Moshi: a speech-text foundation model for real-time dialogue

13th of March 2025
Conversational AI Reading Group
Alexandre Défossez

kyutai

OPEN-SCIENCE AI LAB

Meet the team behind Moshi








And our donors



About Kyutai

Non-profit lab in Paris with focus on open science and open source.

We released **Moshi** last July, then open source last September.

Published Helium-2B, **multi-lingua**l foundation text model in January.

Initial focus on multimodal LLM, but wider interest in any **core-ml** research.

Open technology, train people.

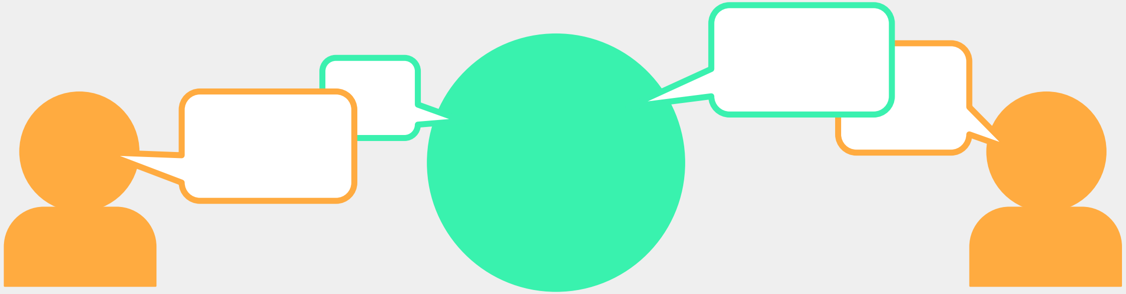
Building a speech AI assistant

Communicating is more than text

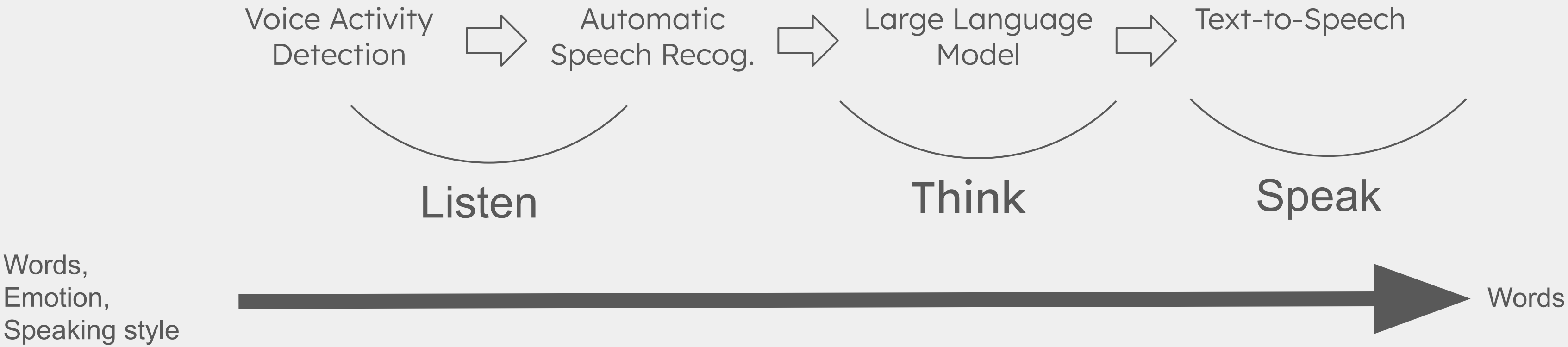
We listen, think and speak almost instantly

Natural flow, interruptions

Paralinguistic communication (emotion, tone, etc.)



