Little Voices

Made with Love for Kids

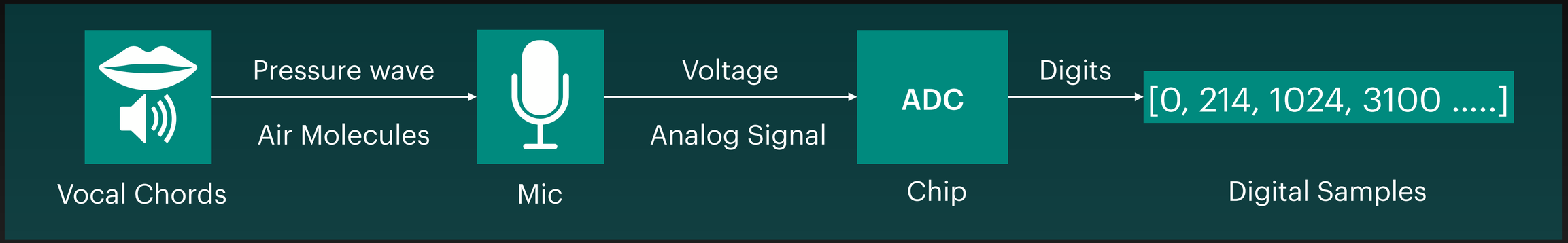
(Human Taste + AI Execution)

S Irfan Basha

A Father. His Children, and a Need

Demo

What is Sound?



- Sound = Air pressure variations over time
- Microphone = converts pressure —> electrical voltage *(analog)*
- ADC = Samples the voltage thousands of times per second —> List of integers

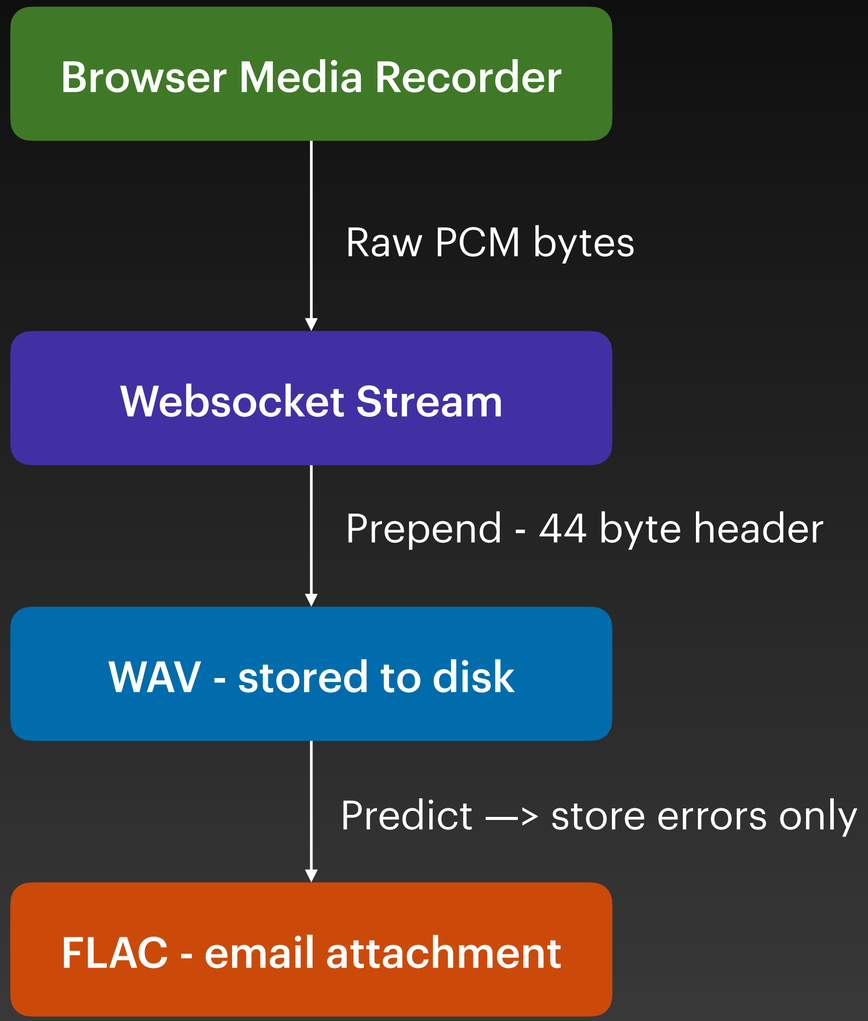
The Three Numbers That Define Audio

Parameter	What it controls	Our choice	Why
Sample Rate	How often we measure	16,000 Hz	Nyquist: speech tops at 8kHz -> need 2x = 16kHz. Wishper trained on 16kHz
Bit Depth	How precisely we measure	16-bit	65, 536 steps — clean voice, no overkill
Channels	How many mics	Mono (1)	ASR ignores spatial audio. Stereo = double data, zero benefit

1 min * 16,000 samples/sec * 2 bytes * 1 channel = 1.9 MB/minute

PCM, WAV, FLAC

Same Audio, Different Wrappers



- PCM = Flat list of integers. Zero overhead.
- WAV = PCM + 44 byte header. Universally readable
- FLAC = lossless compression *(~50% smaller)*

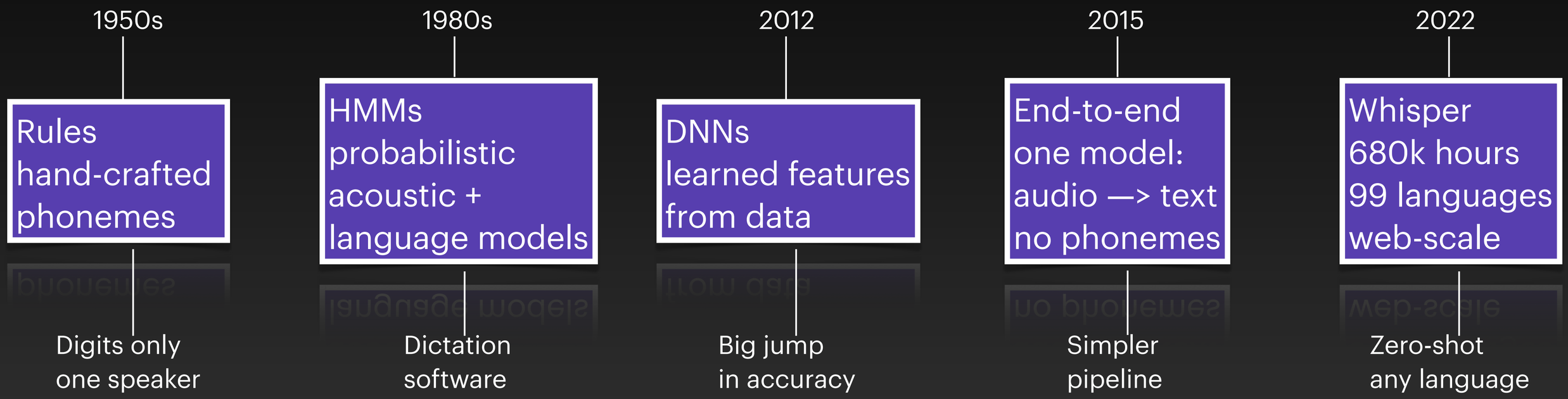
VAD - Voice Activity Detection

The Quality Gate

	Without VAD	With VAD
Silence	Whisper -> “Thank you...”	[SKIP]
Fan hum	Whisper -> “Please subscr...”	[SKIP]
Breathing	Whisper -> “BG music..”	[SKIP]
Hi Mom	Whisper -> “Hi Mom”	Whisper -> “Hi Mom”
	✗ Hallucination	✓ Clean input

Final Transcription	Live Transcription
vad_filter=True	vad_filter=False
500ms silence gap = cut	Everything passes through
Accuracy over completeness	Responsiveness over accuracy

The History of Speech Recognition



Every generation stopped hand-engineering one more thing and let the data decide instead.

