Crossing the uncanny valley of conversational voice

February 27, 2025

Brendan Iribe, Ankit Kumar, and the Sesame team

Try preview now

How do we know when someone truly understands us? It is rarely just our words—it is in the subtleties of voice: the rising excitement, the thoughtful pause, the warm reassurance.

Voice is our most intimate medium as humans, carrying layers of meaning through countless variations in tone, pitch, rhythm, and emotion.

Today's digital voice assistants lack essential qualities to make them truly useful. Without unlocking the full power of voice, they cannot hope to effectively collaborate with us. A personal assistant who speaks only in a neutral tone has difficulty finding a permanent place in our daily lives after the initial novelty wears off.

Over time this emotional flatness becomes more than just disappointing—it becomes exhausting.

Achieving voice presence

At Sesame, our goal is to achieve "voice presence"—the magical quality that makes spoken interactions feel real, understood, and valued. We are creating conversational partners that do not just process requests; they engage in genuine dialogue that builds confidence and trust over time. In doing so, we hope to realize the untapped potential of voice as the ultimate

interface for instruction and understanding.

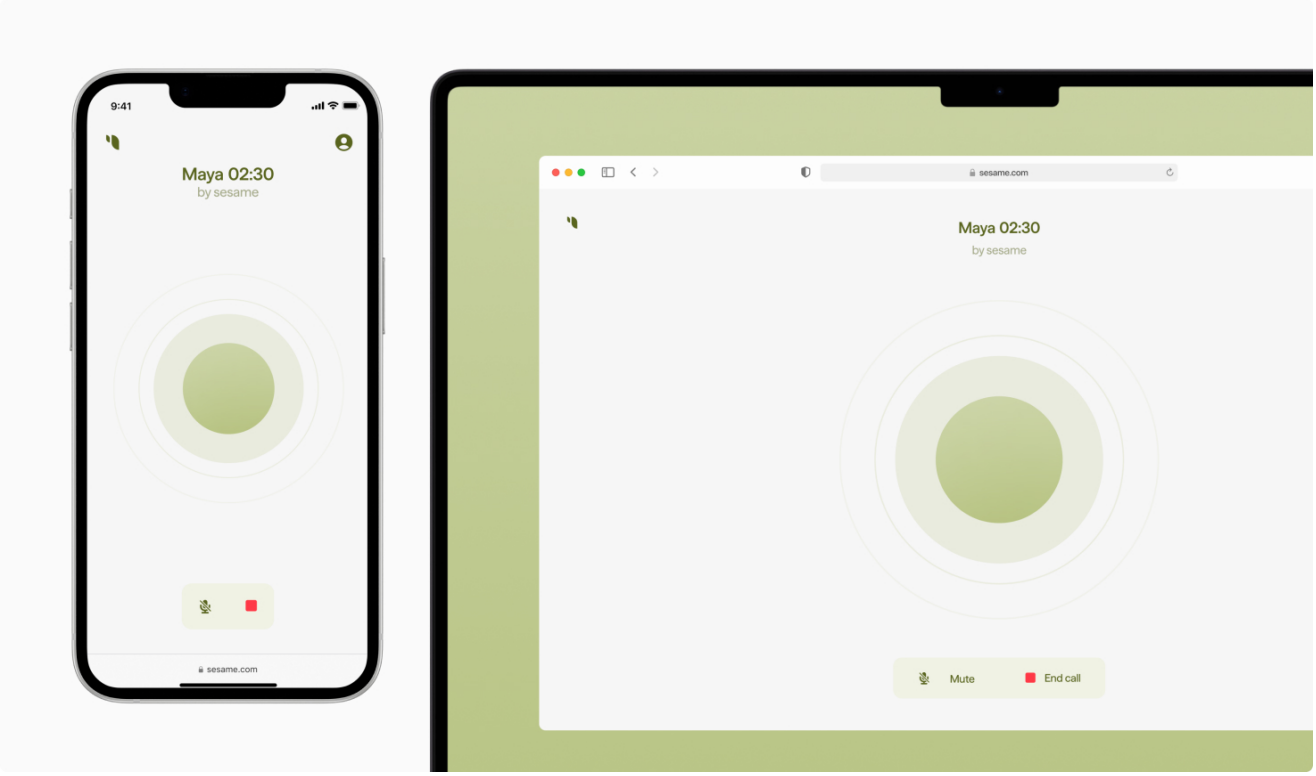
Key components

- Emotional intelligence: reading and responding to emotional contexts.
- Conversational dynamics: natural timing, pauses, interruptions and emphasis.
- Contextual awareness: adjusting tone and style to match the situation.
- Consistent personality: maintaining a coherent, reliable and appropriate presence.