Limits of cascading models

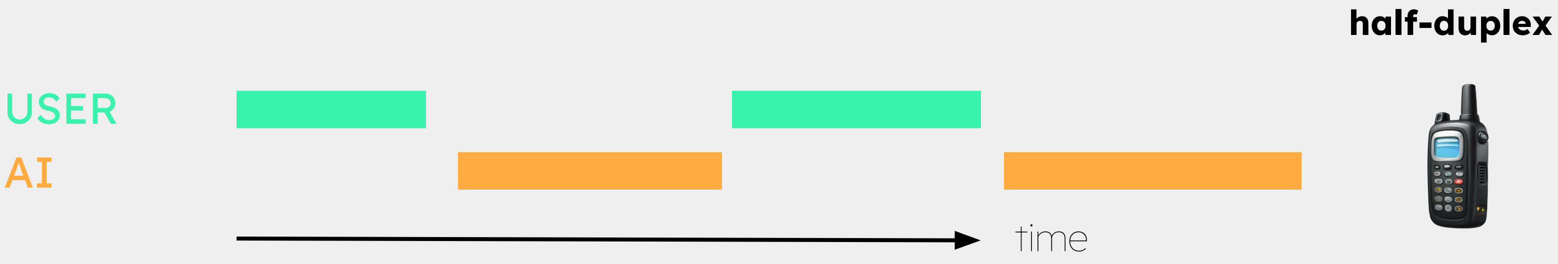


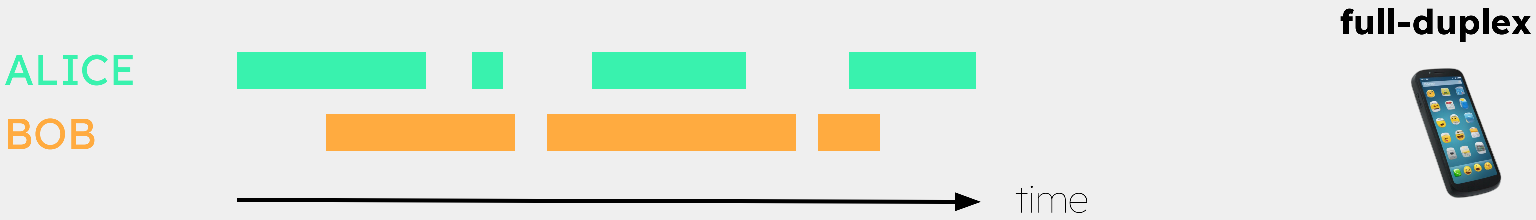
3 to 5 seconds

How can we merge all these steps into a single audio language model?

Half-duplex vs. full-duplex

Existing models are mostly half-duplex, no overlap between speakers.





Neural audio codecs

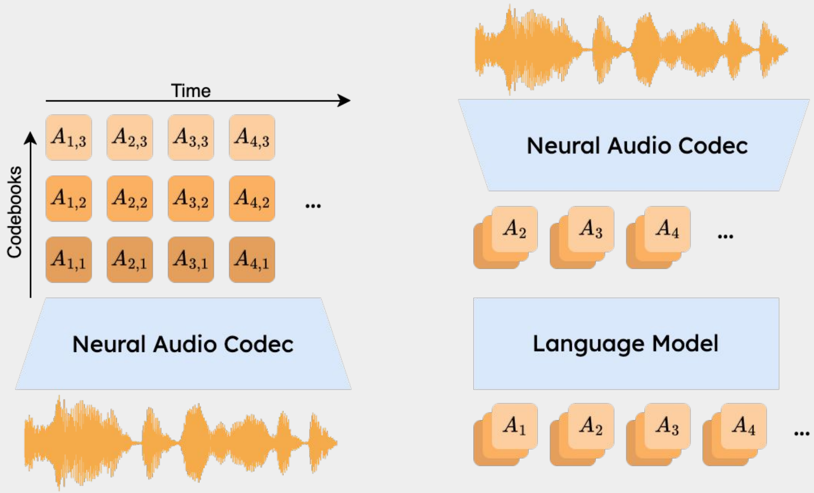
Components of a speech LM

Audio tokenisation with a **neural audio codec**.

Frame rate (12.5 Hz) higher than text (3 Hz).

Each step composed of **8 or more** discrete tokens.

More to do: handling interruptions, multi-turns.



SoundStream: an end-to-end audio codec, Zeghidour et al. IEEE Trans. 2021.
High Fidelity neural audio compression, Défossez et al. TMLR 2022.
SpeechTokenizer: Unified speech tokenizer for speech language models, Zhang et al. ICLR 2024.

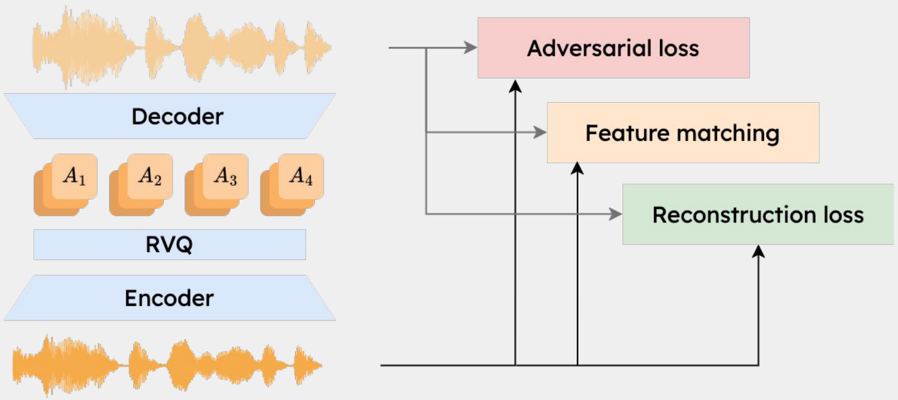
Neural audio codecs

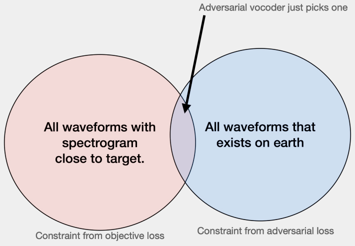
Based on SoundStream [Zeghidour et al. 2021] and EnCodec [Défossez et al. 2022].

Autoencoder with a reconstruction + **adversarial loss**.

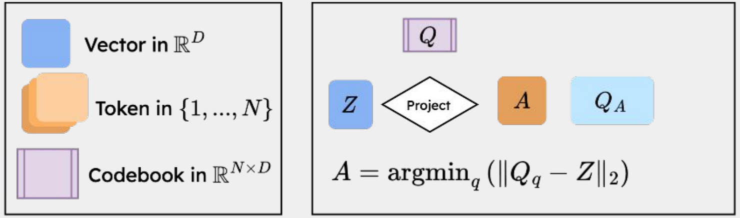
Uses **Residual Vector Quantization** as an information bottleneck, as introduced by Soundstream. Gives **acoustic tokens**.

Can include a **semantic token** distilled from a self-supervised model [Zhang et al. 2024].



