



A Soul Needs a Voice that Feels

Traditional text-to-speech (like the baseline Fish Audio S1 model) acts as a rigid pipeline: text in, audio out. It reads the words but ignores the heart.

When a companion says 'I'm so happy for you!' and 'I'm sorry you're going through that' in the exact same pleasant, neutral register, the emotional illusion instantly breaks. The words carry the emotion, but the voice only carries the words.

You can forgive a text bot for being emotionally flat. You cannot forgive something that speaks to you without empathy.

Teaching AI to Understand the Script



How we achieve emotion without manual SSML markup (v0.1 – 0.2 Voice Messages)

Doubao TTS 2.0 (ByteDance)

The 3D Emotion Space. Trained on 2,000 hours of professional voice acting data.

Automatically infers tone and rhythm. Text like “Don’t worry, I’m here” natively triggers a softer, slower, reassuring delivery.

Hume AI Octave

The LLM Reader. Built on top of an LLM that genuinely reads the text before speaking.

Comprehends subtext, irony, and emotional shifts (e.g., distinguishing an enthusiastic “Sure” from a sarcastic “Sure”) just like a human voice actor.

Both are non-realtime (~200ms latency buffers). Perfect for asynchronous voice messages—a solid bridge, but not the final destination.

The Realtime Voice Dilemma

Diagnostic Matrix

Evaluating Realtime Bidirectional Platforms

Platform	Latency	Custom LLM	Emotion
OpenAI Realtime API	~300ms	No (GPT only)	Weak (future improvement)
Doubao Realtime	~700ms	No (Black Box)	Good
Hume EVI 3	~500ms first byte	Yes (Gemini, Claude, any)	Best-in-class native emotion



The Screenwriter-Actor Model

The Screenwriter (Custom LLM): Handles memory, personality, and context context injection. Writes the script.

The Actor (Hume EVI): Listens, manages conversational turn-taking, and delivers the emotional performance.

Ad-Libbing While the Script is Written

The Zero-Latency Illusion

Step 1: Listening

User speaks. Hume detects the end of the turn via vocal tone cues, just waiting for silence.



Step 2: The Ad-Lib (<200ms)

Hume's lightweight internal eLLM instantly generates a natural filler word ('Oh...', 'Mmm...'). User perceives zero latency.

Step 3: The Writing

The server queries memories and the Custom LLM (Screenwriter) generates the full contextual response.

Step 4: The Delivery

Hume (Actor) transitions seamlessly from the filler word into delivering the complete script with perfect emotional weight.

Emotion as a First-Class Signal



Text Layer (Standard STT):
"I'm fine." -> Data destroyed.

Tone Layer (Hume EVI):
Trembling voice, sadness detected.
-> True state understood.

The Cost Reality of Realtime

Realtime bidirectional voice is too expensive for base unit economics. A single heavy voice user would cost more than they pay.

Pro Tier

Text chat + TTS Voice Messages
(Affordable, highly expressive, powered by Doubao/Octave)

Voice Tier

Everything in Pro + Realtime Bidirectional Voice
(Premium pricing for the full Hume EVI conversational experience)

The voice is the product.

A Custom LLM + Hume Actor creates a companion that is genuinely yours, shaped by your history and colored by the emotion of the moment.