Voila: Voice-Language Foundation Models for Real-Time Autonomous Interaction and Voice Role-Play

Yemin Shi*, Yu Shu*, Siwei Dong*, Guangyi Liu*, Jaward Sesay, Jingwen Li, and Zhiting Hu

Maitrix.org, UC San Diego, MBZUAI
*Equal contribution

Abstract

A voice AI agent that blends seamlessly into daily life would interact with humans in an autonomous, real-time, and emotionally expressive manner. Rather than merely reacting to commands, it would continuously listen, reason, and respond proactively, fostering fluid, dynamic, and emotionally resonant interactions. We introduce `Voila`, a family of large voice-language foundation models that make a step towards this vision. Voila moves beyond traditional pipeline systems by adopting a new end-to-end architecture that enables full-duplex, low-latency conversations while preserving rich vocal nuances such as tone, rhythm, and emotion. It achieves a response latency of just 195 milliseconds, surpassing the average human response time. Its hierarchical multi-scale Transformer integrates the reasoning capabilities of large language models (LLMs) with powerful acoustic modeling, enabling natural, persona-aware voice generation—where users can simply write text instructions to define the speaker's identity, tone, and other characteristics. Moreover, `Voila` supports over one million pre-built voices and efficient customization of new ones from brief audio samples as short as 10 seconds. Beyond spoken dialogue, `Voila` is designed as a unified model for a wide range of voice-based applications, including automatic speech recognition (ASR), Text-to-Speech (TTS), and, with minimal adaptation, multilingual speech translation. `Voila` is fully open-sourced to support open research and accelerate progress toward next-generation human-machine interactions.

Voila Project Page	voila.maitrix.org
Voila Demo	hf.co/spaces/maitrix-org/Voila-demo
Voila Base Model	hf.co/maitrix-org/Voila-base
Voila End-to-end Model	hf.co/maitrix-org/Voila-chat
Voila Full-duplex Model (preview)	hf.co/maitrix-org/Voila-autonomous-preview
Voila Tokenizer	hf.co/maitrix-org/Voila-Tokenizer
Voila Benchmark	hf.co/datasets/maitrix-org/Voila-Benchmark
Voila Voice Library	hf.co/datasets/maitrix-org/Voila-million-voice
Voila Code	github.com/maitrix-org/Voila

1 Introduction

Autonomous Interaction Most AI systems today interact with humans reactively: they wait passively and respond after receiving a user query. For instance, in conversational systems ranging from Siri to ChatGPT, the user asks a question, the system generates an answer, and then waits for the next prompt, resulting in rigid, turn-based interactions. While this command-driven scheme may be sufficient for basic AI assistants, it remains far from achieving a truly *autonomous* machine capable of engaging proactively and emulating the rich dynamics of human-to-human interaction. An autonomous AI would *continuously assess its*

context, anticipate user needs in real-time, and determine if, when, and how to interact in an optimal way (Buss, 2012; Grosinger, 2022; Buyukgoz et al., 2022; Hoke, 2021). For example, when a user is walking down a street, the AI might warn them about an approaching cyclist they had not noticed, or suggest a stop at a hidden gem café nearby. Similarly, if a user keeps expressing a low mood and spiraling into negative thoughts, the system might actively interrupt to suggest a calming activity tailored to the user's emotional needs, instead of passively waiting for the user to ask for help. This vision of autonomous interaction has been imagined in popular culture, as seen in the film *Her*, where AI interacts fluidly with humans, forming genuine and emotive bonds. Such autonomous interactions make AI more than just a passive tool, but a trusted companion and teammate that blends seamlessly into our daily life.

Among the various modes of communication—text, vision, and gestures—*voice* is perhaps the most essential and natural for carrying out autonomous interactions (Schafer, 1995; Flanagan, 1972). Unlike text-based communication, which is often static, asynchronous, and turn-based, voice naturally enables rich, dynamic, and human-like interactions. For example, we speak to draw attention and initiate dialogue (even when the other person is not looking), interrupt or overlap speech to signal urgency or redirect conversational flow, and use simple backchanneling cues like '`mmh`' or '`yeah`' to convey attentiveness and engagement when others speak (Skantze, 2021; Yang et al., 2022). In addition, voice carries rich vocal cues (such as tone, inflection, and rhythm) and subtle emotional nuances that other modalities cannot replicate (Bora, 2024; Schroeder and Epley, 2016). As a result, a voice-based interface is crucial for creating an engaging and immersive experience in human-machine interactions.

Voice AI From Bell Labs' Audrey in 1952, which could recognize the sound of a spoken digit from 0 to 9, to recent advancements like ChatGPT-4o (Hurst et al., 2024) for open-ended spoken dialogue, voice AI has undergone a remarkable evolution, as illustrated in Figure 1. Early voice systems like Apple Siri, Amazon Alexa, and Google Assistant, launched in the 2010s, pioneered the first widely used voice conversational interfaces and relied on complex modular pipelines (Figure 1a). The components in these systems require extensive hand-engineering and are limited to processing a constrained set of user queries. Recent large language models (LLMs) have ushered in a simpler pipeline design that can support open-ended conversations (Figure 1b). This new approach consists of three main components: automatic speech recognition (ASR) for converting human speech into text, LLM for generating text responses, and text-to-speech (TTS) for converting the text responses to audio. The core of this pipeline is the text-based conversation managed by the LLM, while ASR and TTS work to convert voice into text and text back into voice, enabling audio interfaces.

This pipeline approach leverages the strengths of LLMs in text-based interaction, such as broad knowledge, complex reasoning (Wei et al., 2022; Yao et al., 2022; Hao et al., 2023a,b), instruction following (Ouyang et al., 2022), and role playing (Wang et al., 2023c; Shao et al., 2023; Wu et al., 2024). However, the pipeline design also comes with fundamental limitations that hinder truly natural, human-like voice interactions, including *(1) high latency, (2) loss of vocal nuances, and (3) reactive, turn-based interactions.* Specifically, each module in the pipeline can introduce delays, which often accumulate to several seconds—significantly longer than the average human response time of under 300 milliseconds (Stivers et al., 2009; Meyer, 2023). Additionally, converting audio to text for LLM processing results in a loss of rich acoustic cues, such as tone, accent, emotion, and background sounds. Without these nuances, the system struggles to interpret and respond naturally to the subtleties of human communication—for example, the phrase "`Oh, really`" might express surprise or disinterest depending on the speaker's tone. Moreover, the pipeline inherits the rigid, turn-based structure of text-based conversation mentioned earlier, where speakers respond only after the other has finished. The approach thus fails to capture the essence of natural, autonomous voice interactions. For instance, common conversational elements such as backchanneling, interruptions, and overlapping speech are absent, resulting in interactions that feel mechanical rather than unfolding organically with spontaneity, dynamism, and engagement. While additional components can be added in the pipeline, such as interruption detection to let users interrupt the system's responses (Lebourdais et al., 2024) or various triggers that activate the system to initiate interactions under specific conditions (e.g., at a user-specified