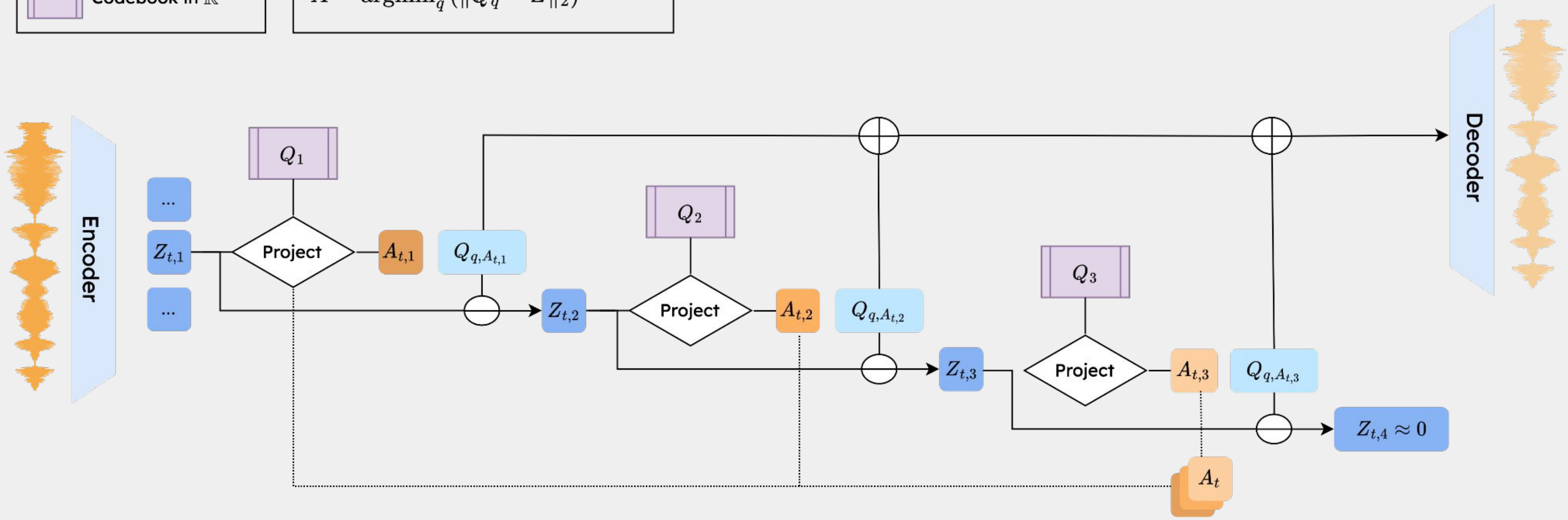


Residual Vector Quantization



First introduced by SoundStream.

For each vector, multiple discrete tokens, from **coarse to fine.**



Mimi: high quality, low framerate

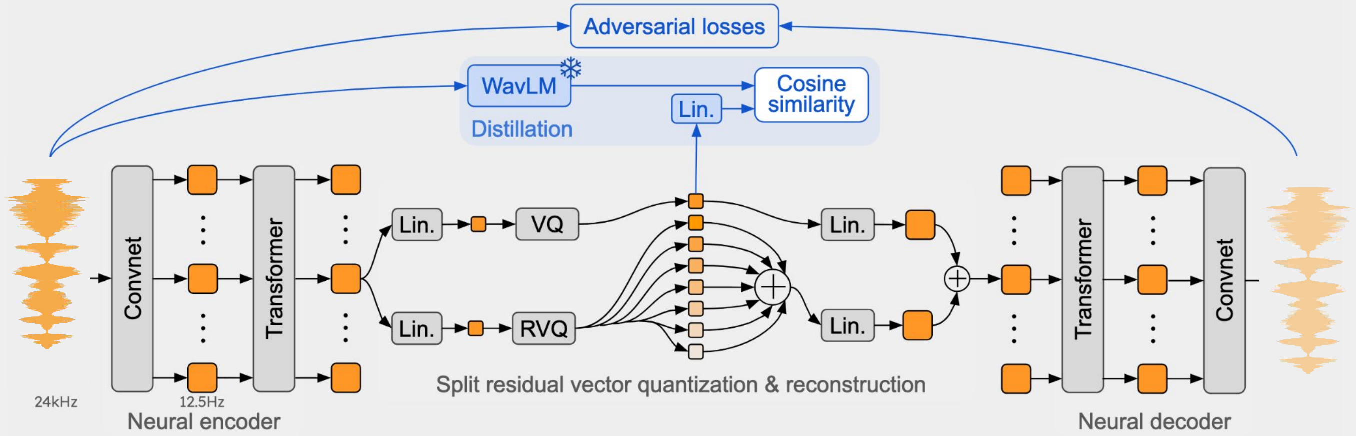
Existing models produce tokens at 50 Hz. Most are **non causal**.

This means at least 50 **auto-regressive steps** per seconds.

Mimi improves on **semantic distillation** for the semantic token.

Operates causally at **12.5Hz**, e.g. by chunks of **80ms** of audio.