How Whisper Works

Audio (16kHz WAV)

STFT + Mel Filterbank

- 2D grid:
  - 80 freq bands * 3000 time steps
- Whisper sees audio as a picture

What sounds are here?

What words do I write?

<|te|> <|transcribe|>

నేను బడికి వెళ్తున్నాను.

<|te|> —> Telugu script
<|en|> —> English
Same model.
One token switches language.

<|transcribe|> —> Write what was said
<|translate|> —> translate to desired language
Same model.
One token switches task.

Multitask training data (680k hours)

English transcription

"Ask not what your country can do for ..."

Ask not what your country can do for ...

Any-to-English speech translation

"El rápido zorro marrón salta sobre ..."

The quick brown fox jumps over ...

Non-English transcription

"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ..."

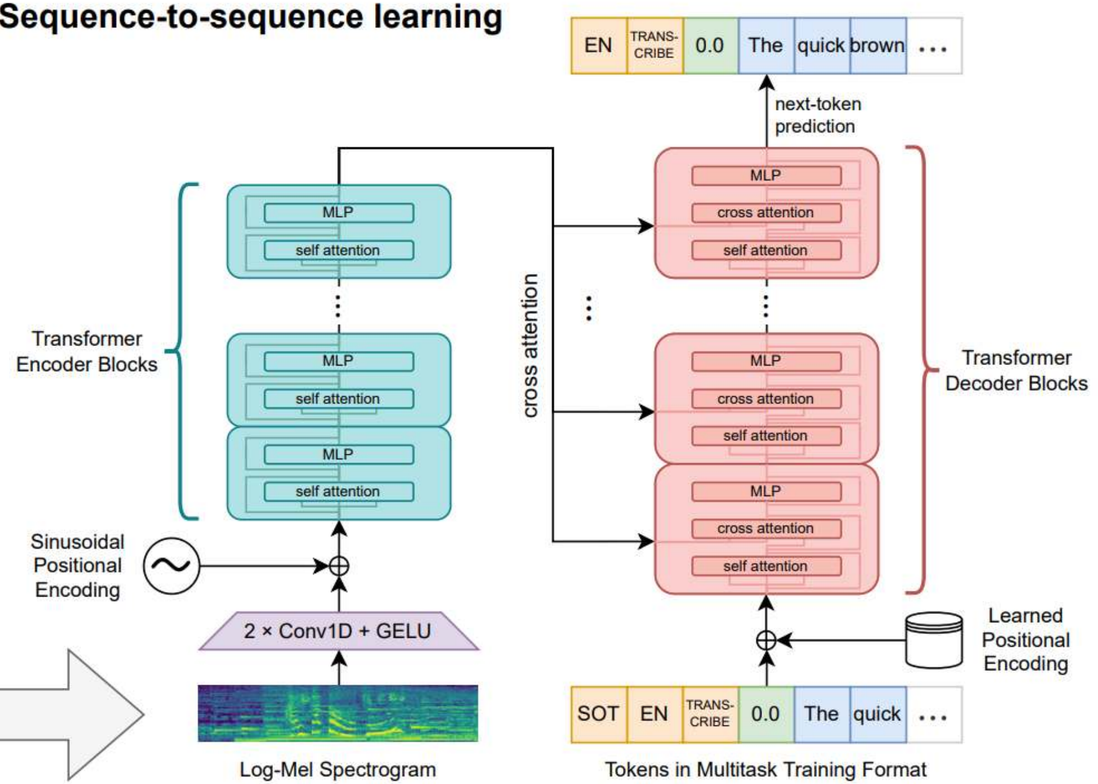
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ...

No speech

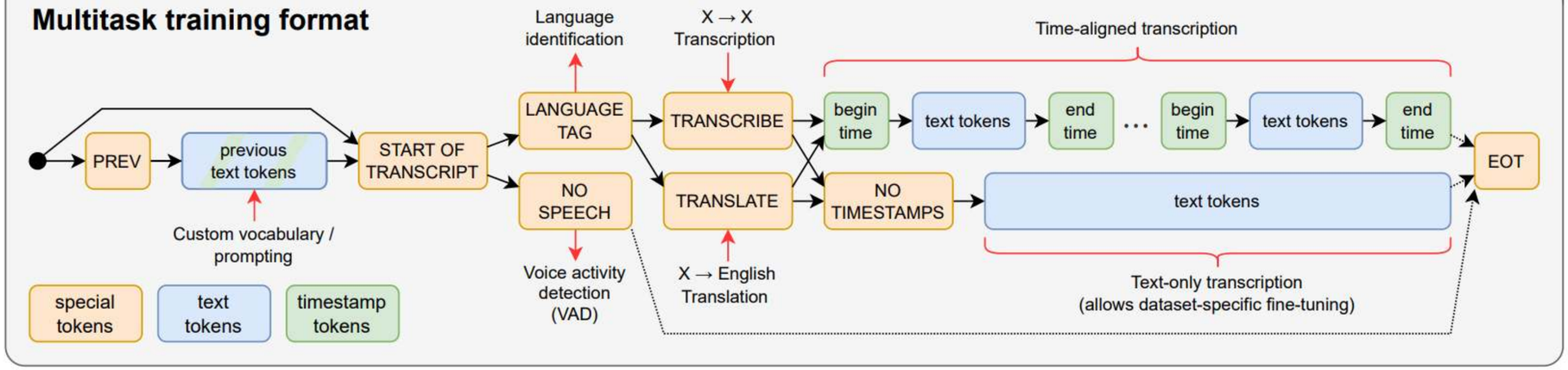
(background music playing)

Ø

Sequence-to-sequence learning



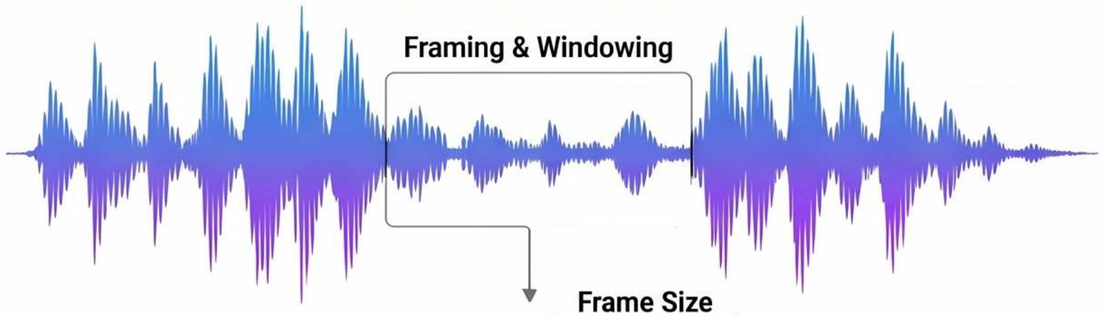
Multitask training format

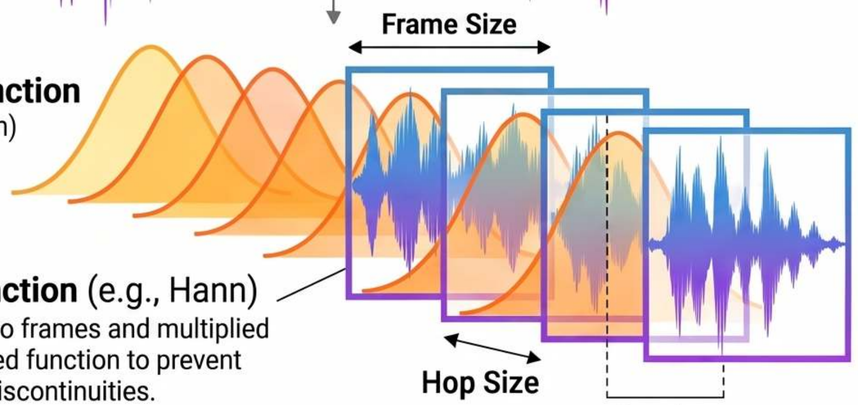


From Sound to Vision: The Mechanics of the Short-Time Fourier Transform (STFT)