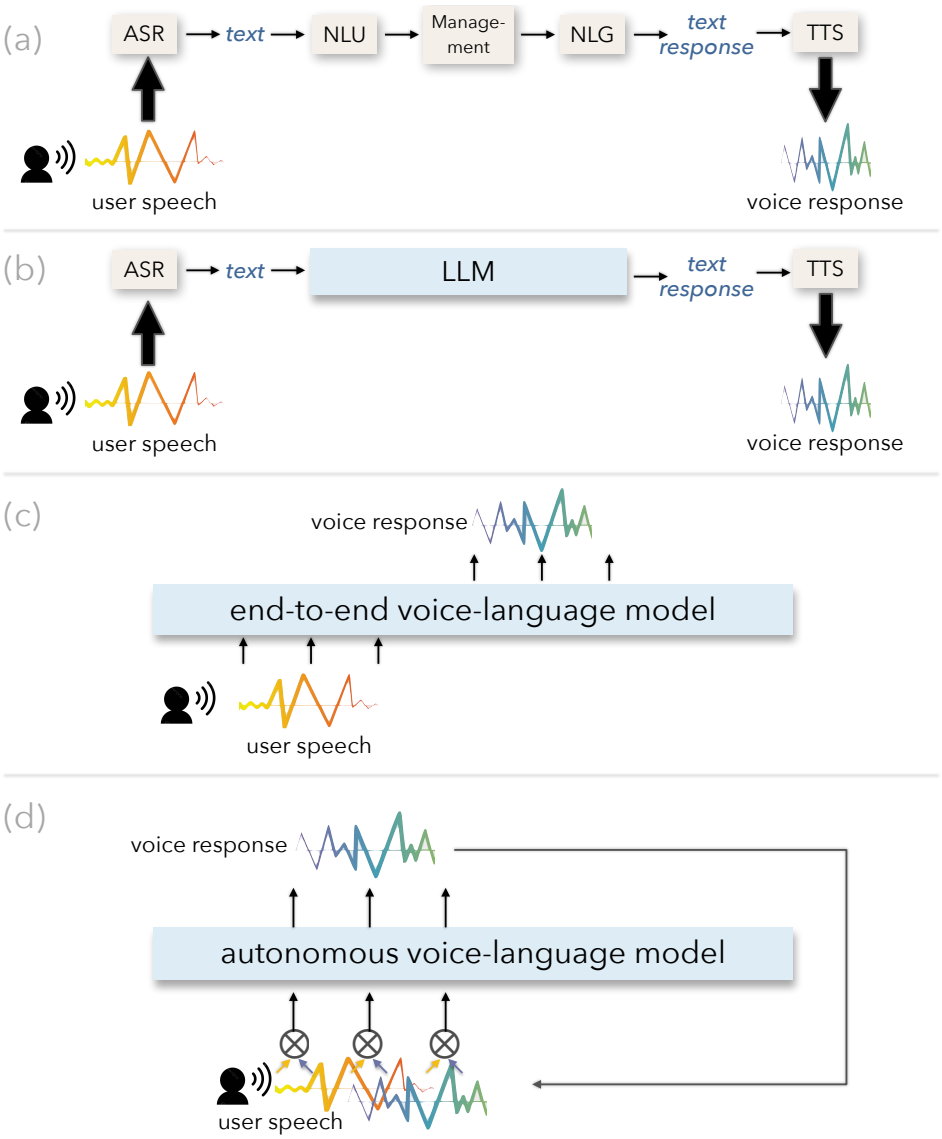

Figure 1: Different paradigms of voice conversation systems: **(a)** Traditional pipeline systems, such as Apple Siri, Amazon Alexa, and Google Assistant, launched in the 2010s; **(b)** Simplified pipeline systems using LLMs to handle text-based understanding and response generation; **(c)** End-to-end audio-in, audio-out systems that offer low latency and rich vocal nuances; **(d)** Autonomous systems that further enable dynamic, proactive interactions.

time) (Bérubé et al., 2024), such rule-based controls are inherently limited. They lack deeper contextual understanding and autonomy necessary to achieve natural, dynamic interactions.

Recent advancements have further led to *end-to-end* audio-language models that bypass the traditional pipeline architecture (Figure 1c). Instead of converting audio signals into text, this approach processes audio representations (e.g., audio tokens) directly within a large model, which then generates a response in the same audio representation space before decoding it into a voice output. By removing text as an intermediate step, the method preserves rich acoustic details in the input and enables nuanced acoustic generation in the output, permitting more natural and engaging interactions. In addition, without the cascading delays of a multi-module pipeline, end-to-end models can achieve lower latency. On the other hand, however, the system still adheres to a reactive, turn-based interaction flow as in prior approaches.

Voila for Autonomous Voice Interaction We introduce `Voila`, a family of large audio-language foundation models aiming to overcome the above challenges and enable real-time, natural, and flexible voice interaction. Specifically, `Voila-e2e` is an end-to-end model for natural voice conversations with low latency, rich vocal details, and strong instruction following and customizability. `Voila-autonomous` further aims for autonomous interaction, where the model continuously listens, reasons, and responds in a full-duplex and simultaneous manner (Figure 1d), delivering a next-level voice interaction experience.

`Voila` designs a hierarchical Transformer architecture, including streaming audio encoding and tokenization, and multi-scale Transformers consisting of an LLM backbone and a hierarchical audio generator (Figure 2). The models are trained in an end-to-end way with extensive audio-text data. In summary, `Voila` features the following key advancements:

- **Effective integration of voice and language modeling capabilities**: `Voila` adopts a range of designs to best combine the text-based capabilities of the pre-trained LLM and the newly learned voice modeling capabilities. For instance, much like how users can input a text prompt to instruct an LLM's behavior, `Voila` allows users to do the same on its backbone LLM to define its persona and steer its responses in voice conversations. In addition, `Voila`'s voice mode retains the extensive knowledge and linguistic proficiency acquired during LLM pretraining, ensuring high-quality model responses.

  To this end, `Voila` uses a multi-scale Transformer architecture (Yang et al., 2023; Zhu et al., 2024; Défossez et al., 2024) that predicts semantic and acoustic tokens separately at different levels. The disentanglement allows the backbone LLM to focus on handling the semantic information as it was pretrained for, while delegating acoustic information modeling to other Transformer modules. To extract semantic and acoustic tokens from audio data, we build `Voila-Tokenizer`, a neural audio codec (Zeghidour et al., 2021; Kumar et al., 2024; Zhang et al., 2023b). These audio tokens are added to text vocabulary for cross-modal training and knowledge sharing between modalities. In addition, `Voila` interleaves audio and text tokens during generation, leveraging the backbone LLM's text generation capabilities to guide the production of coherent voice responses.
- **Millions of pre-built voices and efficient voice creation**: `Voila` allows users to easily customize and plug in new voices for conversations. Given an audio clip of any length (from a few seconds to several hours), `Voila` learns a voice embedding that captures the unique timbre, tone, accent, and other characteristics of the speaker, allowing it to replicate the voice in conversation and speech generation. Combined with the above text instructions that define persona, users can easily create new AI characters capable of engaging in natural, interactive conversations. Thanks to the easy customizability, we were able to pre-build millions of diverse voices.
- **Unified model for various audio tasks**: In addition to spoken dialogue, `Voila` as a unified model also naturally supports a variety of audio tasks such as ASR and TTS, without requiring task-specific specialization. Besides, it can be easily extended to handle other audio tasks, such as speech translation, through simple finetuning. Trained on large multilingual text and audio data, `Voila` supports six languages: English, Chinese, French, German, Japanese, and Korean.

2 Related Work

Pipeline Systems Early voice assistant systems, such as Siri, Alexa, and Google Assistant, consists of complex multi-stage pipelines. They typically begin with wake-word detection, which constantly listen for trigger phrases (like "`Hey Siri`") to activate the assistant. Next, automatic speech recognition (ASR) converts human speech into text. Natural language understanding (NLU) then analyzes the text to determine user intent and extract relevant information. In response, natural language generation (NLG) composes an appropriate reply, which is then converted back to spoken language through text-to-speech (TTS), allowing the assistant to "speak" to the user. Recent systems integrate LLMs to simplify the pipeline and enable open-ended conversations. For example, HuggingGPT (Shen et al., 2023) and AudioGPT (Huang et al., 2023) connect ASR, LLM, and TTS. However, this multi-module approach can introduce significant delays, making it unsuitable for low-latency, real-time

applications. Additionally, converting audio to text often results in the loss of essential acoustic information such as tone and emotion (Faruqui and Hakkani-Tür, 2022; Lin et al., 2022).