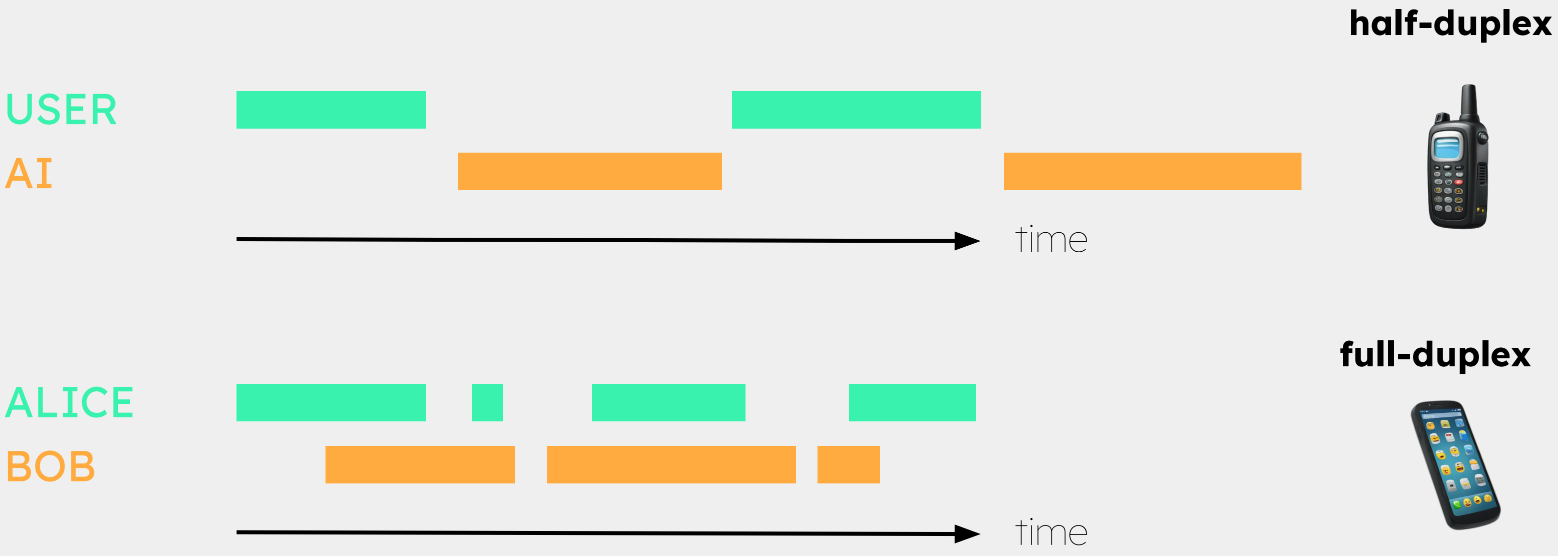
Uses only an adversarial and feature matching loss.



Joint sequence modeling

Half-duplex vs. full-duplex

Existing models are mostly half-duplex, no overlap between speakers.



An early stage prototype

Autoregressive Image Generation using Residual Quantization, Lee et al. 2022.
UniAudio: An Audio Foundation Model Toward Universal Audio Generation, Yang et al. ICLR2024
Simple and Controllable Music Generation, Copet et al. Neurips 2023

Joint sequence modeling

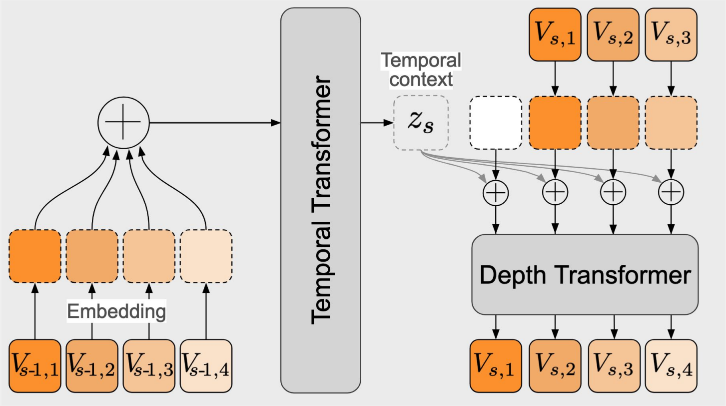
For each chunk of 80ms of audio we get **multiple tokens**: 1 semantic, 7 acoustics.

How to fit that in the **next token prediction** framework?

Flatten? Then 8x more steps...

Or uses **small Transformer** along *depth* dimension: RQ-Transformer, done for image [Lee et al. 2022], and audio [Yang et al. 2024].

We bring improvements: acoustic delay (from [Copet et al. 2023]), per codebook parameters.



Adding more streams

Using **text as foundation** for audio generation explored in Spectron [1], SpeechGPT [2], SpiritLM [3], PSLM [4].

Either full text prefix [1, 2], one modality at a time [3], or single turn [4].