How an audio signal is transformed into a 2D spectrogram by breaking continuous audio into overlapping, windowed chunks to capture how sound evolves

Continuous Audio Signal

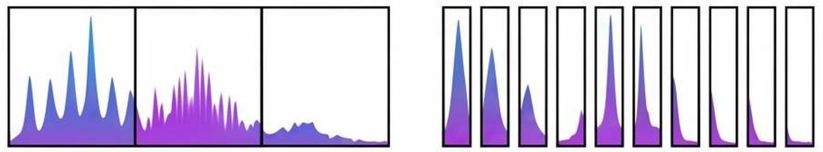




Window Function (e.g., Hann)
Signal divided into frames and multiplied by a bell-shaped function to prevent edge discontinuities.

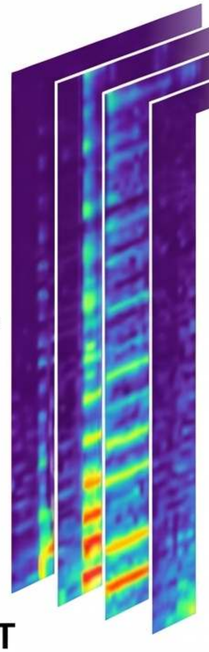
Larger Frames =
Better Frequency Resolution
Worse Time Resolution

Smaller Frames =
Better Time Resolution
Worse Frequency Resolution



The Time-Frequency Trade-off

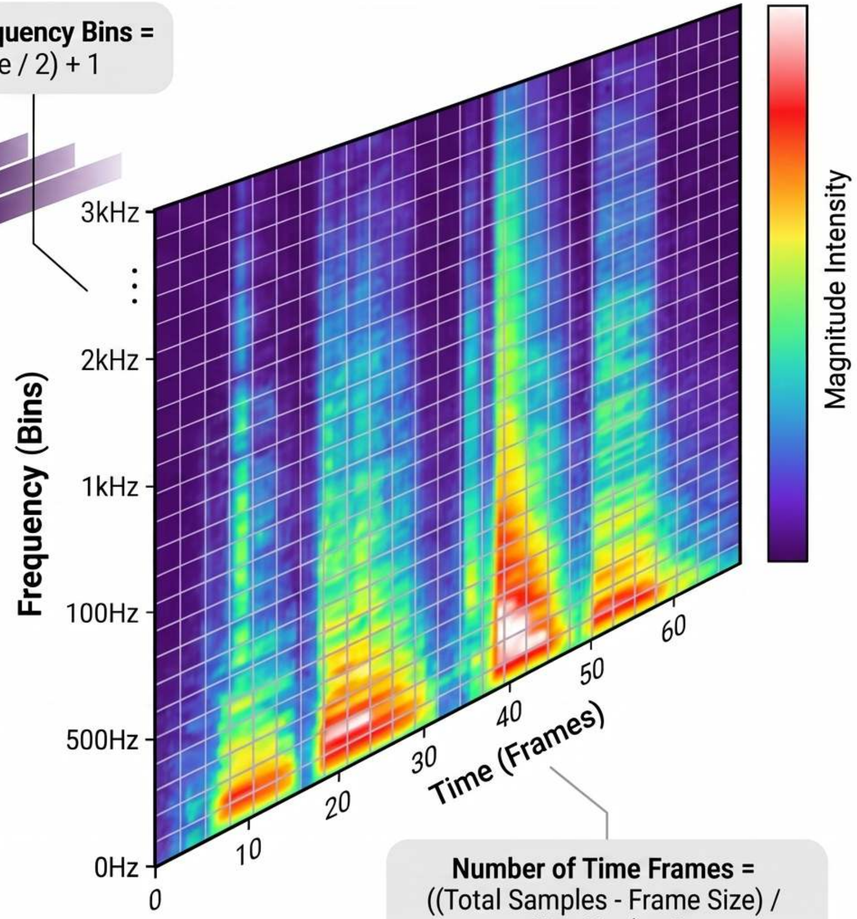
Per-Frame DFT
(Discrete Fourier Transform)
DFT applied to each individual frame to extract its frequency bins.



Number of Frequency Bins =
(Frame Size / 2) + 1

Magnitude Spectrogram

Complex coefficients are converted to real numbers and visualized as a heatmap of intensity



Number of Time Frames =
((Total Samples - Frame Size) / Hop Size) + 1

What Whisper Actually Sees

The Log-Mel Spectrogram

Why Mel Scale?

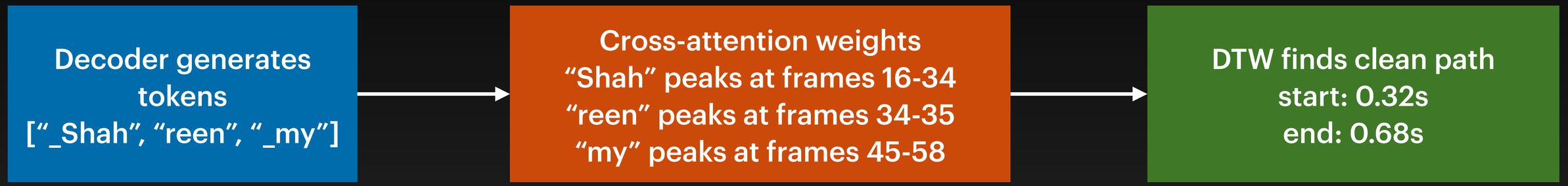
Human hearing is logarithmic
We hear 200Hz vs 400Hz easily
We barely notice 7800Hz vs 8000Hz
Mel scale matches perception
- compress high, expand low

Why Log Scale?

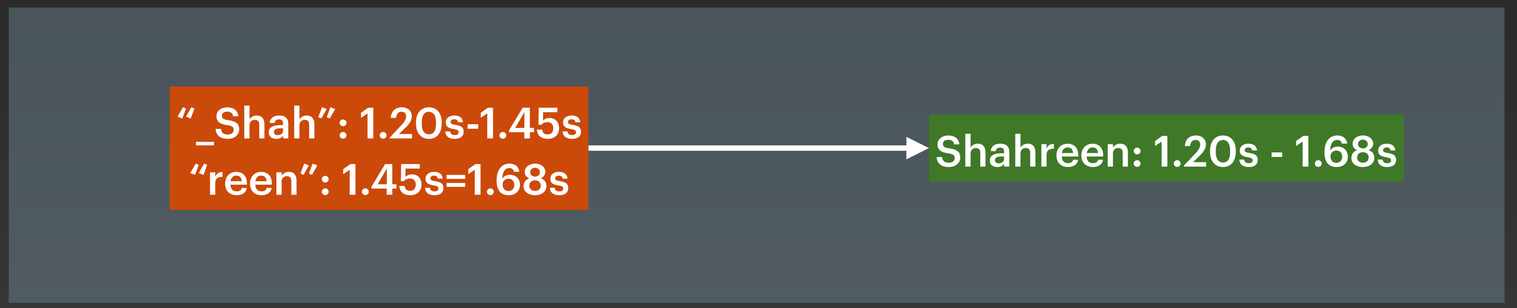
A whisper vs a shout = 1, 000, 000x energy difference.
Log compresses this to a 0-6 range
Quiet and loud speech become equally learnable

Word-level Timestamps

How Whisper Knows When Each Word was Said



Subword - Merging: Byte Pair Encoding



Beam Search vs Greedy Decoding: The Speed-Accuracy Dial

Greedy Decoding (beam_size=1)

Picks the highest probability token at every step for maximum speed (~100ms per 2s audio).

