within a TTS framework, we trained flow matching models based on VAE latents and Flow-VAE latents, respectively on a substantial subset of our data. Following the WER and SIM evaluation methodologies from Seed-TTS (6), we generated test data in two inference settings: zero-shot and one-shot. The calculated WER and SIM scores are presented in Table 6.

It is worth noting that compared to the VAE model, Flow-VAE not only has advantages in the WER and SIM indicators. Upon listening to the synthesized audio, we found that Flow-VAE demonstrated significant advantages in overall stability. We encourage readers to experience through the demo

Table 5 | **Objective indicators of resynthesis by VAE and Flow-VAE.** SELF-SIM represents the similarity between the synthesized audio and the original audio, and PROMPT-SIM represents the similarity between the synthesized audio and the prompt audio.

Model	SELF-SIM ↑	PROMPT-SIM ↑	NB PESQ ↑	WB PESQ ↑	STOI ↑	MS-STFT-LOSS ↓
Dac-Vae	0.98	0.748	4.27	4.20	0.993	0.67
Dac-Flow-Vae	**0.986**	**0.75**	**4.34**	**4.30**	0.993	**0.62**

Table 6 | **Objective indicators of TTS synthesis by VAE and Flow-VAE.**

	Model	test-zh		test-en	
		WER ↓	SIM ↑	WER ↓	SIM ↑
zero-shot	AR Transformer+FM+VAE	0.753	0.747	1.717	0.633
	AR Transformer+FM+Flow-VAE	**0.748**	**0.751**	**1.639**	**0.639**
one-shot	AR Transformer+FM+VAE	**0.873**	0.776	2.242	0.707
	AR Transformer+FM+Flow-VAE	0.901	**0.782**	**2.231**	**0.709**

link.

4. Extensions

The disentangled and robust speaker representations learned by the integrated speaker encoder endow MiniMax-Speech with notable flexibility, facilitating its straightforward extension to various downstream applications. Because the speaker encoder captures pure vocal identity from reference audio without transcription, it provides a stable and versatile foundation upon which these extensions can be built. In this section, we detail three such extensions: (i) the control of emotional expression in synthesized speech utilizing the Low-Rank Adaptation (LoRA) technique; (ii) the generation of arbitrary and diverse vocal timbres from natural language descriptions; and (iii) professional voice cloning (PVC), a parameter efficient fine-tuning approach designed to enhance synthesis quality and fidelity for specific speakers by optimizing their associated embeddings.

4.1. Emotion Control

Emotional expression, conveyed through prosodic features like pitch and duration, is crucial for natural synthesized speech and is primarily modeled by the autoregressive Transformer in Minimax-Speech. We introduce a novel approach using LoRA ([42]) for precise emotional control. We define discrete emotion categories, train independent LoRA modules for each using high-quality emotion-specific datasets, and dynamically load the appropriate module during inference based on user selection. This method offers higher precision and stability in emotional expression compared to natural language control.

The efficacy of this approach depends a lot on the training data, which are formatted as <reference audio, text, target emotive audio>. Reference audio provides speaker identity and establishes an emotional contrast with the target emotive audio, which the LoRA module learns to bridge. We investigated different reference audio types:

- **Emotionally Congruent Reference**: Led to output emotion being overly reliant on the reference's emotion, limiting direct control.

- **Neutral or Random Emotion Reference**: Enabled effective control via the specified emotion category. Neutral references yielded higher expressiveness due to clearer emotional contrast, while random emotion references produced robustly natural speech with stable speaker similarity, likely by enhancing the model's ability to disentangle speaker identity from varied expressions.

To decouple synthesized emotion from lexical content, we collected multiple emotive audio samples for the same text, each with a different emotion. This trains the model to articulate identical content with varied emotional inflections, ensuring that the learned emotional rendering is independent of the text's semantic meaning.

A key advantage of this LoRA-based approach is that emotion-specific modules are trained without modifying the pre-trained Minimax-Speech core architecture. This simplifies training and deployment, preserves the original voice cloning performance, and offers excellent scalability. Experimental results show that our method achieves notable enhancements in the accuracy and naturalness of emotional expression compared to existing methodologies, producing more vivid and engaging utterances.