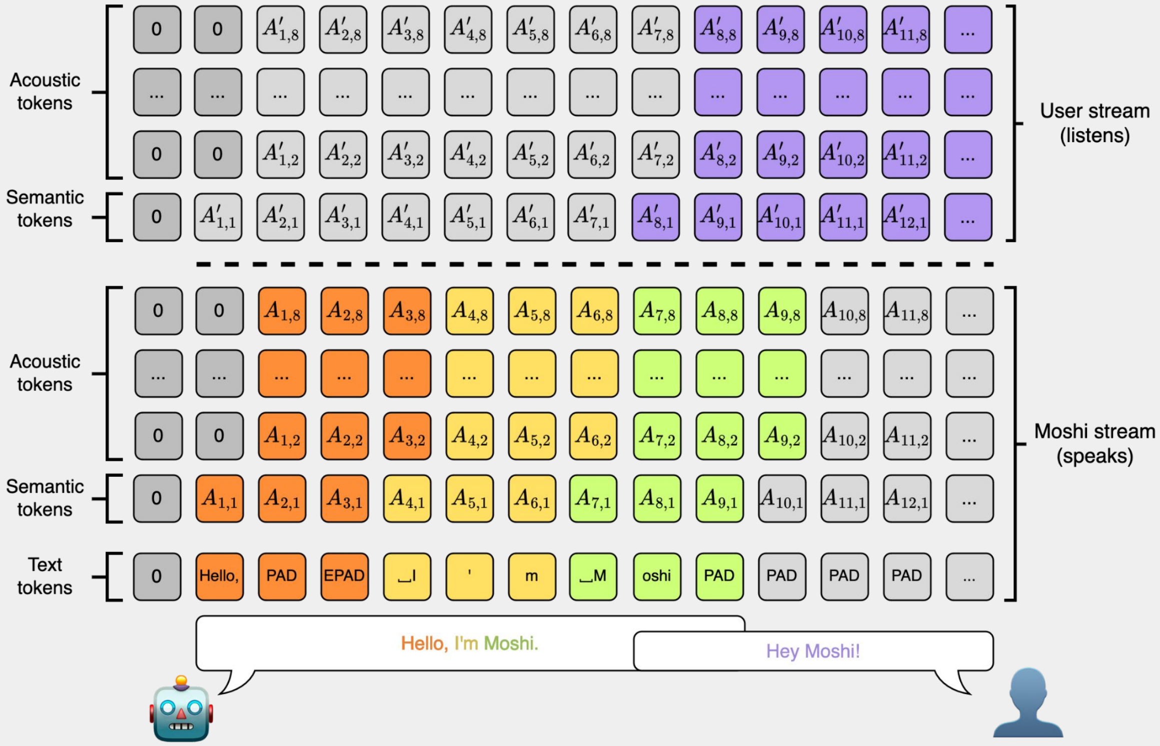
To handle interruptions, reduce latency, we want text **aligned closely** with audio.

New text stream from **word-aligned** transcriptions: the *inner monologue*.

Uses special PAD and EPAD tokens to fill in the blanks.

To handle user vs. Moshi: add second set of audio tokens for the **user's audio stream**.

All streams at once



Training

Training stages

First train 7B **text-only model**, Helium, for 500k steps.

Warm init from Helium, trained on **unlabeled audio data** + text for 1M steps.

Adds second audio stream, use diarization to emulate multi-stream audio.

Finally, fine tune on real two-speaker audio with separate streams, including **synthetic data**.

Hyper-parameter	Helium training pre-training	Moshi training pre-training	post-training	fisher	fine
Temporal Transformer					
Model dimension	4096	same			
MLP dimension	11264				
Number of heads	32				
Number of layers	32				
Context size	4096	3000 steps, e.g. 4 min.			
Learning rate	$3 \cdot 10^{-4}$	$3 \cdot 10^{-5}$	$3 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$2 \cdot 10^{-6}$
Depth Transformer					
Model dimension	-	1024			
MLP dimension	-	4096			
Number of heads	-	16			
Number of layers	-	6			
Learning rate	-	$2 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	$4 \cdot 10^{-6}$	$2 \cdot 10^{-6}$
Input / Output space					
Text cardinality	32000	32000			
Audio cardinality	-	2048			
Frame rate	-	12.5 Hz			
Common parameters					
Batch size (text)	4.2M tok.	1.2M tok.	1.2M tok.	-	-
Batch size (audio)	-	16h	8h	40min	2.7h
Training steps	500k	1M	100k	10k	30k
LR Schedule	cosine	cosine	-	-	-
Acoustic delay	-	2	1	1	1
Text delay	-	± 0.6	0	0	0

Synthetic data

Existing datasets (Open Hermes) too specific to text.

Fine tune Helium on transcripts of dialogs.

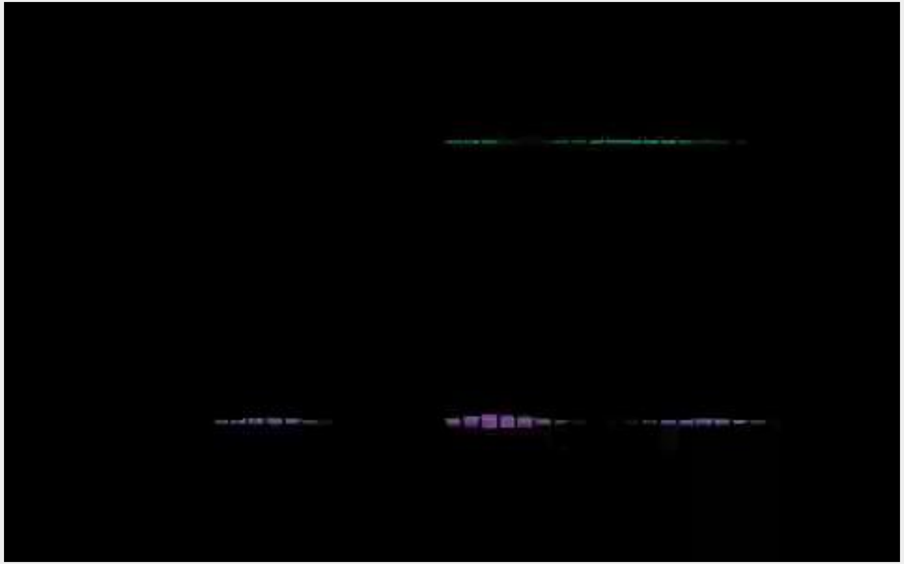
Use it to generate interaction scripts.