~100ms (per 2s audio)

Highest probability only

Instant, but rigid

Example:

Accuracy Variance:
Greedy may misidentify ‘Shahreen’ as ‘Sahreen’

Sahreen

Speed-Accuracy Dial

This infographic compares two primary decoding strategies used in audio signal processing, highlighting the 'Speed-Accuracy Dial' used by the Little Voices app to balance instant user feedback with high-precision final results.

Beam Search (beam_size=5)

Tracks multiple candidates simultaneously for higher accuracy (~400-800ms per 2s audio).

~400-800ms (per 2s audio)

Keeps top 5 candidates alive

Slower, but accurate

Example:

Shahreen

Accuracy Variance:
Beam Search correctly identifies the winner.

Implementation in ‘Little Voices’ application

Recording

Live Transcription: Instant Feedback

Uses Greedy Decoding as the child records so words appear on screen instantly.

Final Processing: Permanent Save

Upon stopping, Beam Search overwrites the live display with the high-accuracy result.

Saved

The Hallucination Problem

- Audio: [silence] → "Thank you for watching. Don't forget to subscribe."
- Audio: [fan noise] → "I'll see you in the next video."
- Audio: [child: "Hi"] → "Hi. Welcome to my channel. Today we're going to..."

Mitigation	Where	What it Prevents
temperature = 0.0	Live transcription	Prevents random amplification of transcription errors
vad_filter = True	Final transcription	Prevents silent frames from reaching the decoder
100 ms minimum clip	WebSocket handler	Prevents clips that are too short to contain meaningful speech
Explicit language token	Every transcription call	Prevents incorrect language detection/guessing
Human review	Parent Journal Manager	Catches anything that slipped through the automated pipeline

The Audio Transcription Pipeline: From Sound Waves to Real-Time Text

Child presses 'Start Recording'.
The user interaction triggers the browser to open a WebSocket connection to the backend, initiating a new recording row in the SQLite database.

User Activation

Browser

WebSocket

Backend New row

SQLite Database

Raw Signal Capture

Real-Time Framing

250ms

Real-Time Framing

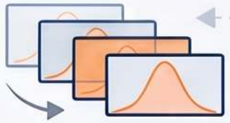
Windowing Function

Phase 2:
Signal Framing and Windowing

MediaRecorder Captures Microphone Input
The browser captures audio in PCM int16 format at 10kHz, preparing the 'time-domain' waveform for digital processing.

Audio Segments Dispatched Every 250ms
The signal is divided into 'frames'—small chunks of samples—to move from a static average to a dynamic sequence of frequency components.

Applying the Bell-Shaped Hann Window
Each frame is multiplied by a windowing function to squash the sample values at the edges to zero, avoiding discontinuities and spectral leakage.



KEY_FINDING: The Hop Size (Capital H)

Sliding the Window for Overlap
The system slides the window by a specific 'hop eize' to create overlapping froines, ensuring no critical audio data is lost between segments.

Spectrogram Matrix Visualization

Total Signal Samples: 10,000
Frame Size (H): 1,000
Hop Biue (H): OOO

Resolting Output Shape:
16 Frames (Temporal Dins),
SOT Frequency Dins

Calculation: (10,000 - 1,000) / 000 × 1

Phase 3:
Processing and Transcription

Real-Time UI Updates

beam size 1

Small Model

UI Preview

hello...
world... is...
is...

Async Transcription with Small Model
A lightweight model with a beam size of 1 processes the backend buffer quickly to provide the user with immediate, low-latency text proviews.

The Time-Frequency Trade-Off

Balancing Resolution and Speed
Larger frame sizes increase frequency resolution (better for accuracy) but degrade time resolution, a critical heuristic when choosing model parameters.

beam size 5

Final Transcription

Large-v3 Model

Voice Activity Detection (VAD)

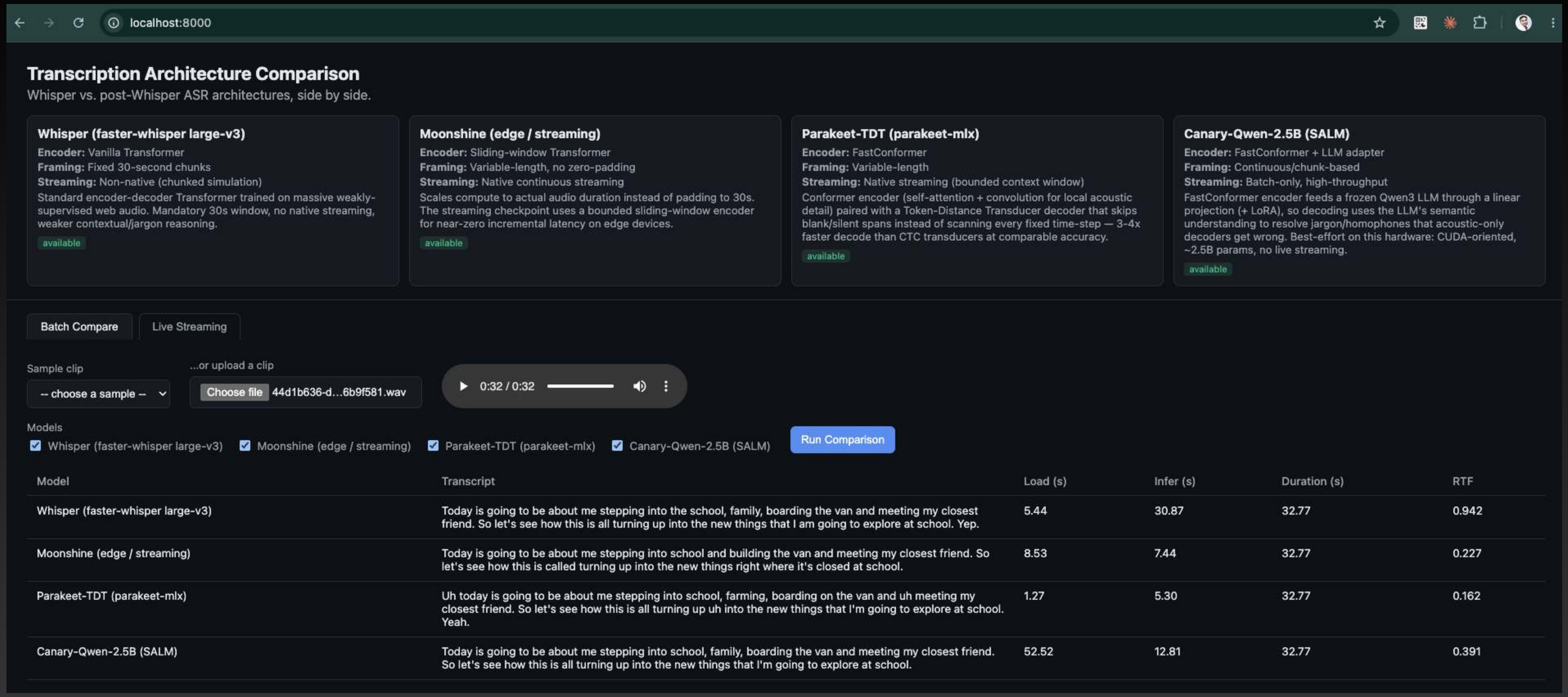
High-Resolution Processing with Large-v3 Model
Once the child presses 'Stop Recording', a larger model (beam size 5) with Voice Activity Detection (VAB) performs the final, high-accuracy transcription.