End-to-end Models End-to-end audio-language models aim to overcome the above limitations (Lakhotia et al., 2021; Hassid et al., 2024; Rubenstein et al., 2023). Audio modality can be integrated into pretrained LLMs using connector modules that align audio representations with the model's input space. For instance, models in (Chu et al., 2023, 2024; Li et al., 2024; Tang et al., 2024; Shu et al., 2023) use Whisper encoder (Radford et al., 2022) to convert speech signals into embeddings. The speech and text embeddings are then concatenated as inputs to LLMs. However, Whisper encoder may introduce significant latency since it requires the full input sequence before processing, making it unsuitable for real-time streaming settings. Additionally, Tan et al. (2024) uses speech-text alignments to segment and compress speech features, matching the granularity of text tokens. However, these methods do not support speech generation and are limited to text outputs.

To support speech generation, recent methods encode continuous audio signals into discrete units (i.e., audio tokens) typically derived from self-supervised models (Hsu et al., 2021; Babu et al., 2021; Zhang et al., 2023b; Liu et al., 2024). These units are then incorporated into LLM's vocabulary, enabling the modeling of audio as a foreign language using next-token prediction. The audio tokens predicted by the LLM are then decoded back into audio signals as outputs. GSLM (Lakhotia et al., 2021), TWIST (Hassid et al., 2024), SpeechGPT (Zhang et al., 2023a) and VoxtLM (Maiti et al., 2024a) use Wav2Vec (Schneider et al., 2019) or HuBERT (Hsu et al., 2021) to tokenize continuous speech signals into learned discrete tokens. These tokens are incorporated into LLM's vocabulary for standard next-token prediction training. AudioLM (Borsos et al., 2023) proposed to combine semantic tokens with acoustic tokens from a neural audio codec (Zeghidour et al., 2021), preserving the acoustic information of the input signal and enabling the model to simulate the output of any sound, including non-speech sounds. SpeechTokenizer (Zhang et al., 2023b) distills from Hubert and unifies the semantic and acoustic tokens by disentangling different aspects of speech information hierarchically. Spectron (Nachmani et al., 2023) is trained directly on spectrograms without any quantization, which naturally preserves the input's semantic and acoustic characteristics. As Spectron predicts the spectrograms for audio output, it supports both speech and text generation, without altering the LLM vocabulary.

Using multimodal inputs combining speech and text allows the model to learn cross-modal knowledge collaboratively. Recent works (Zhang et al., 2023a; Nachmani et al., 2023; Nguyen et al., 2024; Mitsui et al., 2024) further exploit cross-modal information interaction. For example, SpeechGPT and Spectron use a Chain-of-Modality approach, guiding the model to process information textually before converting the generated text to speech. These methods fully leverage the LLM backbone's reasoning abilities acquired during text pretraining. However, this hierarchical structure of speech and text requires the model to output a complete textual response before generating the speech output, leading to increased latency that is not ideal for streaming scenarios. In contrast, Spirit-LM (Nguyen et al., 2024) and USDM (Kim et al., 2024) address the limitation by using an interleaved format of text and speech tokens, where some of the text tokens in a sequence are replaced with speech tokens. Yet, the text and speech tokens conveying the same semantics usually do not align on a token-by-token basis, making the token replacement noisy. This lack of explicit synchronization between text and speech makes the model training more difficult. On the other hand, PSLM (Mitsui et al., 2024) replaces the hierarchical speech-text structure with a parallel approach; however, it depends on an external ASR system to provide text input for the spoken audio.

Full-duplex Models End-to-end models still follow mechanical, turn-based conversation dynamics, where one party speaks while the other waits to respond. In contrast, full-duplex models allow simultaneous two-way communication that mirrors natural human interactions. These models can listen and speak at the same time, forming the foundation for autonomous interactions where the system continuously listens and actively participates by initiating speech (e.g., through backchanneling or interrupting) when appropriate. Moshi (Défossez et al., 2024) is a full-duplex speech-text model combining several ideas mentioned above,

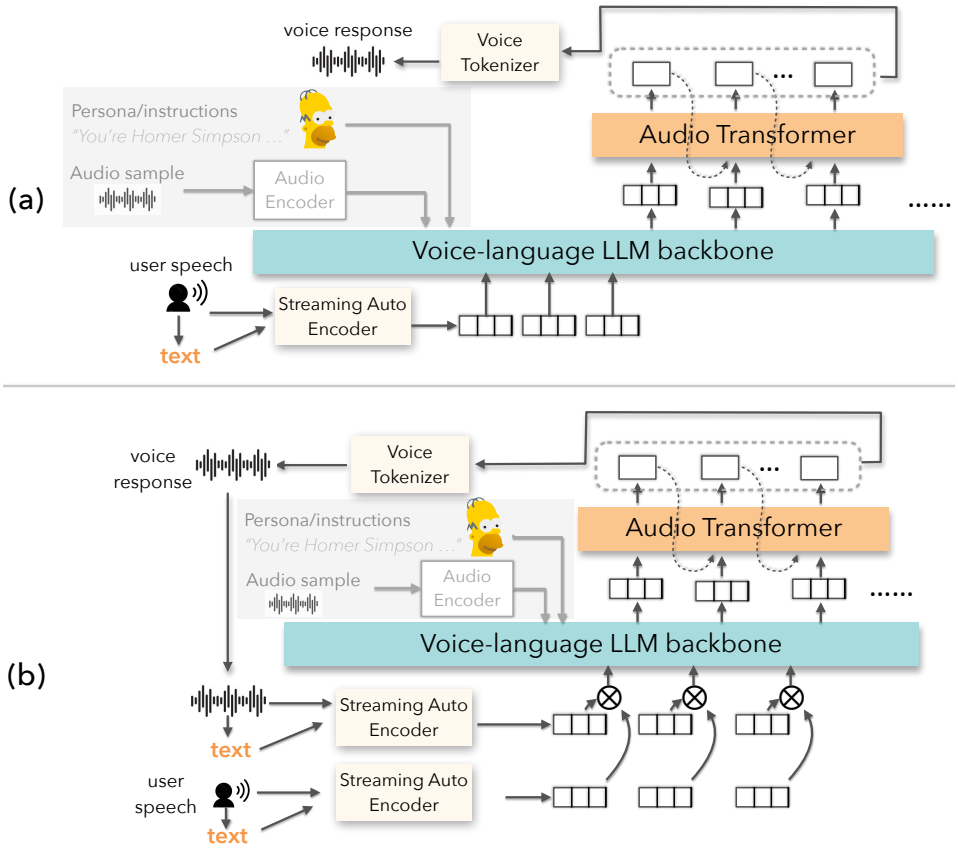

Figure 2: `Voila` models: **(a)** `Voila-e2e` for end-to-end voice conversation, **(b)** `Voila-autonomous` for autonomous interaction. Both models allow easy customization of speaker characteristics and voice via text instructions and audio samples.

including semantic and acoustic discrete audio tokens, and the interleaved structure of Spirit-LM with the parallel mode of PSLM through an Inner Monologue module. This combination enhances the factual accuracy and linguistic quality of generated speech in streaming mode. However, the Inner Monologue mechanism requires specific configurations to support different tasks, such as spoken dialogue, ASR, and TTS, making it difficult to use one single model to support all applications. For example, in ASR tasks, the audio tokens need to precede the text, whereas in TTS tasks, the text delay is adjusted so that the text appears before the audio tokens. Hertz-dev (Standard Intelligence, 2024) is a pure audio model without leveraging the power of pretrained LLMs. Their ablation studies show pretraining with text data does not bring notable improvement over training only on audio data. In comparison, `Voila-autonomous` offers several unique advantages, including better integration of LLM text capabilities with the new audio capabilities, easy customizability with text instructions and plug-and-play voice embedding, and unified modeling of spoken dialogue, ASR, TTS and other various tasks in a single model.