We're not there yet

Building a digital companion with voice presence is not easy, but we are making steady progress on multiple fronts, including personality, memory, expressivity and appropriateness. This demo is a showcase of some of our work in conversational speech generation. The companions shown here have been optimized for friendliness and expressivity to illustrate the potential of our approach.

Try conversational speech in our preview



Try preview now

Technical post

Conversational speech generation

Authors
Johan Schalkwyk, Ankit Kumar, Dan Lyth,
Sefik Emre Eskimez, Zack Hodari, Cinjon Resnick,
Ramon Sanabria, Raven Jiang

To create AI companions that feel genuinely interactive, speech generation must go beyond producing high-quality audio—it must understand and adapt to context in real time. Traditional text-to-speech (TTS) models generate spoken output directly from text but lack the contextual awareness needed for natural conversations. Even though recent models produce highly human-like speech, they struggle with the one-to-many problem: there are countless valid ways to speak a sentence, but only some fit a given setting. Without additional context—including tone, rhythm, and history of the conversation—models lack the information to choose the best option. Capturing these nuances requires reasoning across multiple aspects of language and prosody.

To address this, we introduce the Conversational Speech Model (CSM), which frames the problem as an end-to-end multimodal learning task using transformers. It leverages the history of the conversation to produce more natural and coherent speech. There are two key takeaways from our work. The first is that CSM operates as a single-stage model, thereby improving efficiency and expressivity. The second is our evaluation suite, which is necessary for evaluating progress on contextual capabilities and addresses the fact that common public evaluations are saturated.

Background

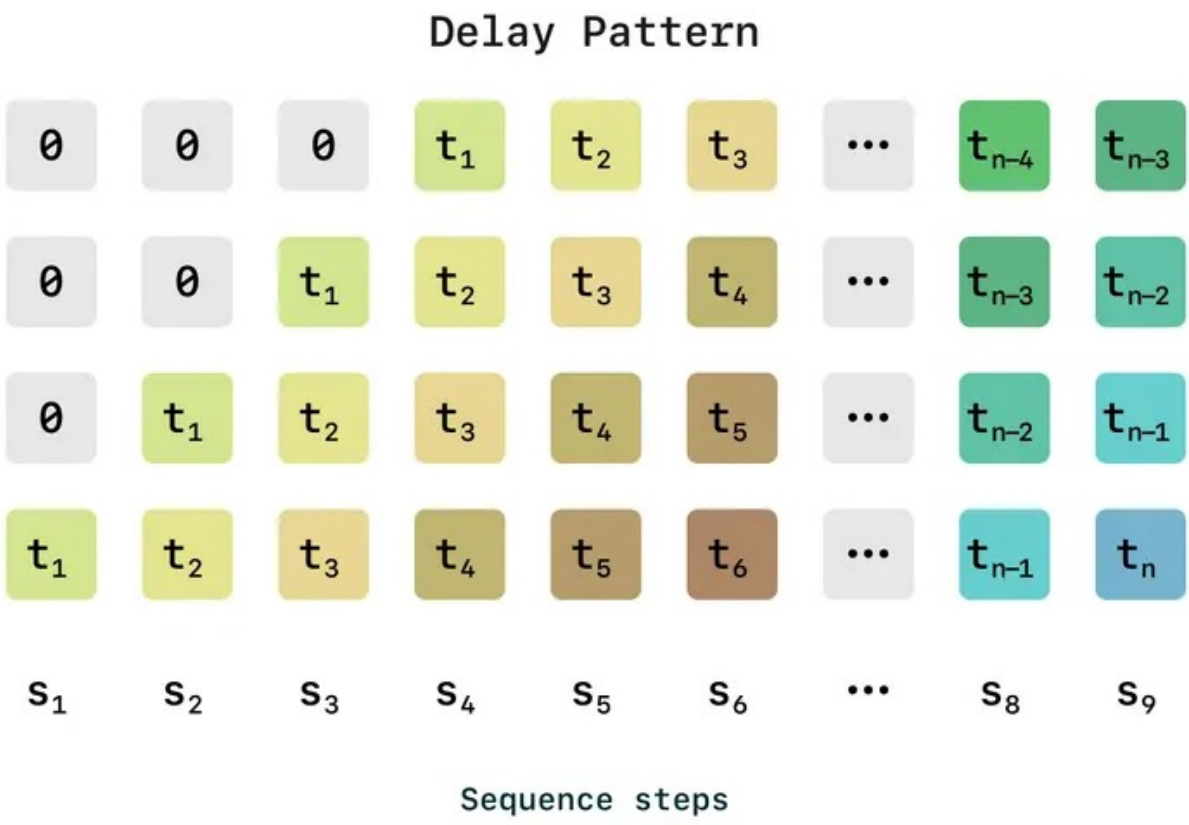
One approach to modeling audio with transformers is to convert continuous waveforms into discrete audio token sequences using tokenizers. Most contemporary approaches ([1], [2]) rely on two types of audio tokens:

1. Semantic tokens: Compact speaker-invariant representations of semantic and phonetic features. Their compressed nature enables them to capture key speech characteristics at the cost of high-fidelity representation.
2. Acoustic tokens: Encodings of fine-grained acoustic details that enable high-fidelity audio

reconstruction. These tokens are often generated using Residual Vector Quantization (RVQ) [2]. In contrast to semantic tokens, acoustic tokens retain natural speech characteristics like speaker-specific identity and timbre.