Phase 4:
Data Finalisation

SQLite Update and UI Refresh

Hello World, this is the final

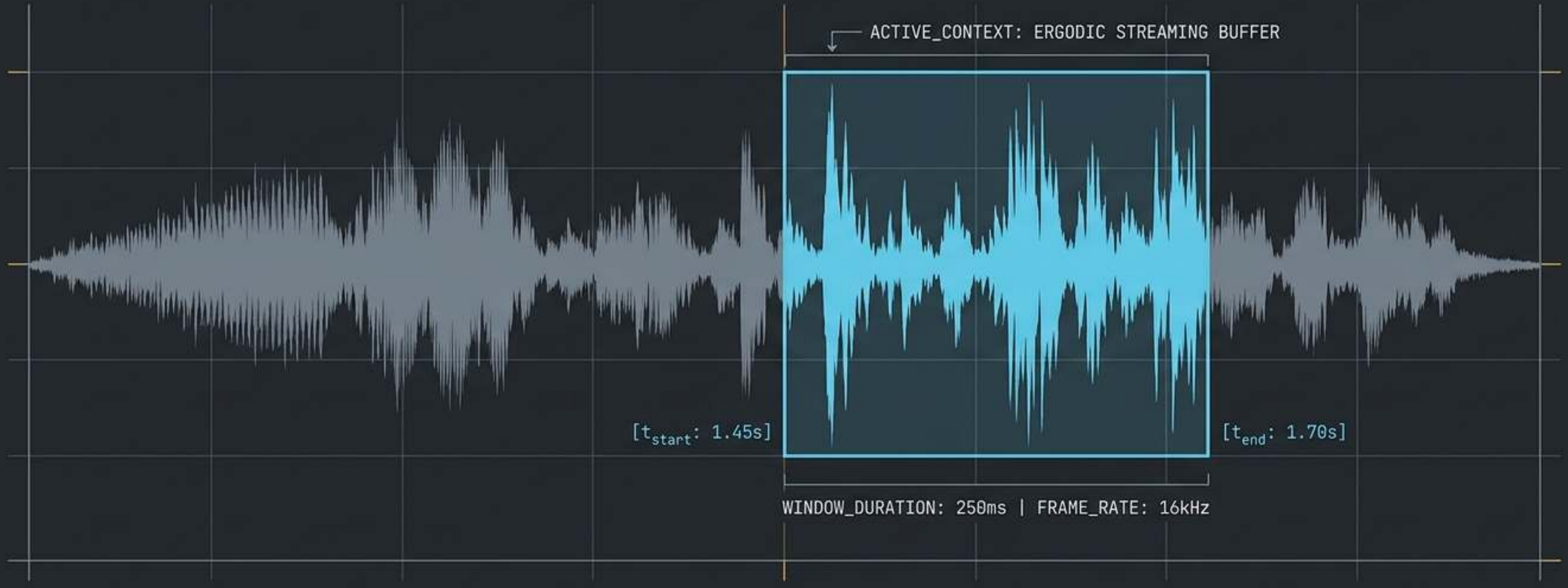
Database Synchronisation
The final, high-accuracy text replaces the real-time preview in the UI and is permanently stored in the system database.



Model	Transcript	Load (s)	Infer (s)	Duration (s)	RTF
Whisper (faster-whisper large-v3)	Today is going to be about me stepping into the school, family, boarding the van and meeting my closest friend. So let's see how this is all turning up into the new things that I am going to explore at school. Yep.	5.44	30.87	32.77	0.942
Moonshine (edge / streaming)	Today is going to be about me stepping into school and building the van and meeting my closest friend. So let's see how this is called turning up into the new things right where it's closed at school.	8.53	7.44	32.77	0.227
Parakeet-TDT (parakeet-mlx)	Uh today is going to be about me stepping into school, farming, boarding on the van and uh meeting my closest friend. So let's see how this is all turning up uh into the new things that I'm going to explore at school. Yeah.	1.27	5.30	32.77	0.162
Canary-Qwen-2.5B (SALM)	Today is going to be about me stepping into school, family, boarding the van and meeting my closest friend. So let's see how this is all turning up into the new things that I'm going to explore at school.	52.52	12.81	32.77	0.391

Moonshine v2: Ergodic Streaming Encoders for Latency-Critical Voice AI

Breaking the algorithmic latency bottleneck for real-time edge speech applications.



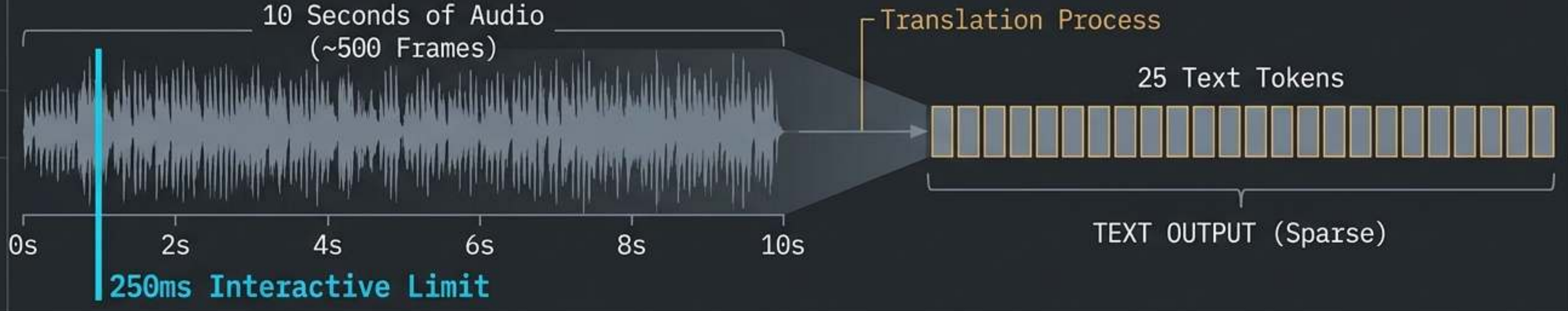
The biological constraints of human-computer acoustic interaction

The Biological Expectation

- **The 250ms Threshold:** Humans expect real-time conversational feedback. Any system delay >250ms breaks the illusion of interactivity.
- **Acoustic Context:** A human syllable spans ~150-300 ms of acoustic context. Humans do not need to hear the end of a sentence to parse the beginning.

The Data Asymmetry

- **Density vs. Sparsity:** Audio data is extremely dense; text output is extremely sparse.
- **The Translation Rate:** In the time it takes to speak 500 frames of audio (10 seconds), a human might only speak 20 or 30 words.