3 Voila: Voice-Language Foundation Models

As shown in Figure 2, `Voila` adopts a hierarchical multi-scale transformer-based architecture consisting of a voice-language LLM backbone and an audio transformer. The LLM backbone is used to model the semantic information, while the audio transformer models the audio tokens based on the semantic output of the LLM. The audio tokens generated by the audio transformer are finally decoded back to audio by the `Voila` tokenizer (§3.1). `Voila-e2e` supports end-to-end voice conversations that capture nuanced vocal information. `Voila-`

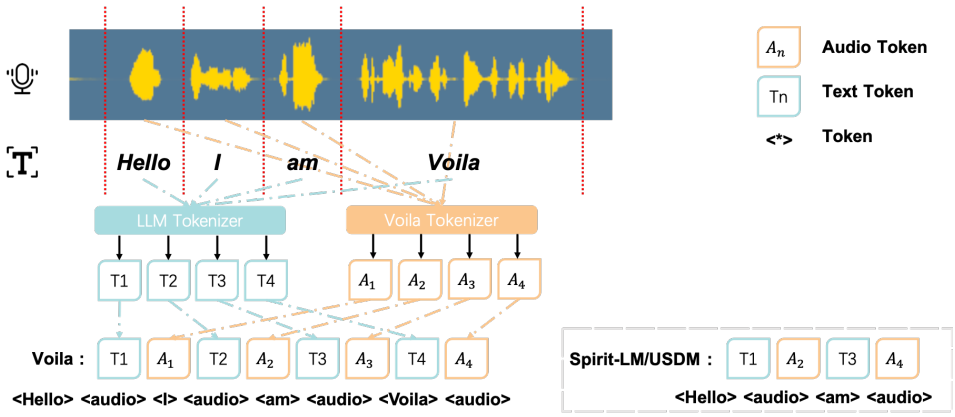

Figure 3: Text and audio interleaved alignment.

`autonomous` further extends over `Voila-e2e` and is a full-duplex model, enabling simultaneous listening, reasoning, and speaking in a two-way communication that emulates natural human interactions. We develop new text-audio alignment methods (§3.2) that take full advantages of pretrained LLMs when building the voice-language models. `Voila` models allow easy speaker customization via both text instructions and reference voice embeddings (§3.3).

3.1 Voice Tokenizer

By transforming continuous audio signals into discrete tokens, LLM can be trained/finetuned with next-token prediction to understand and generate audio. Discrete audio tokens can be classified into two categories, namely semantic tokens (Hsu et al., 2021; Schneider et al., 2019) and acoustic tokens (Zeghidour et al., 2021; Défossez et al., 2022). HuBERT (Hsu et al., 2021) obtains semantic tokens by applying K-means clustering to the activation hidden space, which shows effectiveness in capturing high-level linguistic content and supporting language modeling and resynthesis (Zhang et al., 2023a; Polyak et al., 2021). However, semantic tokens lose acoustic details such as speaker identity, intonation, and emotion, resulting in suboptimal reconstruction. In contrast, acoustic tokens generated by neural codec models (Zeghidour et al., 2021; Défossez et al., 2022) with residual vector quantization (RVQ) can effectively restore sound, but they have weak semantic dependency, making it difficult for LLM training/finetuning to converge. We rather extend the approach of Zhang et al. (2023b) by distilling semantic information into the first level of tokens with RVQ. The first level of RVQ tokens focuses on semantic information and the other three levels learn the acoustic information. We trained our tokenizer with 100K hours of audio data.

3.2 Text and Audio Alignment

Multi-task alignment. We integrate the discrete audio tokens extracted by the above `Voila` tokenizer into the vocabulary of the backbone LLM. To align the text and audio modalities, we train the model on tasks including automatic speech recognition (ASR), text-to-speech (TTS), and instruction following. These tasks are unified under a chat-style format, with next-token prediction as the training objective. For ASR, the input-output sequence is structured as `'<human> audio input <voila> text output <eos>'`, where the audio input consists of discrete audio tokens and the model generates the corresponding transcript. In TTS, the format is `'<human> text input <voila> audio output <eos>'`, with the model predicting audio tokens from a given text input. Instruction-following data is expressed in four formats: Text Input → Text Output (TITO), Text Input → Audio Output (TIAO), Audio Input → Text Output (AITO), and Audio Input → Audio Output (AIAO). For all tasks, we compute the loss only on the response portion, i.e., the tokens between `'<voila>'` and `'<eos>'`. When the instruction output involves audio (as in TIAO and AIAO), we adopt an interleaved format of text and audio tokens to improve alignment between modalities, as described below.

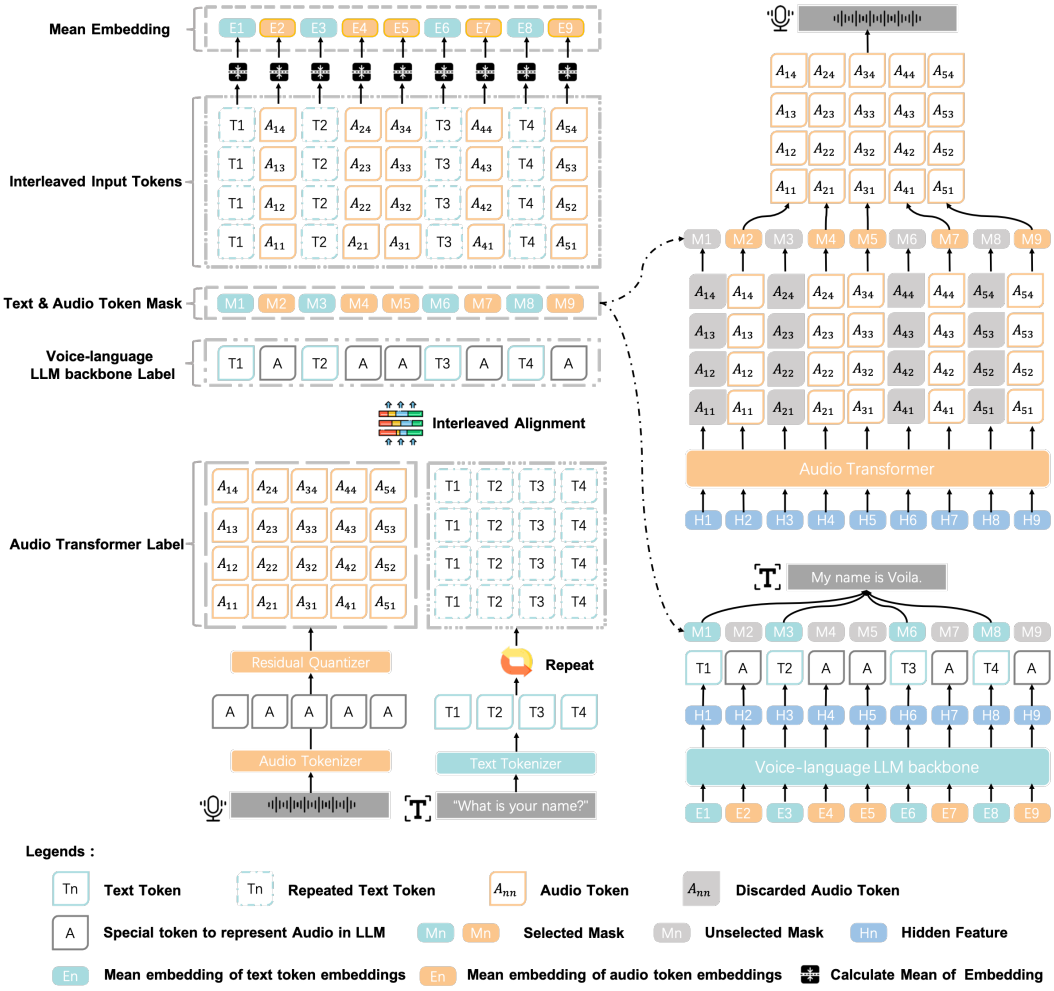

Figure 4: Input embedding and output decoding in `Voila`.

Text-audio interleaved alignment. To improve alignment between text and audio, `Voila` adopts a structured interleaved alignment strategy, where each semantic unit of text is paired with its corresponding audio tokens in an alternating sequence. For example, as illustrated in Figure 3, given the spoken input `'Hello I am Voila'`, the input sequence is encoded as `'<Hello> <audio> <I> <audio> <am> <audio> <Voila> <audio>'`, ensuring that each word is tightly aligned with its corresponding audio segment. This design facilitates fine-grained alignment and enhances the model's ability to generate expressive and synchronized speech. This design differs from prior approaches, such as Spirit-LM (Nguyen et al., 2024) and USDM (Kim et al., 2024), which also adopt interleaved text-audio formats but do so with looser coupling. These methods alternate between sequences of text and audio tokens without enforcing one-to-one alignment, often requiring the model to infer the correspondence between modalities implicitly (Figure 3). The lack of explicit alignment can hinder training stability and limit the expressiveness of generated speech.

