

What can Voice AI do?

Ling Lin

Goal: To identify the best AI software/APIs for our project.

Emotion Analysis

Tools that understand how something is said, not just what is said.

Hume AI

What it does

It is the industry leader in "Prosody Analysis." It analyzes vocal tone, pitch, rhythm, and timbre to detect 53+ distinct emotions (e.g., Sarcasm, Desire, Awkwardness, Distress).

Why it matters for us

It provides real-time data values (e.g., Anger: 0.8) that can be fed into platform to create reactive effects.

Pros

Unique capability; provides rich emotional data.

Cons

Newer API, requires learning their specific WebSocket protocol.

Text-to-Speech (TTS) & Generation

Tools that generate realistic, human-like AI voices.

ElevenLabs

What it does

The current gold standard for AI voice generation. It offers high-fidelity, emotionally expressive voices and "Voice Cloning" (copying a specific actor's voice).

Why it matters for us

Essential if our AI character needs a specific personality (e.g., a terrifying god, a whispering lover, or a child).

Pros

Best audio quality; highly customizable.

Cons

Can have slight latency (1-2 seconds) in live generation.

OpenAI TTS (Text-to-Speech)

What it does

Generates clean, standard AI speech very quickly.

Why it matters for us

Good for projects where speed is more important than unique character acting.

Pros

Fast, reliable, cost-effective.

Cons

Only 6 preset voices; cannot clone voices or control extreme emotions.

Speech-to-Text

Tools that convert the actor's speech into text for the system to process.

OpenAI Whisper

What it does

Converts speech to text with incredibly high accuracy.

Why it matters for us

It works well even in a noisy environment.
It handles accents and mumbled speech very well.

Pros

Most accurate model available.

Cons

Standard API has slight latency.

Deepgram

What it does

The fastest speech-to-text API in the world.

Why it matters for us

If we do a Live Improv or rapid-fire interaction, we need the AI to respond instantly (under 300ms). Deepgram is built for this.

Pros

Extremely low latency (Speed).

Cons

Slightly less accurate than Whisper in complex contexts.

"All-in-One": Real-time Conversation

Emerging tech that handles listening, thinking, and speaking in one stream.

OpenAI GPT-4o Realtime API

What it does

A new "End-to-End" model. It hears audio and outputs audio directly without converting to text first.

Why it matters for us

It allows for interruptions. The actors can cut the AI off, and it will stop talking instantly.

Pros

It feels the most "human."

Cons

Expensive; acts as a "Black Box" (harder to extract specific data).

Creative Transformation

Tools for creative transformation and visual performance.

RVC (Retrieval-based Voice Conversion)

What it does

Real-time AI Voice Changing. Unlike standard pitch-shifters, RVC uses deep learning to completely replace the actor's vocal timbre with another identity (e.g., changing a male actor into a young girl, a monster, or a specific celebrity) while keeping the original emotion and intonation.

Why it matters for us

An actor can play multiple characters simply by switching the AI model.
The character might be "possessed" by the AI, and their voice physically changes in real-time.

Pros

Extremely low latency (can be used live); Open Source (Free); High fidelity.

Cons

Requires a decent GPU to run locally; Setup is more technical than a web API.

NVIDIA Audio2Face

What it does

Generates expressive 3D facial animation solely from an audio source. You feed it audio (from an actor or AI), and it animates a 3D character's mouth, eyes, and emotions to match perfectly.

Why it matters for us

Can project a giant digital avatar on stage. An actor backstage (or the AI) speaks, and the avatar performs the lines instantly without wearing a motion capture suit.

Pros

Industry-standard quality; works with Unreal Engine and Unity; No complex mocap gear needed.

Cons

Requires an NVIDIA RTX Graphics Card; Setup involves 3D software knowledge.