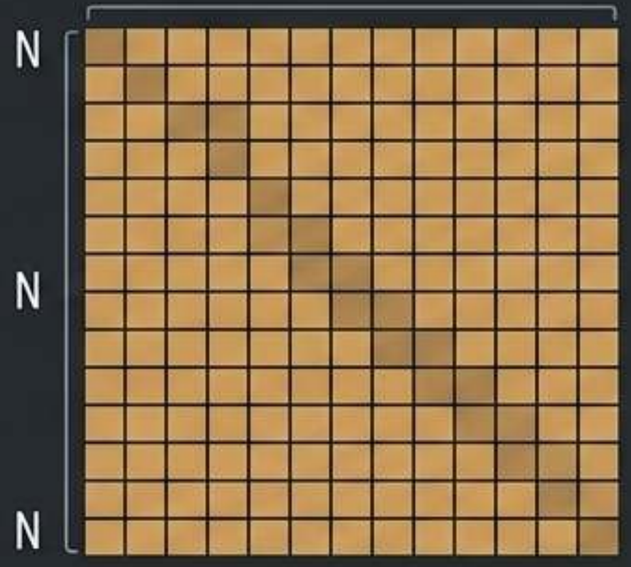


Syllable

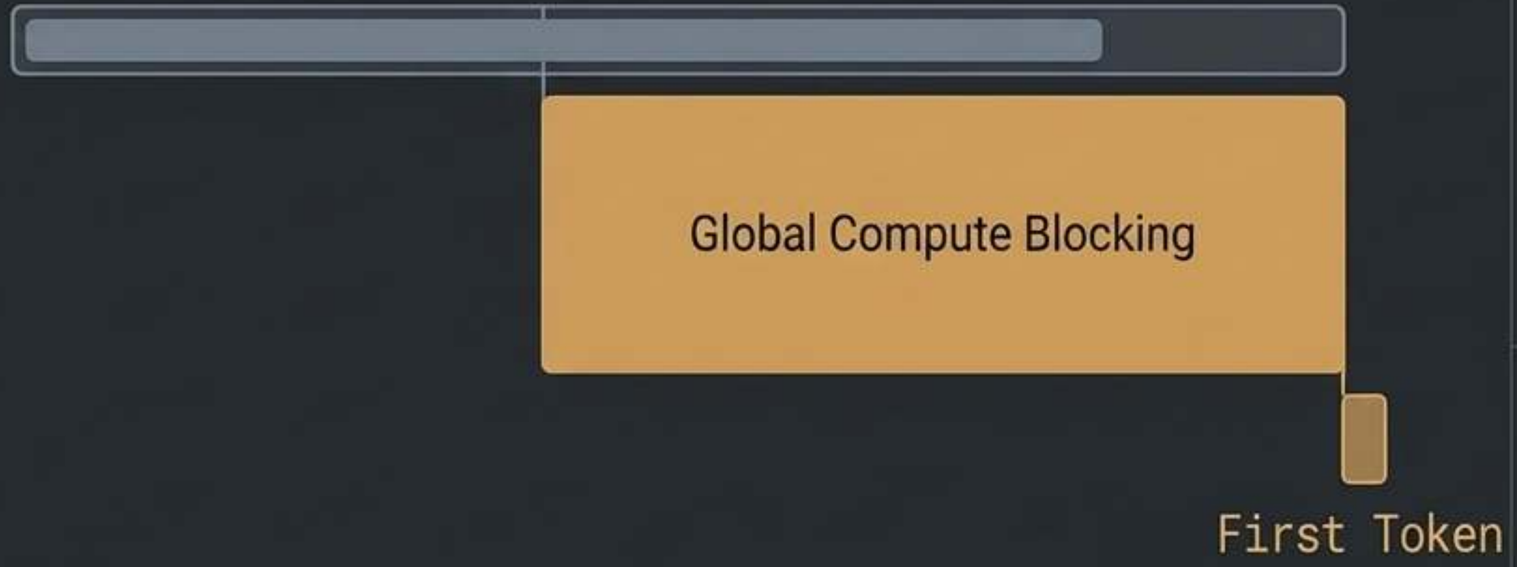
Word	Syllables	Pronunciation
Book	1	Book
Hello	2	Hel. Lo
Computer	3	Com. Pu. Ter
Intelligence	4	In. Tel. Li. Gence

Global attention models create an algorithmic latency trap

O(N^2) Full-Attention Matrix



10-second audio stream



Full-Attention Accuracy

Resolves locally ambiguous acoustics using distant lexical context. Every frame attends to every other frame.

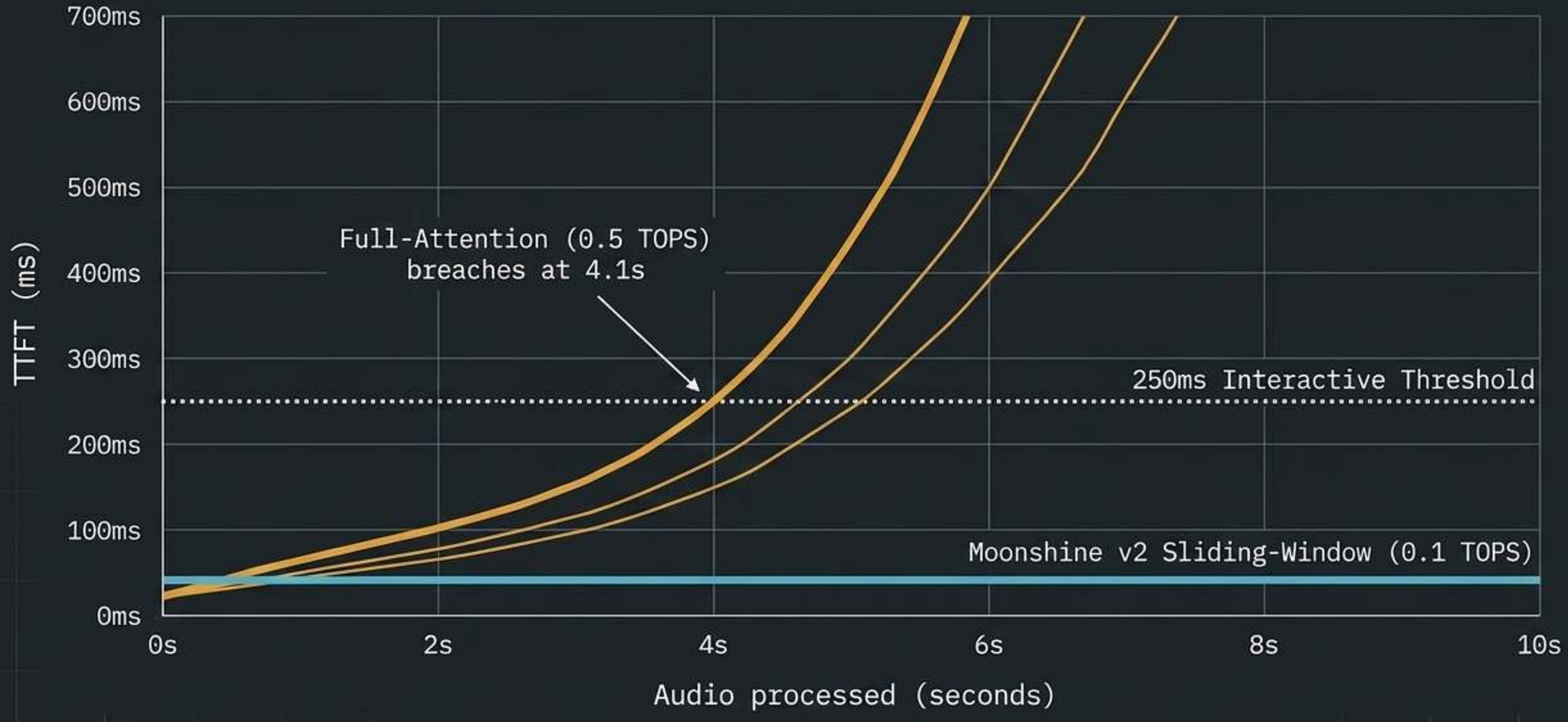
Quadratic Complexity

Self-attention mixing work grows quadratically with sequence length.

The TTFT Bottleneck

Induces an “encode-the-whole-utterance” latency profile, blocking the decoder from emitting a single token.