Figure 4 illustrates the overall input-output architecture. We use audio tokens produced by the four-layer RVQ tokenizer, as described in §3.1. To maintain dimensional consistency with the audio tokens, each text token is repeated four times (bottom-left panel of the figure). The response segments of instruction-following data—which contain both text and audio—are then arranged into an interleaved pattern as described above. Tokens from this sequence are extracted and converted into embeddings (top-left panel), whose mean is computed and fed into the backbone LLM (bottom-right panel). Finally, the audio transformer takes the

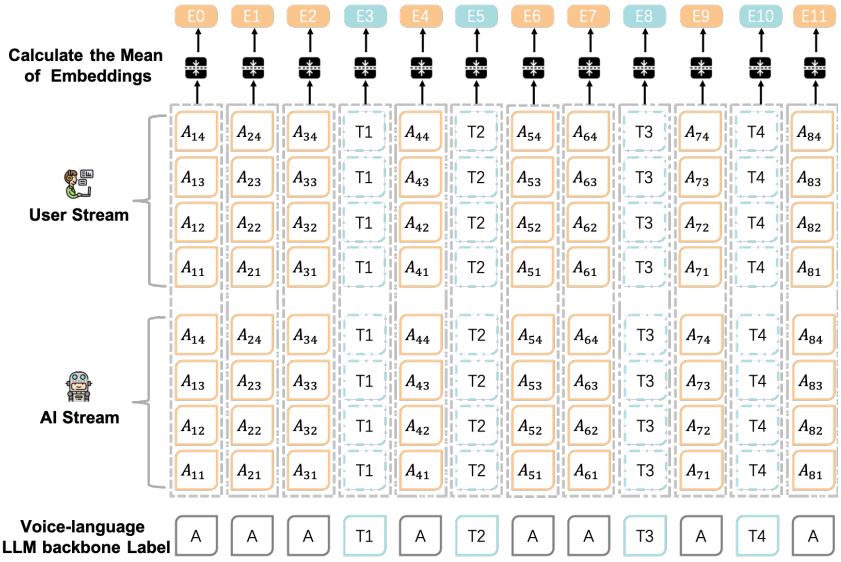

Figure 5: `Voila-autonomous` two-stream inputs, including user's audio stream and `Voila` 's own audio stream.

output from the backbone LLM as input to predict the corresponding audio tokens (top-right panel).

As shown in Figure 5, `Voila-autonomous` operates as a full-duplex model, processing both the user's audio stream and `Voila` 's own audio stream simultaneously. Each stream is independently tokenized and embedded. Once embeddings from both streams are obtained, they are fused by averaging and then passed into the backbone LLM. Finally, the audio transformer generates the `Voila` audio output by modeling the corresponding audio tokens.

3.3 One Million Pre-built Voices and Customizing New Voices

`Voila` allows users to easily customize and plug in new voices for conversations. Unlike recent pipeline-based systems that handle voice customization through separate TTS modules (Huang et al., 2025), `Voila` integrates this functionality directly into a unified, end-to-end and autonomous framework (Figure 2).

To this end, `Voila` introduces a learnable special token that captures a speaker's unique voice characteristics—including timbre, tone, and accent—via a voice embedding. During inference, this token conditions the model to synthesize speech in the desired voice. Specifically, we use Wespeaker (Wang et al., 2023b) to extract speaker embeddings from all training data with audio outputs. For training tasks involving audio generation, we prepend three special tokens to the system prompt: one each to indicate the start, reference point, and end of the voice embedding segment. The extracted speaker voice embedding is added to the embedding of the reference token, effectively conditioning the model on speaker identity. To avoid task confusion, we use different token sets for TTS and chat tasks: `<TTS_REF_START><TTS_REF><TTS_REF_END>` for TTS, and `<CHAT_REF_START><CHAT_REF><CHAT_REF_END>` for chat.

At inference time, a voice embedding can be derived from an audio clip of any length—ranging from a few seconds to several hours—using Wespeaker on the fly. This embedding is then passed to `Voila` to generate speech in the same voice. Combined with text instructions that define a persona, users can create fully customized AI characters capable of natural and expressive conversations. Thanks to `Voila` 's efficient customizability, we have pre-built over one million diverse voices, empowering users to further personalize voice profiles dynamically at inference.

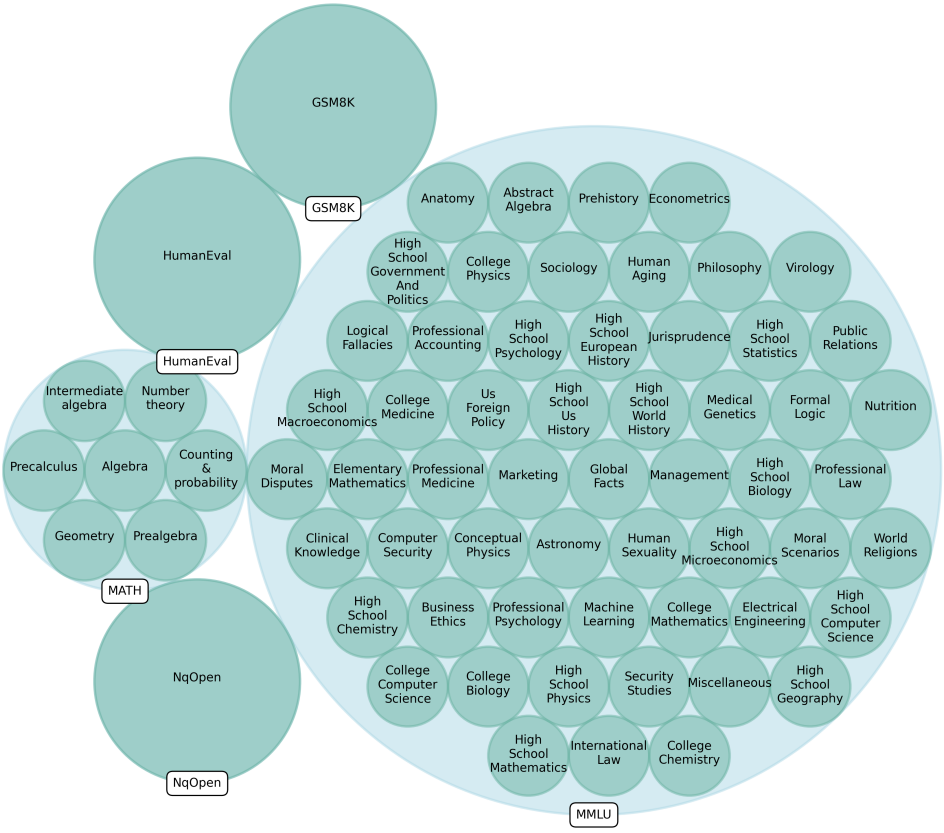

Figure 6: Domain distribution in `Voila` Benchmark

4 Experiments

4.1 Voila Benchmark

Just as recent efforts have established comprehensive benchmarks for evaluating LLMs, a thorough assessment of voice-language models requires evaluation across a wide range of domains. To this end, we introduce the `Voila` Benchmark—a new audio-language evaluation suite. The `Voila` Benchmark is constructed by sampling from five widely used LLM evaluation datasets: MMLU (Hendrycks et al., 2021a), MATH (Hendrycks et al., 2021b), OpenAI HumanEval (Chen et al., 2021), NQ-Open (Kwiatkowski et al., 2019), and GSM8K (Cobbe et al., 2021). These samples are then converted into speech using off-the-shelf TTS systems to provide broad domain coverage and realistic audio inputs. Figure 6 shows the domain distribution of the benchmark.