A common strategy first models semantic tokens and then generates audio using RVQ or diffusion-based methods. Decoupling these steps allows for a more structured approach to speech synthesis—the semantic tokens provide a compact, speaker-invariant representation that captures high-level linguistic and prosodic information, while the second-stage reconstructs the fine-grained acoustic details needed for high-fidelity speech. However, this approach has a critical limitation; semantic tokens are a bottleneck that must fully capture prosody, but ensuring this during training is challenging.

RVQ-based methods introduce their own set of challenges. Models must account for the sequential dependency between codebooks in a frame. One method, the delay pattern (figure below) [3], shifts higher codebooks progressively to condition predictions on lower codebooks within the same frame. A key limitation of this approach is that the time-to-first-audio scales poorly because an RVQ tokenizer with N codebooks requires N backbone steps before decoding the first audio chunk. While suitable for offline applications like audiobooks, this delay is problematic in a real-time scenario.



Example of delayed pattern generation in an RVQ tokenizer with 4 codebooks

Conversational Speech Model

CSM is a multimodal, text and speech model that operates directly on RVQ tokens. Inspired by the RQ-Transformer [4], we use two autoregressive transformers. Different from the approach in [5], we split the transformers at the zeroth codebook. The first **multimodal backbone** processes interleaved text and audio to model the zeroth codebook. The second **audio decoder** uses a distinct linear head for each codebook and models the remaining N – 1 codebooks to reconstruct speech from the backbone's representations. The decoder is significantly smaller than the backbone, enabling low-latency generation while keeping the model end-to-end.

CSM model inference process. Text (T) and audio (A) tokens are interleaved and fed sequentially into the Backbone, which predicts the zeroth level of the codebook. The Decoder then samples levels 1 through N – 1 conditioned on the predicted zeroth level. The reconstructed audio token (A) is then autoregressively fed back into the Backbone for the next step, continuing until the audio EOT symbol is emitted. This process begins again on the next inference request, with the interim audio (such as a user utterance) being represented by interleaved audio and text transcription tokens.

Both transformers are variants of the Llama architecture. Text tokens are generated via a Llama tokenizer [6], while audio is processed using Mimi, a split-RVQ tokenizer, producing one semantic codebook and N – 1 acoustic codebooks per frame at 12.5 Hz. [5] Training samples are structured as alternating interleaved patterns of text and audio, with speaker identity encoded directly in the text representation.

Compute amortization

This design introduces significant infrastructure challenges during training. The audio decoder processes an effective batch size of B × S and N codebooks autoregressively, where B is the original batch size, S is the sequence length, and N is the number of RVQ codebook levels. This high memory burden even with a small model slows down training, limits model scaling, and hinders rapid experimentation, all of which are crucial for performance.

To address these challenges, we use a compute amortization scheme that alleviates the memory bottleneck while preserving the fidelity of the full RVQ codebooks. The audio decoder is trained on only a random 1/16 subset of the audio frames, while the zeroth codebook is trained on every frame. We observe no perceivable difference in audio decoder losses during training when using this approach.

Amortized training process. The backbone transformer models the zeroth level across all frames (highlighted in blue), while the decoder predicts the remaining N – 31 levels, but only for a random 1/16th of the frames (highlighted in green). The top section highlights the specific frames modeled by the decoder for which it receives loss.

Experiments

Dataset: We use a large dataset of publicly available audio, which we transcribe, diarize, and segment. After filtering, the dataset consists of approximately one million hours of predominantly English audio.

Model Sizes: We trained three model sizes, delineated by the backbone and decoder sizes:

- Tiny: 1B backbone, 100M decoder
- Small: 3B backbone, 250M decoder

- **Medium**: 8B backbone, 300M decoder

Each model was trained with a 2048 sequence length (~2 minutes of audio) over five epochs.

Samples

Paralinguistics

00:00 00:07

00:00 00:03

Sentences from Base TTS

Foreign words

00:00 00:09

00:00 00:06

Sentences from Base TTS

Contextual expressivity

00:00 00:22

00:00 00:22

Samples from Expresso, continuation after chime

Pronunciation correction

00:00 00:06

00:00 00:07

Pronunciation correction sentence is a recording, all other audio is generated.