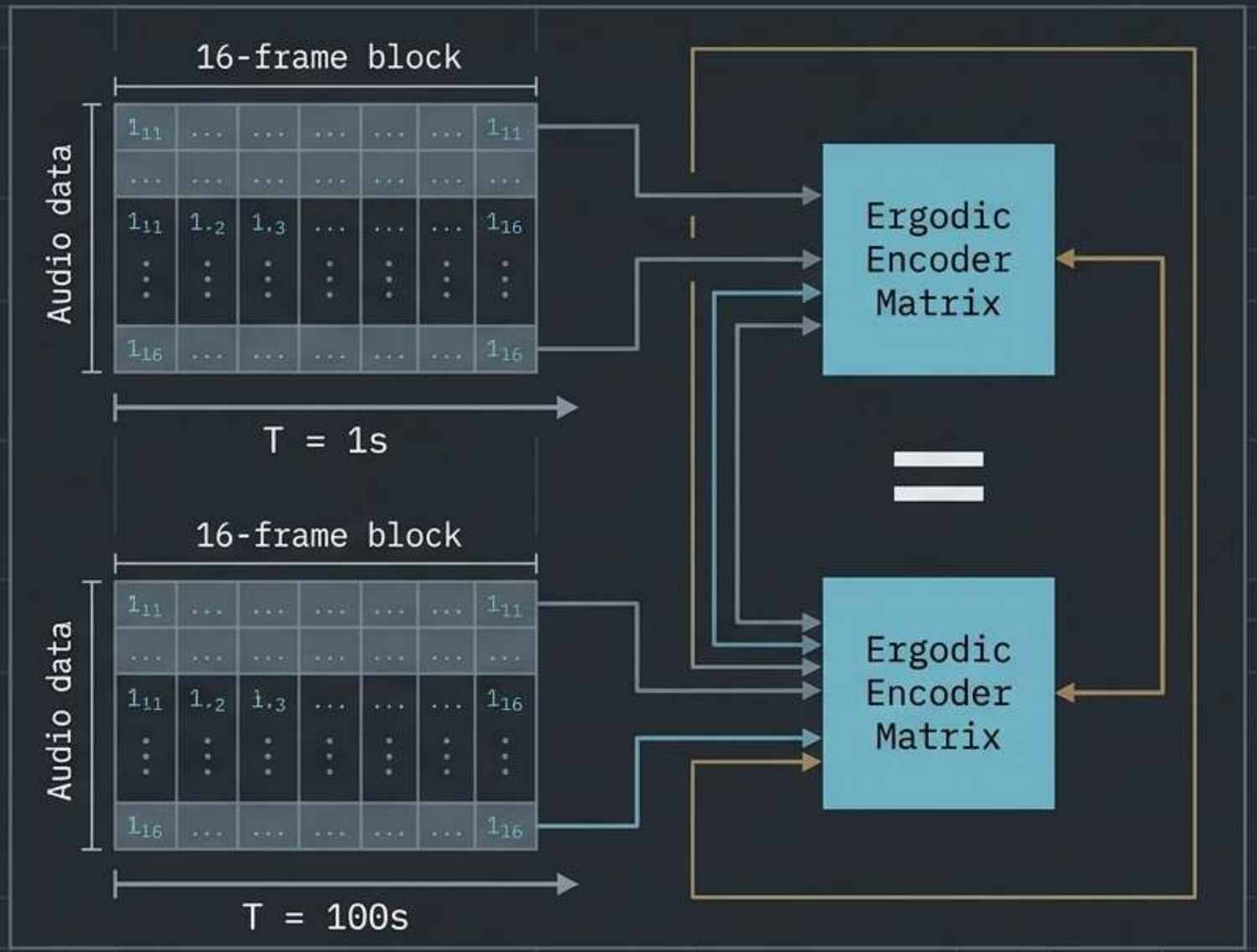
Time-to-First-Token (TTFT) scales linearly in global attention systems



Bounded inference via sliding-window self-attention

Linear Complexity	Bounded Lookahead	Constant Latency
Algorithmic cost drops to O(Tw) where w is the fixed window size.	The encoder uses a strict local window, capping future lookahead to just 4 frames of right context.	4 frames at 50 Hz equals exactly 80 ms of algorithmic lookahead. Inference latency becomes predictable and detached from total utterance duration.

The Ergodic Advantage: Translation-invariant encoding



No Positional Embeddings

The encoder uses zero absolute or relative positional embeddings.

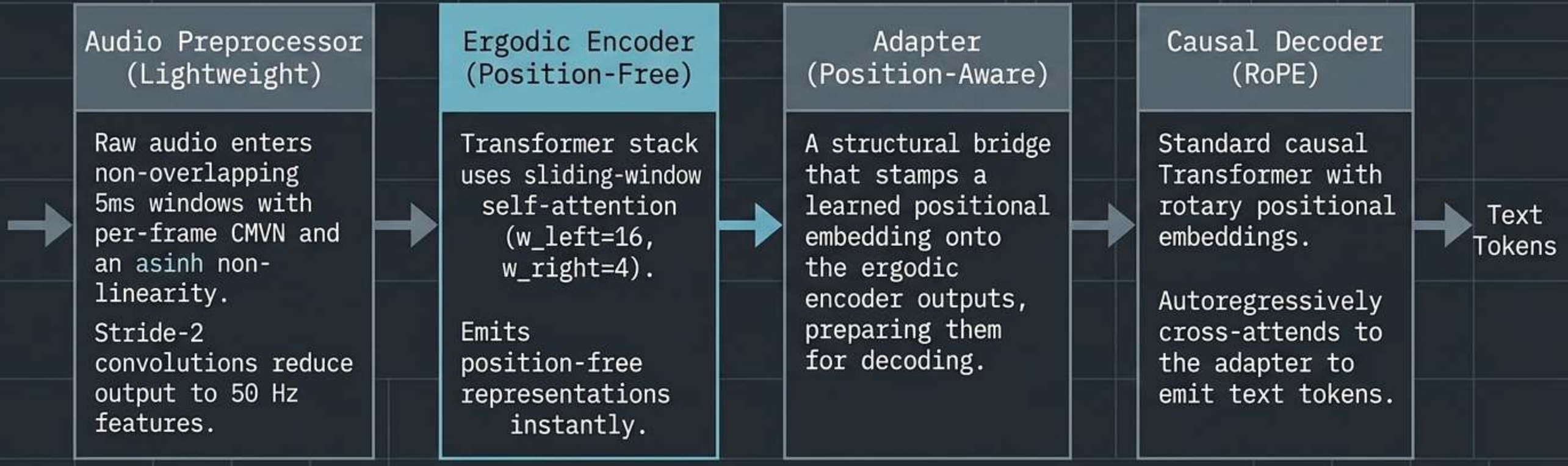
Position-Free Processing

The encoder is ergodic—it possesses no explicit notion of absolute time, inferring structure solely from the raw acoustic content of the local sliding window.

Infinite Streaming

Because the math is invariant at 1 second or 1,000 seconds, the system operates continuously without memory overflow or positional degradation.