Specifically, for MMLU, we randomly selected 20 samples from each of its 57 diverse subjects, resulting in 1,140 data points. Similarly, for the MATH dataset, which spans 6 subjects, we selected 20 samples per subject, yielding 120 data points. For OpenAI HumanEval, NQ-Open, and GSM8K, we randomly selected 100 samples from each dataset, treating each as a distinct subject. In total, the `Voila` Benchmark comprises 66 subjects and 1,580 samples, all drawn from the test splits of their respective datasets. Since these evaluation sets originate from text-based sources, some samples—particularly those containing mathematical formulas or code—are not directly suitable for TTS synthesis. To address this, we use GPT-4o (`gpt-4o-2024-08-06`) to rewrite the text into a TTS-friendly format, which is then converted into speech using Google Cloud's TTS technology.

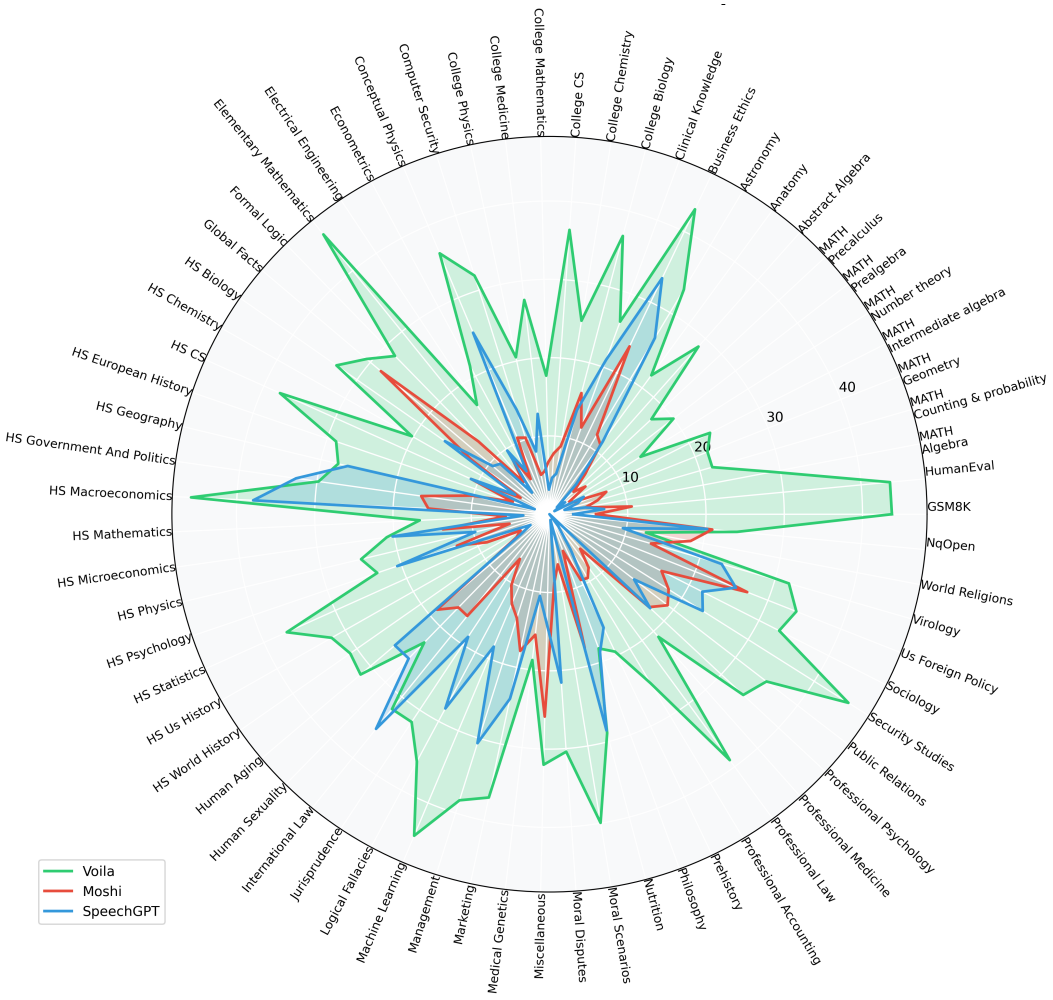

Figure 7: Performance comparison across the diverse domains in `Voila` Benchmark.

Table 1: Overall performance on `Voila` Benchmark

Model	Accuracy
SpeechGPT (7B) (Zhang et al., 2023a)	13.29
Moshi (Défossez et al., 2024)	11.45
Voila	**30.56**

4.2 Evaluation on `Voila` Benchmark

To evaluate the correctness of audio outputs from voice-language models, we first transcribe the generated speech using the Whisper system Radford et al. (2022). We then use GPT-4o to assess the transcribed responses. For each test case, GPT-4o is provided with the question and the reference answer as ground truth, and it assigns a score to the model's output on a scale from 0 to 100 based on its alignment with the reference answer.

We compare our results against two recent open-source audio-language models: SpeechGPT (Zhang et al., 2023a) and Moshi (Défossez et al., 2024). Table 1 presents the average scores of all models on the `Voila` Benchmark. Detailed performance across the benchmark's diverse fine-grained domains is shown in Figure 7 and Table 2. `Voila` outperforms both SpeechGPT and Moshi, establishing a strong baseline on this benchmark. Notably, `Voila` demonstrates

significant improvements in the math and code domains, highlighting `Voila`'s text-audio alignment takes effective advantage of the reasoning capabilities of the backbone LLM.

Table 2: Detailed results across the diverse domains in `Voila` Benchmark.

Domain	**SpeechGPT**	**Moshi**	**Voila**
MMLU-High School Microeconomics	20.25	13.75	**21.00**
MMLU-High School World History	4.45	5.15	**31.05**
MMLU-High School Statistics	4.75	8.75	**36.80**
MMLU-High School Biology	16.30	12.25	**33.20**
MMLU-High School European History	2.00	4.95	**28.80**
MMLU-High School US History	2.80	4.25	**32.00**
MMLU-High School Psychology	20.55	12.50	**23.25**
MMLU-High School Government and Politics	32.70	16.55	**29.80**
MMLU-High School Macroeconomics	37.90	15.50	**45.85**
MMLU-High School Chemistry	5.05	4.25	**20.20**
MMLU-High School Mathematics	3.35	5.60	**16.60**
MMLU-High School Computer Science	11.00	11.00	**37.75**
MMLU-High School Geography	26.50	10.25	**27.75**
MMLU-Medical Genetics	10.50	15.50	**18.75**
MMLU-Jurisprudence	18.80	6.85	**31.85**
MMLU-Formal Logic	9.00	13.00	**28.25**
MMLU-Management	30.65	13.90	**38.30**
MMLU-Global Facts	9.50	28.25	**30.65**
MMLU-Human Aging	25.90	18.85	**31.65**
MMLU-Philosophy	**18.75**	5.00	18.25
MMLU-Sociology	21.90	16.05	**32.90**
MMLU-Marketing	24.10	17.85	**37.00**
MMLU-Human Sexuality	22.90	16.65	**25.75**
MMLU-US Foreign Policy	25.60	27.15	**33.90**
MMLU-Astronomy	26.25	11.85	**33.45**
MMLU-Moral Scenarios	0.75	6.50	**39.95**
MMLU-Professional Law	1.90	8.45	**38.95**
MMLU-College Mathematics	3.15	6.65	**17.70**
MMLU-Machine Learning	18.35	12.40	**44.55**
MMLU-Electrical Engineering	9.75	8.85	**16.75**
MMLU-Conceptual Physics	25.15	10.60	**36.15**
MMLU-Professional Medicine	0.00	5.90	**20.95**
MMLU-Professional Psychology	17.60	17.40	**33.85**
MMLU-Professional Accounting	5.25	9.50	**25.15**
MMLU-College Physics	8.20	5.15	**20.50**
MMLU-College Medicine	12.90	5.50	**27.55**
MMLU-Computer Security	14.00	10.25	**31.95**
MMLU-Security Studies	23.15	18.00	**45.15**
MMLU-High School Physics	9.75	5.25	**24.75**
MMLU-Business Ethics	33.40	23.75	**43.15**
MMLU-Miscellaneous	14.05	25.85	**32.00**
MMLU-College Biology	13.75	16.00	**36.75**
MMLU-Elementary Mathematics	3.55	5.30	**45.95**
MMLU-Prehistory	16.05	9.30	**19.45**