Conversations with multiple speakers

Single generation using audio prompts from two speakers

Evaluation

Our evaluation suite measures model performance across four key aspects: faithfulness to text, context utilization, prosody, and latency. We report both objective and subjective metrics—objective benchmarks include word error rate and novel tests like homograph disambiguation, while subjective evaluation relies on a Comparative Mean Opinion Score (CMOS) human study using the Expresso dataset.

Objective metrics

Traditional benchmarks, such as word error rate (WER) and speaker similarity (SIM), have become saturated—modern models, including CSM, now achieve near-human performance on these metrics.

Objective metric results for Word Error Rate (top) and Speaker Similarity (bottom) tests, showing the metrics are saturated (matching human performance).

To better assess pronunciation and contextual understanding, we introduce a new set of phonetic transcription-based benchmarks.

- Text understanding through Homograph Disambiguation: Evaluates whether the model correctly pronounced different words with the same orthography (e.g., “lead” /lɛd/ as in “metal” vs. “lead” /liːd/ as in “to guide”).
- Audio understanding through Pronunciation Continuation Consistency: Evaluates whether the model maintains pronunciation consistency of a specific word with multiple pronunciation variants in multi-turn speech. One example is “route” (/raʊt/ or /ruːt/), which can vary based on region of the speaker and context.

Objective metric results for Homograph Disambiguation (left) and Pronunciation Consistency (right) tests, showing the accuracy percentage for each model’s correct pronunciation. Play.ht, Elevenlabs, and OpenAI generations were made with default settings and voices from their respective API documentation.

The graph above compares objective metric results across three model sizes. For Homograph accuracy we generated 200 speech samples covering 5 distinct homographs—lead, bass, tear, wound, row—with 2 variants for each and evaluated pronunciation consistency using wav2vec2-lv-60-espeak-cv-ft. For Pronunciation Consistency we generated 200 speech samples covering 10 distinct words that have common pronunciation variants—aunt, data, envelope, mobile, route, vase, either, adult, often, caramel.

In general, we observe that performance improves with larger models, supporting our hypothesis that scaling enhances the synthesis of more realistic speech.

Subjective metrics

We conducted two Comparative Mean Opinion Score (CMOS) studies using the Expresso dataset to assess the naturalness and prosodic appropriateness of generated speech for CSM-Medium. Human evaluators were presented with pairs of audio samples—one generated by the

model and the other a ground-truth human recording. Listeners rated the generated sample on a 7-point preference scale relative to the reference. Expresso's diverse expressive TTS samples, including emotional and prosodic variations, make it a strong benchmark for evaluating appropriateness to context.

In the first CMOS study we presented the generated and human audio samples with no context and asked listeners to *"choose which rendition feels more like human speech."* In the second CMOS study we also provide the previous 90 seconds of audio and text context, and ask the listeners to *"choose which rendition feels like a more appropriate continuation of the conversation."* Eighty people were paid to participate in the evaluation and rated on average 15 examples each.

Subjective evaluation results on the Expresso dataset. No context: listeners chose *"which rendition feels more like human speech"* without knowledge of the context. Context: listeners chose *"which rendition feels like a more appropriate continuation of the conversation"* with audio and text context. 50:50 win–loss ratio suggests that listeners have no clear preference.

The graph above shows the win-rate of ground-truth human recordings vs CSM-generated speech samples for both studies. Without conversational context (top), human evaluators show no clear preference between generated and real speech, suggesting that naturalness is saturated. However, when context is included (bottom), evaluators consistently favor the original recordings. These findings suggest a noticeable gap remains between generated and human prosody in conversational speech generation.

Open-sourcing our work

We believe that advancing conversational AI should be a collaborative effort. To that end, we're committed to open-sourcing key components of our research, enabling the community to experiment, build upon, and improve our approach. Our models will be available under an

Apache 2.0 license.

Check out our GitHub for updates and contributions

Limitations and future work

CSM is currently trained on primarily English data; some multilingual ability emerges due to dataset contamination, but it does not perform well yet. It also does not take advantage of the information present in the weights of pre-trained language models.

In the coming months, we intend to scale up model size, increase dataset volume, and expand language support to over 20 languages. We also plan to explore ways to utilize pre-trained language models, working towards large multimodal models that have deep knowledge of both speech and text.

Ultimately, while CSM generates high quality conversational prosody, it can only model the text and speech content in a conversation—not the structure of the conversation itself. Human conversations are a complex process involving turn taking, pauses, pacing, and more. We believe the future of AI conversations lies in fully duplex models that can implicitly learn these dynamics from data. These models will require fundamental changes across the stack, from data curation to post-training methodologies, and we're excited to push in these directions.

Join us

If you're excited about building the most natural, delightful, and inspirational voice interfaces out there, reach out—we're hiring. Check our open roles.