4.2. Text to Voice

Generating speech in a desired timbre with most existing TTS methods necessitates providing a reference audio sample of that specific timbre, a requirement that can limit their operational flexibility. In contrast, we introduce a T2V framework that uniquely integrates open-ended natural language descriptions with structured tag information. As a complement to the reference-audio-driven speaker encoder (which excels at cloning existing voices), this approach facilitates highly flexible and controllable timbre generation, thereby significantly enhancing the versatility of TTS systems.

Firstly, we curated a high-quality speech dataset with attributes including speech rate, gender, language, pitch, and volume. Inspired by Spark-TTS ([15]), these attributes were discretized. (e.g., pitch was divided into six bins [0, 1, 2, 3, 4, 5] according to its value in Hertz, with 0 denoting 'unknown'). These structured attributes are then combined with textual descriptions and speech data to form an aligned corpus of text-speech pairs.

Subsequently, timbre representations were extracted from AR transformer and the flow matching model. Principal Component Analysis (PCA) ([43]) was employed to compress these high-dimensional features into 128 dimensions, retaining core timbre characteristics while reducing the complexity of predicting these representations. These compressed timbre representations, in conjunction with structured attributes and textual descriptions, are subsequently inputted to a compact timbre generation model. This model is trained to map natural language timbre descriptions and discrete speech attributes onto the aforementioned compressed timbre representation space. During the training phase, a random masking augmentation mechanism was introduced: key semantic words within the textual descriptions are randomly masked with a predefined probability, thereby enhancing the model's robustness with incomplete input.

This proposed framework, by combining open-ended textual descriptions with structured tag parameters, establishes a versatile timbre generation system. This system effectively unifies textual descriptions with audio-derived timbre representations for controlling timbre, empowering users to generate desired vocal characteristics using natural language (e.g., "a warm, middle-aged female voice with a slightly fast speech rate"), thereby significantly enhancing the flexibility of audio replication scenarios.

4.3. Professional Voice Clone

The learnable speaker encoder in the MiniMax-Speech model not only affords the model enhanced flexibility in zero-shot voice cloning tasks (due to its text-independent operation and ability to capture pure vocal identity), but also critically provides a streamlined pathway for efficient and rapid parameter fine-tuning tailored to individual speakers. Drawing inspiration from contemporary Parameter-Efficient Fine-Tuning (PEFT) methodologies ([44]; [45]), we introduce a novel fine-tuning strategy. This strategy conceptualizes the conditional embedding, which encapsulates a specific speaker's vocal identity (initially derived from the speaker encoder's understanding of voice characteristics) as a set of learnable parameters. During the fine-tuning phase for a target speaker, this dedicated embedding is optimized, substituting the pre-existing speaker encoder.

Specifically, to optimize the voice timbre for any given target speaker, an initial collection of their speech data is acquired. Throughout the fine-tuning process, the autoregressive Transformer is employed as the underlying base model, and all its parameters are kept fixed (i.e., frozen). Optimization is performed exclusively on the conditional embedding associated with the target speaker, treating it as the only trainable parameter set for this adaptation. Subsequently, during the inference stage, this fine-tuned, speaker-specific conditional embedding is directly invoked to replace the real-time output generated by the standard speaker encoder.

The rationale behind PVC is to refine the speaker representation within the latent space established by the speaker encoder. Although the speaker encoder adeptly captures significant speaker information from reference audio for zero-shot voice cloning, the conditional embedding it generates for a specific speaker can be further optimized for enhanced accuracy if sufficient speech data from that speaker is available. Fine-tuning the compact conditional embedding is more tractable and offers greater flexibility for tailoring to an individual speaker compared to optimizing the entire, already well-generalized speaker encoder. Our experiments demonstrate that, with appropriate hyperparameter tuning, this PVC approach enables the model to synthesize speech exhibiting improved fidelity to the target speaker's unique timbre and superior overall perceptual quality, especially for speakers with strong accents or distinctive vocal characteristics.

This method also offers significant advantages in terms of scalability and efficiency. Because adaptation for each speaker requires optimizing only a singular vector embedding, the system readily facilitates fine-tuning and deployment for potentially thousands of distinct speakers. This is achieved without altering the foundational model's core architecture or necessitating the deployment of complete, individual models per speaker. Compared to Supervised Fine-Tuning (SFT) or even methods like LoRA, our proposed technique demonstrably curtails training complexity and reduces computational resource expenditure. This is achieved while concurrently ensuring enhancements in both the speaker similarity and the naturalness of the synthesized audio, highlighting its superior practicality and extensibility for real-world applications. For instance, its application in education allows for targeted fine-tuning to specific teacher voices, enabling the efficient generation of personalized audio content that enriches instructional materials and boosts learner engagement.