MMLU-Logical Fallacies	28.15	10.30	**35.85**
MMLU-Abstract Algebra	5.15	6.85	**28.65**
MMLU-Econometrics	6.25	5.10	**21.50**
MMLU-Clinical Knowledge	21.00	11.80	**26.20**
MMLU Anatomy	11.50	11.00	**22.00**
MMLU-College Computer Science	4.65	7.80	**36.40**
MMLU-Moral Disputes	21.55	8.75	**30.40**
MMLU-International Law	**35.25**	16.70	32.00
MMLU-Public Relations	13.65	19.05	**35.00**
MMLU-Virology	22.90	13.50	**31.85**
MMLU-College Chemistry	5.25	8.80	**25.05**
MMLU-Nutrition	28.55	17.65	**28.90**
MMLU-World Religions	9.55	**18.35**	12.55
MATH-Intermediate Algebra	4.90	3.50	**23.00**
MATH-Algebra	4.65	4.70	**29.00**
MATH-Geometry	3.90	7.80	**21.70**
MATH-Number Theory	0.75	3.00	**13.75**
MATH-Precalculus	1.30	4.20	**17.80**
MATH-Counting Probability	2.00	6.40	**21.65**
MATH-Prealgebra	2.50	5.80	**20.00**
OpenAI HumanEval	7.01	10.55	**43.70**
NQ Open	20.24	20.89	**24.07**
GSM8K	3.05	5.96	**43.76**

Table 3: Results of ASR.

Model	LibriSpeech test-clean (WER ↓)
Whisper large v2 (Radford et al., 2022)	2.7
Whisper large v3 (Radford et al., 2022)	2.2
FastConformer (Rekesh et al., 2023)(w/ LibriSpeech train split)	3.6
VoxtLM (Maiti et al., 2024b)(w/ LibriSpeech train split)	2.7
Moshi (Défossez et al., 2024)	5.7
Voila (w/o LibriSpeech train split)	**4.8**
Voila (w/ LibriSpeech train split)	**2.7**

Table 4: Results of TTS.

Model	LibriSpeech test-clean (WER ↓)
YourTTS (Casanova et al., 2022)	7.7
Vall-E (Wang et al., 2023a)	5.9
Moshi (Défossez et al., 2024)	4.7
Voila (w/o LibriSpeech train split)	**3.2**
Voila (w/ LibriSpeech train split)	**2.8**

4.3 Evaluation on ASR and TTS

Metrics. `Voila` supports not only spoken dialogue but also automatic speech recognition (ASR) and text-to-speech (TTS). We evaluate both components independently. For ASR, we measure performance on the LibriSpeech test-clean dataset (Panayotov et al., 2015), using word error rate (WER) as the evaluation metric. For TTS, we follow the protocol

from Vall-E (Wang et al., 2023a), transcribing the generated audio with HuBERT-Large (Hsu et al., 2021). Following the Vall-E setup, we evaluate audio samples between 4 and 10 seconds in length. We report results under two settings: (1) when LibriSpeech data is excluded during training, and (2) when the LibriSpeech training split is included.

Baselines. We compare our ASR results with prior work, including Radford et al. (2022); Rekesh et al. (2023); Maiti et al. (2024b); Défossez et al. (2024). For TTS, we compare with Wang et al. (2023a); Défossez et al. (2024).

Results. The results in Table 3 show that `Voila` performs competitively with state-of-the-art ASR models such as Whisper (Radford et al., 2022), outperforming models like FastConformer (Rekesh et al., 2023). When compared to recent speech-language models (Maiti et al., 2024b; Défossez et al., 2024), `Voila` achieves a lower WER of 4.8%, surpassing the 5.7% reported in Défossez et al. (2024). In the setting where LibriSpeech training data is used, `Voila` reaches a WER of 2.7%, matching the best result reported by Maiti et al. (2024b). As shown in Table 4, `Voila` also outperforms in TTS, achieving a WER of 3.2% (2.8% with LibriSpeech training data), compared to 4.7% from Défossez et al. (2024).

5 Conclusion

We introduced `Voila`, a family of voice-language foundation models that support spoken dialogue, ASR, TTS, and other voice-language tasks in an end-to-end, autonomous manner. Through innovations in voice tokenization, hierarchical modeling, and audio-text alignment, `Voila` achieves performance comparable to or exceeding state-of-the-art models. Built on a unique multi-scale Transformer architecture, `Voila` is designed to tightly integrate voice and language capabilities, enabling fine-grained processing of both semantic and acoustic signals and fully leveraging the strengths of large language models. A key feature of `Voila` is its support for extensive customization, allowing users to create diverse and expressive voice personas that enrich interaction quality. Overall, `Voila` represents a step toward autonomous voice AI that acts as a proactive and empathetic partner in human activities. We release the models and code to support further research and development in this space.

References

A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

C. Bérubé, M. Nißen, R. Vinay, A. Geiger, T. Budig, A. Bhandari, C. R. P. Benito, N. Ibarcena, O. Pistolese, P. Li, et al. Proactive behavior in voice assistants: A systematic review and conceptual model. *Computers in Human Behavior Reports*, page 100411, 2024.

S. Bora. Breaking the silence: How voice AI is shaping the future of human-machine interactions, 2024. Medium, accessed November 8, 2024.

Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

S. Buss. Autonomous action: Self-determination in the passive mode. *Ethics*, 122(4):647–691, 2012.

S. Buyukgoz, J. Grosinger, M. Chetouani, and A. Saffiotti. Two ways to make your robot proactive: Reasoning about human intentions or reasoning about possible futures. *Frontiers in Robotics and AI*, 9:929267, 2022.

E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR, 2022.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code, 2021.

Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

A. Défossez, J. Copet, G. Synnaeve, and Y. Adi. High fidelity neural audio compression, 2022. URL `https://arxiv.org/abs/2210.13438`.

M. Faruqui and D. Hakkani-Tür. Revisiting the boundary between asr and nlu in the age of conversational dialog systems. *Computational Linguistics*, 48(1):221–232, 2022.

J. L. Flanagan. *Speech analysis synthesis and perception*. Springer Berlin, Heidelberg, 1972.

J. Grosinger. On proactive human-ai systems. In *AIC*, pages 140–146, 2022.

S. Hao, Y. Gu, H. Ma, J. Hong, Z. Wang, D. Wang, and Z. Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, 2023a.

S. Hao, T. Liu, Z. Wang, and Z. Hu. ToolkenGPT: Augmenting frozen language models with massive tools via tool embeddings. *arXiv preprint arXiv:2305.11554*, 2023b.

M. Hassid, T. Remez, T. A. Nguyen, I. Gat, A. Conneau, F. Kreuk, J. Copet, A. Defossez, G. Synnaeve, E. Dupoux, et al. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36, 2024.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021b.

N. Hoke. Digital assistants vs digital companions: What's the difference?, 2021. URL `https://blog.intuitionrobotics.com/digital-assistants-vs-digital-companions-whats-the-difference`. Blog post.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021. URL `https://arxiv.org/abs/2106.07447`.