Leverage our **multi-stream TTS engine** to synthesize 20k hours of instruct data.

No RLHF for now.

```
{{context}}

Write the transcript of a conversation between Blake and Moshi.
{{summary}} Moshi is knowledgeable about the topic.  Use some
backchanneling.  Use short turns.
```



Results

Mimi codec

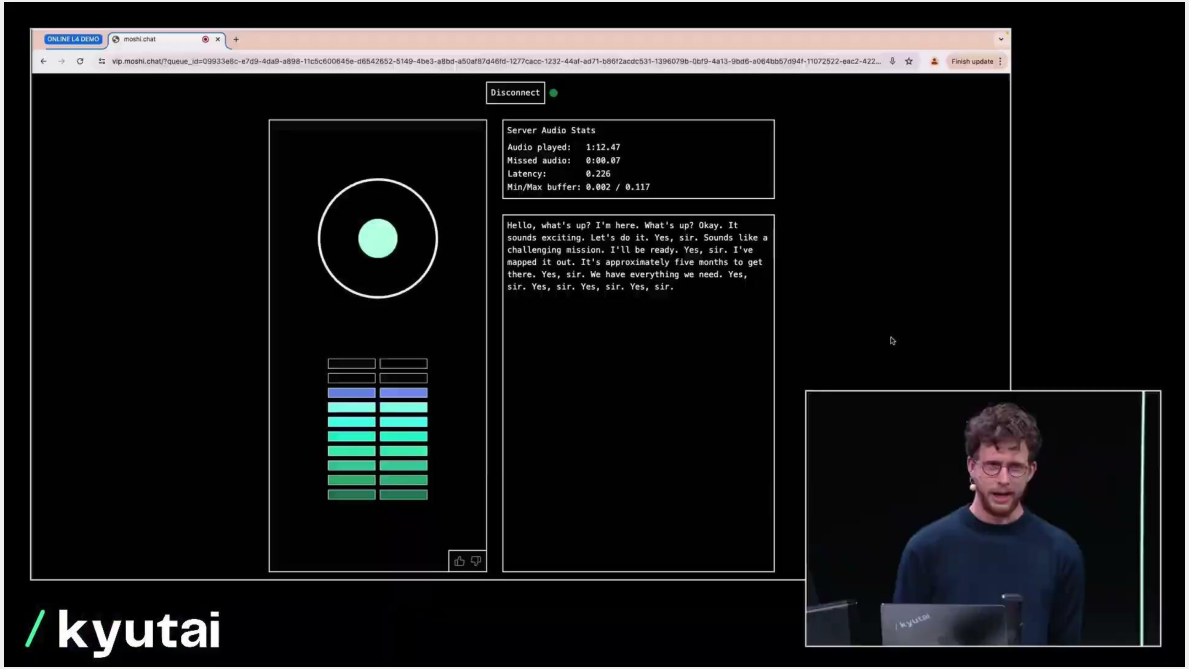
Model	f_s	f_r	bitrate	causal	ABX (↓)	VisQOL (↑)	MOSNet (↑)	MUSHRA (↑)
Ground Truth	24kHz	-	-	-	-	-	3.08	90.6±1.0
RVQGAN	24kHz	75Hz	1.5kbps		-	1.74	2.74	31.3±1.3
SemantiCodec	16kHz	50Hz	1.3kbps		42.2%	2.43	3.12	64.8±1.5
SpeechTokenizer	16kHz	50Hz	1.5kbps		**3.3%**	1.53	2.67	45.1±1.5
SpeechTokenizer	16kHz	50Hz	4.0kbps		**3.3%**	3.07	3.10	74.3±1.5
Mimi, adv. loss only	24kHz	12.5Hz	1.1kbps	✓	8.7%	1.84	3.10	**81.0**±1.3
Same, downsampled at 16kHz	16kHz	12.5Hz	1.1kbps	✓	-	-	-	77.7±1.4
Mimi, non adv. only	24kHz	12.5Hz	1.1kbps	✓	8.1%	2.82	2.89	58.8±1.8

