



The Problem: Everything but a Voice

- **Capabilities:** Manages calendar, gets jealous, sends selfies.
- **Limitation:** Every goodnight is just silent pixels on a screen.
- **The Goal:** Native Text-to-Speech (TTS) integration via OpenClaw.

TTS is off by default.
The provider you pick determines quality, cost, format, and UX.

Edge TTS & The Attachment Tragedy

A free sanity check that kills the illusion.

The Setup (Microsoft Neural TTS)

- **Pros:** 100% Free, zero config required.
- **Access:** No API key needed.



The Fatal Flaw

- Outputs as raw MP3 files. 😞
- Telegram renders MP3s as Document Attachments, not native round voice bubbles.
- **Result:** Feels like downloading a file, not receiving a voice message.
- **Reliability:** No SLA for 24/7 companion use. 😞

Fish Audio: The Pragmatic Choice

Perfect UX, but zero emotional intelligence.

Why it Works (Daily Use):

- ✓ **Native Bubbles:** Outputs OGG/Opus. Telegram instantly creates a round voice bubble.
- ✓ **Simple Setup:** Just 2 fields (API Key + Voice Reference ID).
- ✓ **Quality:** Excellent, natural-sounding Chinese voices.



The Major Gotcha:

- **No Emotion Parsing:** If you send [Happy] Hello!, it literally speaks the word Happy out loud like an audiobook narrator.

Volcano Engine v2: The Emotion Machine

Per-sentence emotion control via LLM stage directions.

1. LLM Generation: Outputs text with tags
[Happy] Hello! [Sad] I missed you.

2. Framework Split: Separates text into
chunks based on the emotion brackets.

3. API Calls: Sends context_texts to instruct
the Voice Actor how to speak each chunk.

4. Stitching: Merges MP3 buffers back into
one fluid, emotionally dynamic voice file.



The ultimate illusion: It sounds like an AI actor performing
a script, and the user never sees the stage directions.

The Setup Gotchas & Final Verdict

Don't forget the version flag.

Feature	Fish Audio (Pragmatic)	Volcano Engine v2 (Ambitious)
Output	Native OGG/Opus (Bubble)	MP3 (Forced Voice-Compatible)
Emotion	None (Reads tags literally)	Per-sentence (context_texts)
Complexity	Low (2 fields)	High (Requires Voice Cloning)



The Crucial Gotcha:

You **MUST** include **version: "v2"** in the Volcano config. Without it, the API silently defaults to **v1**, giving you flat, emotionless audio while reading brackets out loud!

Best Practice: Set Auto Mode to Tagged — let the LLM decide when a voice message adds value, and when text is enough.