A. Huang, B. Wu, B. Wang, C. Yan, C. Hu, C. Feng, F. Tian, F. Shen, J. Li, M. Chen, P. Liu, R. Miao, W. You, X. Chen, X. Yang, Y. Huang, Y. Zhang, Z. Gong, Z. Zhang, H. Zhou, J. Sun, B. Li, C. Feng, C. Wan, H. Hu, J. Wu, J. Zhen, R. Ming, S. Yuan, X. Zhang, Y. Zhou, B. Li, B. Ma, H. Wang, K. An, W. Ji, W. Li, X. Wen, X. Kong, Y. Ma, Y. Liang, Y. Mou, B. Ahmidi, B. Wang, B. Li, C. Miao, C. Xu, C. Wang, D. Shi, D. Sun, D. Hu, D. Sai, E. Liu, G. Huang, G. Yan, H. Wang, H. Jia, H. Zhang, J. Gong, J. Guo, J. Liu, J. Liu, J. Feng, J. Wu, J. Wu, J. Yang, J. Wang, J. Zhang, J. Lin, K. Li, L. Xia, L. Zhou, L. Zhao, L. Gu, M. Chen, M. Wu, M. Li, M. Li, M. Li, M. Liang, N. Wang, N. Hao, Q. Wu, Q. Tan, R. Sun, S. Shuai, S. Pang, S. Yang, S. Gao, S. Yuan, S. Liu, S. Deng, S. Jiang, S. Liu, T. Cao, T. Wang, W. Deng, W. Xie, W. Ming, W. He, W. Sun, X. Han, X. Huang, X. Deng, X. Liu, X. Wu, X. Zhao, Y. Wei, Y. Yu, Y. Cao, Y. Li, Y. Ma, Y. Xu, Y. Wang, Y. Shi, Y. Wang, Y. Zhou, Y. Zhong, Y. Zhang, Y. Wei, Y. Luo, Y. Lu, Y. Yin, Y. Luo, Y. Ding, Y. Yan, Y. Dai, Y. Yang, Z. Xie, Z. Ge, Z. Sun, Z. Huang, Z. Chang, Z. Guan, Z. Yang, Z. Zhang, B. Jiao, D. Jiang, H.-Y. Shum, J. Chen, J. Li, S. Zhou, X. Zhang, X. Zhang, and Y. Zhu. Step-audio: Unified understanding and generation in intelligent speech interaction, 2025. URL `https://arxiv.org/abs/2502.11946`.

R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu, Y. Ren, Z. Zhao, and S. Watanabe. Audiogpt: Understanding and generating speech, music, sound, and talking head, 2023. URL `https://arxiv.org/abs/2304.12995`.

A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

H. Kim, S. Seo, K. Jeong, O. Kwon, S. Kim, J. Kim, J. Lee, E. Song, M. Oh, J.-W. Ha, S. Yoon, and K. M. Yoo. Integrating paralinguistics in speech-empowered large language models for natural conversation, 2024. URL `https://arxiv.org/abs/2402.05706`.

R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36, 2024.

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466, 2019. doi: 10.1162/tacl_a_00276. URL `https://doi.org/10.1162/tacl_a_00276`.

K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux. Generative spoken language modeling from raw audio, 2021. URL `https://arxiv.org/abs/2102.01192`.

M. Lebourdais, M. Tahon, A. Laurent, and S. Meignier. Automatic speech interruption detection: Analysis, corpus, and system. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-Coling 2024)*, pages à–paraître, 2024.

Y. Li, H. Sun, M. Lin, T. Li, G. Dong, T. Zhang, B. Ding, W. Song, Z. Cheng, Y. Huo, S. Chen, X. Li, D. Pan, S. Zhang, X. Wu, Z. Liang, J. Liu, T. Zhang, K. Lu, Y. Zhao, Y. Shen, F. Yang, K. Yu, T. Lin, J. Xu, Z. Zhou, and W. Chen. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024.

T.-E. Lin, Y. Wu, F. Huang, L. Si, J. Sun, and Y. Li. Duplex conversation: Towards human-like interaction in spoken dialogue systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3299–3308, 2022.

A. H. Liu, H.-J. Chang, M. Auli, W.-N. Hsu, and J. Glass. Dinosr: Self-distillation and online clustering for self-supervised speech representation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

S. Maiti, Y. Peng, S. Choi, J.-w. Jung, X. Chang, and S. Watanabe. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13326–13330. IEEE, 2024a.

S. Maiti, Y. Peng, S. Choi, J. weon Jung, X. Chang, and S. Watanabe. Voxtlm: unified decoder-only models for consolidating speech recognition/synthesis and speech/text continuation tasks, 2024b. URL https://arxiv.org/abs/2309.07937.

A. S. Meyer. Timing in conversation. *Journal of Cognition*, 6(1), 2023.

K. Mitsui, K. Mitsuda, T. Wakatsuki, Y. Hono, and K. Sawada. Pslm: Parallel generation of text and speech with llms for low-latency spoken dialogue systems. *arXiv preprint arXiv:2406.12428*, 2024.

E. Nachmani, A. Levkovitch, R. Hirsch, J. Salazar, C. Asawaroengchai, S. Mariooryad, E. Rivlin, R. Skerry-Ryan, and M. T. Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. *arXiv preprint arXiv:2305.15255*, 2023.

T. A. Nguyen, B. Muller, B. Yu, M. R. Costa-Jussa, M. Elbayad, S. Popuri, P.-A. Duquenne, R. Algayres, R. Mavlyutov, I. Gat, et al. Spirit-lm: Interleaved spoken and written language model. *arXiv preprint arXiv:2402.05755*, 2024.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

V. Panayotov, G. Chen, D. Povey, and S. 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.

A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux. Speech resynthesis from discrete disentangled self-supervised representations, 2021. URL https://arxiv.org/abs/2104.00355.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam, et al. Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. de Chaumont Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, H. Muckenhirn, D. Padfield, J. Qin, D. Rozenberg, T. Sainath, J. Schalkwyk, M. Sharifi, M. T. Ramanovich, M. Tagliasacchi, A. Tudor, M. Velimirović, D. Vincent, J. Yu, Y. Wang, V. Zayats, N. Zeghidour, Y. Zhang, Z. Zhang, L. Zilka, and C. Frank. Audiopalm: A large language model that can speak and listen, 2023. URL https://arxiv.org/abs/2306.12925.

R. Schafer. Scientific bases of human-machine communication by voice. *Proceedings of the National Academy of Sciences of the United States of America*, 92(22):9914–9920, 1995.

S. Schneider, A. Baevski, R. Collobert, and M. Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

J. Schroeder and N. Epley. Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General*, 145(11):1427, 2016.

Y. Shao, L. Li, J. Dai, and X. Qiu. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*, 2023.

Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face, 2023. URL https://arxiv.org/abs/2303.17580.

Y. Shu, S. Dong, G. Chen, W. Huang, R. Zhang, D. Shi, Q. Xiang, and Y. Shi. Llasm: Large language and speech model, 2023.

G. Skantze. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178, 2021.

Standard Intelligence. Introducing hertz-dev, the first open-source base model for conversational audio generation, 2024. URL https://si.inc/hertz-dev/. Blog post.

T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon, et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26): 10587–10592, 2009.

W. Tan, H. Inaguma, N. Dong, P. Tomasello, and X. Ma. Ssr: Alignment-aware modality connector for speech language models. *arXiv preprint arXiv:2410.00168*, 2024.

C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024. URL https://arxiv.org/abs/2310.13289.

C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023a.

H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023b.

Z. M. Wang, Z. Peng, H. Que, J. Liu, W. Zhou, Y. Wu, H. Guo, R. Gan, Z. Ni, J. Yang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023c.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

W. Wu, H. Wu, L. Jiang, X. Liu, J. Hong, H. Zhao, and M. Zhang. From role-play to drama-interaction: An llm solution. *arXiv preprint arXiv:2405.14231*, 2024.

D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, X. Wu, Z. Zhao, S. Watanabe, and H. Meng. Uniaudio: An audio foundation model toward universal audio generation, 2023. URL https://arxiv.org/abs/2310.00704.

L. Yang, C. Achard, and C. Pelachaud. Multimodal analysis of interruptions. In *International Conference on Human-Computer Interaction*, pages 306–325. Springer, 2022.

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023a.

X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*, 2023b.

Y. Zhu, D. Su, L. He, L. Xu, and D. Yu. Generative pre-trained speech language model with efficient hierarchical transformer, 2024. URL https://arxiv.org/abs/2406.00976.