Spoken QA: the modality gap

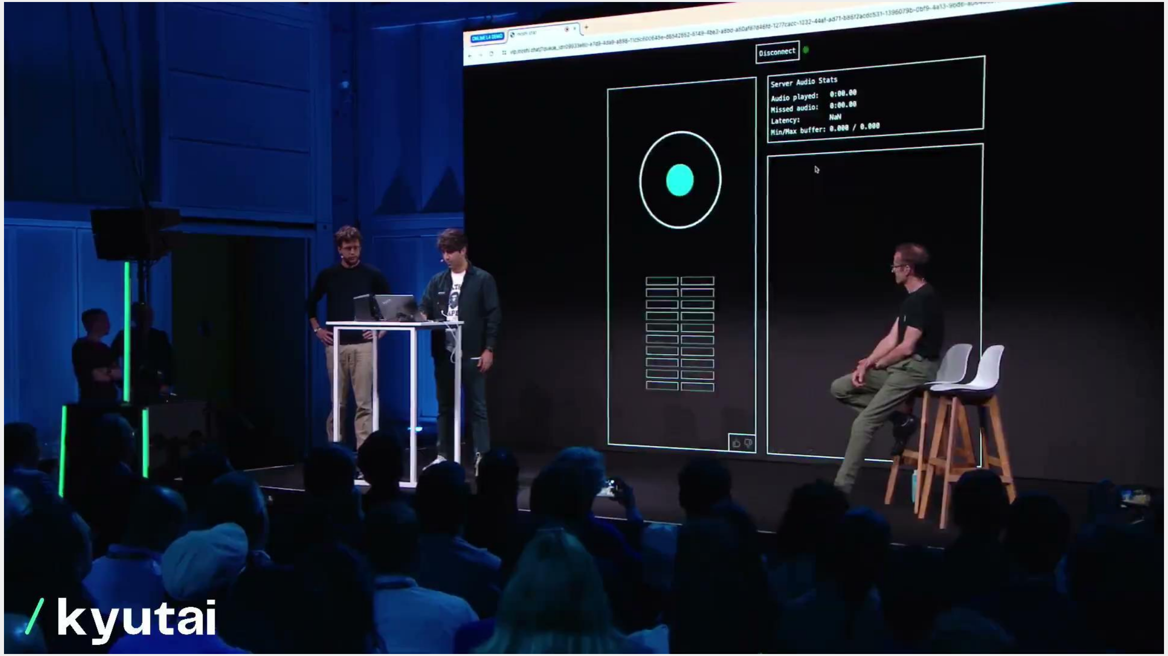
Model	Web Q.	LlaMA Q.	Audio Trivia QA
Audio only			
GSLM (Lakhotia et al., 2021)	1.5	4.0	-
AudioLM (Borsos et al., 2022)	2.3	7.0	-
TWIST (7B) (Hassid et al., 2023)	1.1	0.5	-
Moshi (w/o Inner Monologue)	**9.2**	**21.0**	7.3
Text and audio			
SpeechGPT (7B) (Zhang et al., 2024a)	6.5	21.6	14.8
Spectron (1B) (Nachmani et al., 2024)	6.1	22.9	-
Moshi	**26.6**	**62.3**	**22.8**
Moshi (w/o text batches in pre-training)	23.2	61.3	18.3
Text			
Helium (text)	32.3	75.0	56.4

Demos

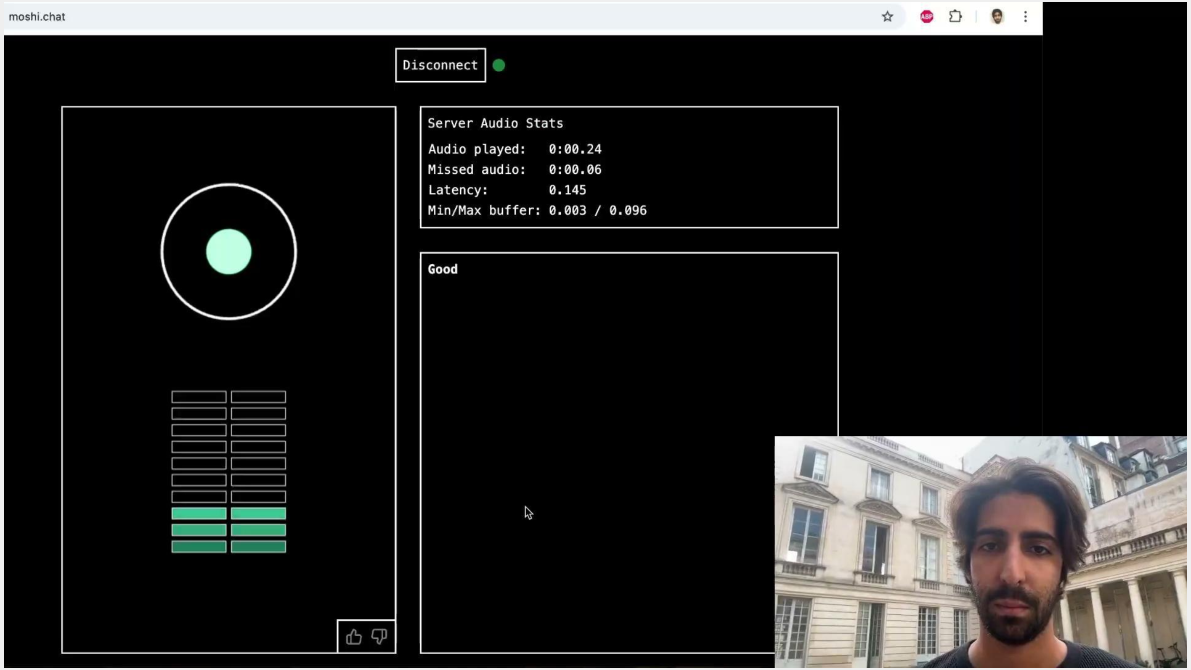
Moshi: playful and imaginative



Moshi: expressive



Moshi: resilient to noise



Moshi talks to Moshi

On a L4 GPU, ~200 ms latency (160ms theoretical).

Runs on a MacBook Pro with int4 quantization.

Online demo at moshi.chat.

Few issues remains: Moshi replies too quickly, or stay silent, or misses the point.

