

The Chinese STT Dilemma

Why standard models stumble on everyday speech

The Input: Casual speech, dialects, colloquialisms, and filler words.

Standard STT
(e.g., Whisper)



吃我不
我是
不是的
楚

The Output:
Mangled sentences, wrong
homophones, and entirely
dropped clauses.



The Verdict:
Formal models struggle with
the messy reality of
conversational Chinese audio.

Doubao STT: The Seed ASR Solution

Perfecting casual Mandarin and dialects at a fraction of the cost

	Standard STT	Doubao (Seed ASR)
Mandarin Accuracy	Formal only	Excellent (Casual + Filler)
Dialect Support	Limited	Strong Regional Coverage
Cost	Baseline	~2.5x Cheaper
Timestamps	Per-segment	Per-utterance



Navigating the Console Maze

Finding the correct Bigmodel service among identically named traps



Out With the Old (API)

Stripping away V2 legacy cruft for V3 REST simplicity

Toss in the Trash:

🗑️ The “Secret Key”
(Not needed for V3!)

🗑️ Bearer Tokens

🗑️ WebSocket Endpoints
(Will return 400s)

🗑️ Old V2 Documentation

Keep for V3 REST:

✨ APP ID
(DOUBAO_STT_APP_ID)

✨ Access Token
(DOUBAO_STT_ACCESS_KEY)

The Async Submit-Then-Poll Flow

Do not panic when you see empty braces

Step 1: Submit (POST)

Send audio + X-API-Request-Id (UUID).

Gotcha: Returns HTTP 200 with {}. The UUID is your Task ID.

Step 2: Query (POST)

Poll using the exact same UUID.

Gotcha: Returns {} continuously while processing.
No 'in-progress' status.

Step 3: Success!

Wait 2-4 seconds...

The full JSON transcript suddenly arrives. Keep polling!