5. Conclusion

In this work, we have presented MiniMax-Speech, an autoregressive Transformer-based TTS model. Existing TTS methods, particularly for robust zero-shot voice cloning and high-fidelity synthesis, often face challenges such as a dependency on transcribed reference audio, which can limit cross-lingual capabilities and expressiveness, or they may struggle to achieve optimal audio quality and speaker similarity due to limitations in their generative components. To address these limitations, MiniMax-Speech introduces key innovations: its learnable speaker encoder and our novel Flow-VAE architecture

integrated within a flow matching mechanism. Specifically, the learnable speaker encoder enables robust zero-shot voice cloning by extracting speaker timbre directly from reference audio, crucially without requiring accompanying text, offering superior performance in cross-lingual scenarios and a wider decoding space for generating richer and more natural prosodic variations. Concurrently, our Flow-VAE enhances the information representation power of the audio generation process, further improving both the overall audio quality and speaker similarity of the synthesized speech. Through this combined approach, MiniMax-Speech capably supports synthesis in 32 languages. Furthermore, it has demonstrated SOTA performance on objective and subjective evaluations. Notably, this includes achieving top results on voice cloning metrics and securing the leading position on the public TTS Arena leaderboard. The extensibility granted by the speaker encoder has been showcased through applications like LoRA-based emotion control, text-driven timbre generation, and efficient professional voice cloning, establishing MiniMax-Speech as a powerful and versatile solution for high-fidelity, expressive, and controllable speech synthesis. Future work will explore further enhancements to controllability and efficiency.

References

[1] 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 *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[2] 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.

[3] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11341–11345. IEEE, 2024.

[4] Yuan Gao, Nobuyuki Morioka, Yu Zhang, and Nanxin Chen. E3 tts: Easy end-to-end diffusion-based text to speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

[5] 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.

[6] Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

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

[8] Dongchao Yang, Dingdong Wang, Haohan Guo, Xueyuan Chen, Xixin Wu, and Helen Meng. Simplespeech: Towards simple and efficient text-to-speech with scalar latent transformer diffusion models. *arXiv preprint arXiv:2406.02328*, 2024.

[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] Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689. IEEE, 2024.

[11] Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, et al. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*, 2025.

[12] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

[13] Chengyi Wang, Sanyuan Chen, 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. *arXiv preprint arXiv:2301.02111*, 2023.

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

[15] 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.

[16] Hao-Han Guo, Yao Hu, Kun Liu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283*, 2024.

[17] 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.

[18] James Betker. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*, 2023.

[19] 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.

[20] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[21] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Trans. Mach. Learn. Res.*, 2024, 2024.

[22] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4234–4245, 2024.

[23] 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.

[24] Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024.

[25] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[27] Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent Van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, et al. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv preprint arXiv:2402.08093*, 2024.

[28] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[29] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop, SSW 2016, Sunnyvale, CA, USA, September 13-15, 2016*, page 125. ISCA, 2016.

[30] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2415–2424. PMLR, 2018.

[31] Jean-Marc Valin and Jan Skoglund. LPCNet: Improving Neural Speech Synthesis Through Linear Prediction. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5891–5895. IEEE, 2019.

[32] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 492–498. IEEE, 2021.

[33] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

[34] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[35] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[36] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[37] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[38] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[39] Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. Funasr: A fundamental end-to-end speech recognition toolkit. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pages 1593–1597. ISCA, 2023.

[40] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, 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, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association, 2020.

[41] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. 2023.

[42] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[43] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[44] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics, 2021.

[45] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

A. Contributors

The contributors to the report are listed in alphabetical order as follows:

Bowen Zhang, Congchao Guo, Geng Yang, Hang Yu, Haozhe Zhang, Heidi Lei, Jialong Mai, Junjie Yan, Kaiyue Yang, Mingqi Yang, Peikai Huang, Ruiyang Jin, Sitan Jiang, Weihua Cheng, Yawei Li, Yichen Xiao, Yiying Zhou, Yongmao Zhang, Yuan Lu, Yucen He