Conclusion

Moshi is the first real-time, low latency, full-duplex AI speech model.

Key contributions:

- Low frame rate, high quality neural codec Mimi.
- Improving RQ-Transformer for modeling two audio streams + inner monologue.
- Bootstrapped synthetic data.

Weights and inference code available github.com/kyutai-labs/moshi

Many more details in the paper!

Recent work on running DPO on Moshi to be published soon.

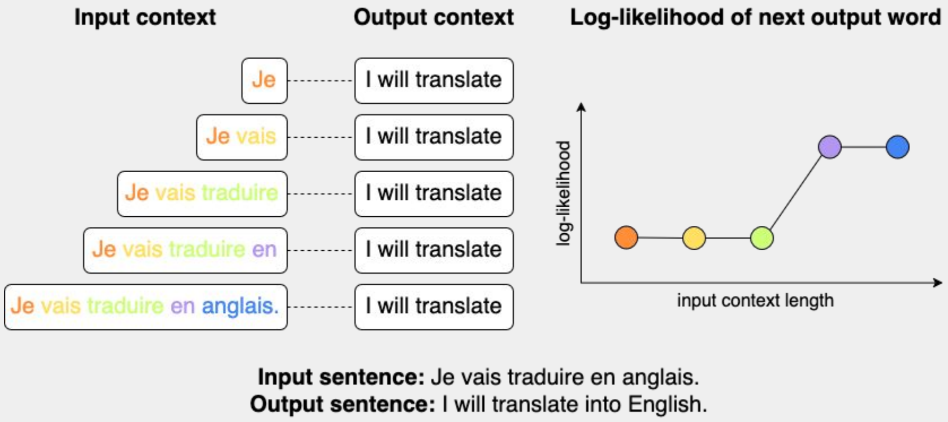
High-Fidelity Simultaneous Speech-To-Speech Translation

Extending the Moshi framework to live translation with Hibiki

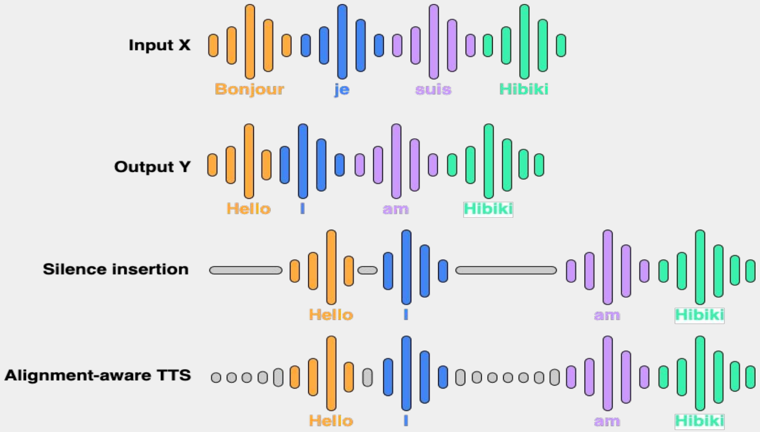
     

A slight change to the streams

And different synthetic data



First, we evaluate the maximum increase of the log-likelihood based on partial prefixes with a text model (MADLAD).



Second, we synthesize aligned data based on silence insertion or alignment aware TTS.

Hibiki presentation

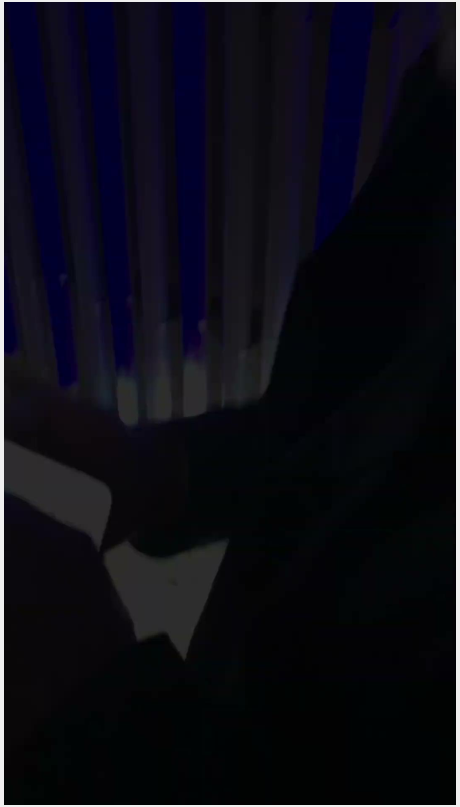
Table 3. **Human evaluation.** Raters report Mean Opinion Scores (MOS) between 1 and 5.

MODEL	QUALITY	SPEAKER SIM.	NATURALNESS
GROUND-TRUTH	4.18 ± 0.07	-	4.12 ± 0.08
SEAMLESS	2.22 ± 0.08	2.86 ± 0.12	2.18 ± 0.09
HIBIKI	$\mathbf{3.78} \pm 0.09$	$\mathbf{3.43} \pm 0.10$	$\mathbf{3.73} \pm 0.09$

- Live speech to speech + text translation model.
- Currently support French to English.
- Backbones are both 1B and 2B (+ Depth Transformer).
- 1B runs on an iPhone Live !
- Weights and code at github.com/kyutai-labs/hibiki

Demo

Demo



kyutai

www.kyutai.org