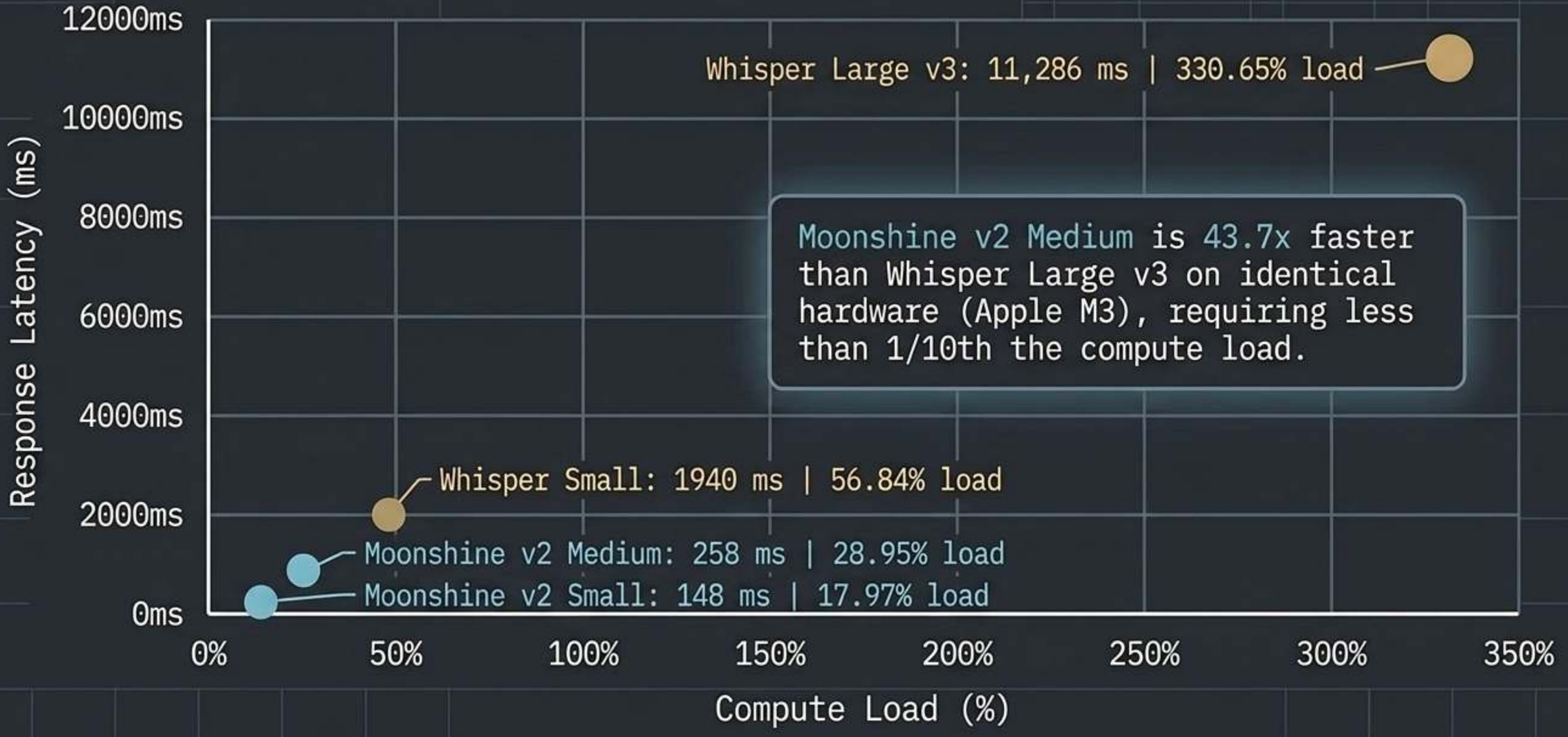
The Moonshine v2 system architecture pipeline



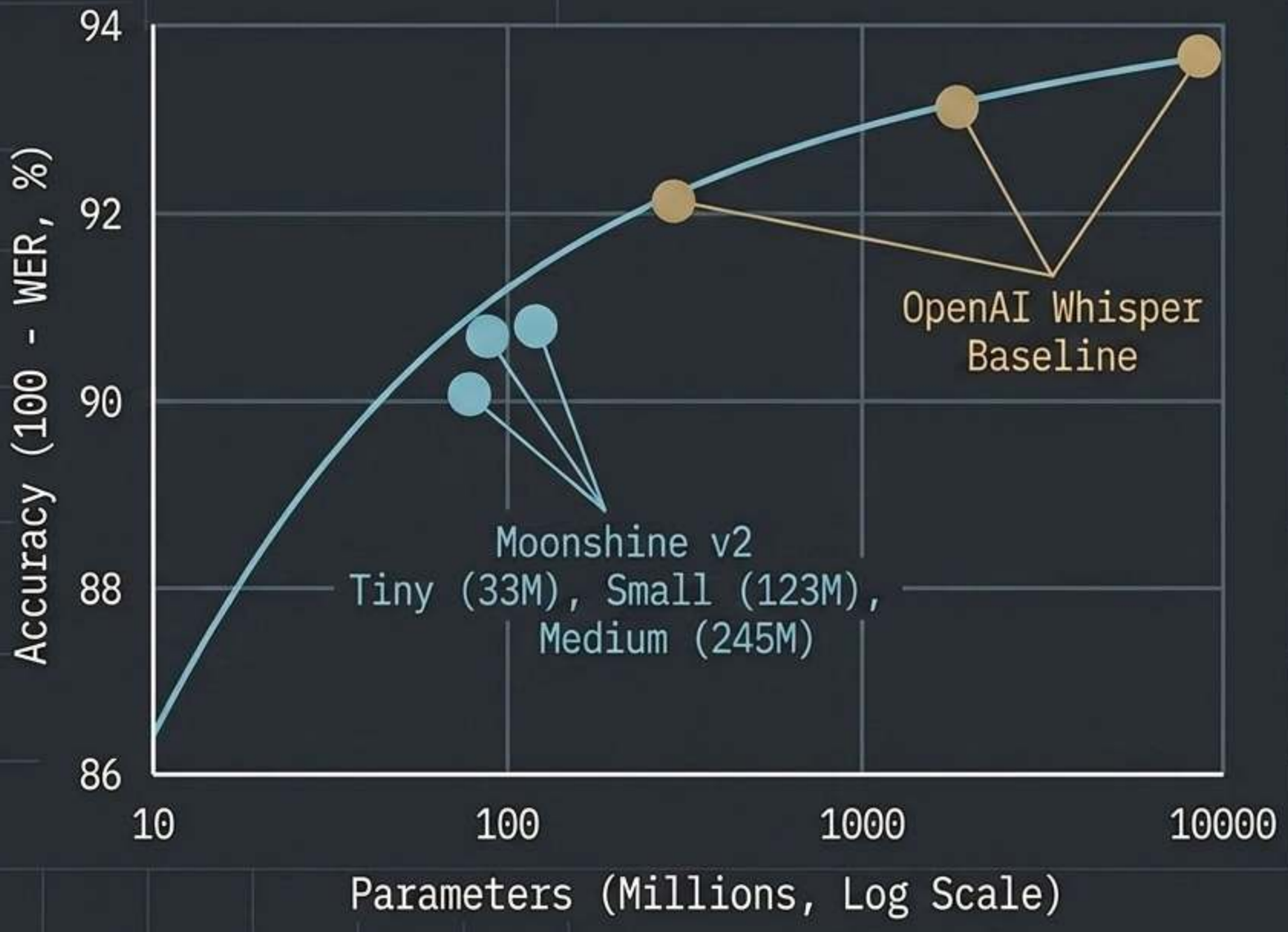
Continuous streaming via provisional and finalised states

The architecture never stalls. Live caption displays instantly render provisional encoder states, which naturally resolve and harden into high-accuracy tokens as the 320ms right-context window naturally fills.

Moonshine v2 breaks the Whisper latency bottleneck



Preserving full-attention accuracy at a fraction of the size



The Optimal Trade-Off

Moonshine v2 attains accuracy on-par with models 6x their size. By optimising for the 0.1–1 TOPS compute envelope, it successfully masters the low-parameter edge quadrant without the multi-GB memory footprint required by GPU-heavy architectures.

Enabling real-time voice AI at the edge

Offline Voice Assistants

Requirement

Zero network round-trip latency; strict memory constraints.

Moonshine Advantage

Ergodic encoder allows continuous, unbounded listening loops on sub-1 GB memory footprints without any TTFT degradation over time.

Live Accessibility Captioning

Requirement

Immediate visual feedback for users; continuous continuous streaming.

Moonshine Advantage

Bounded constant lookahead allows the emission of instant provisional tokens, refining dynamically as right-context audio naturally arrives.

Privacy-Critical On-Device AI

Requirement

High-accuracy transcription that never leaves the hardware.

Moonshine Advantage

Matches the WER of cloud-dependent, full-attention models while running efficiently on 0.1–1 TOPS edge processors without inducing thermal throttling.