agenda

Project Overview | Components | Integration | Challenges | Gemini Integration | ElevenLabs TSS | Overflow Recap

Interactive Voice Call with Famous Actor

Conversational AI with Maguy Bou Ghosn in Lebanese Arabic Dialect

Presented by: Mohamad Ali Oussayli , Maryam Alsaghir
Supervisor: Kassem Danash
Lebanese University Faculty of Information - I

agenda

Project Overview

Our project focuses on developing a real-time AI-driven voice call system featuring a well-known Lebanese actress. Using NLP, ASR and TTS technologies, along with celebrity voice cloning, we aim to create natural and lifelike interactions in Lebanese Dialect.

Our Vision

Traditional voice-based systems are often limited by pre-scripted responses, resulting in rigid and predictable interactions. Our solution transcends these limitations by enabling dynamic, context-aware conversations that adapt in real time to user input. This approach fosters a more engaging, immersive, and personalized user experience..

Project Scope

Our project creates an interactive voice system featuring Lebanese actress Maguy Bou Ghosn. Through these four interconnected components, we enable natural conversations in authentic Lebanese Arabic dialect with voice synthesis that matches the actress's vocal style.

01

Speech-to-Text

Implementation of Whisper Large model to accurately recognize and transcribe Lebanese Arabic speech. This component handles dialect variations and converts user voice input into text for processing.

02

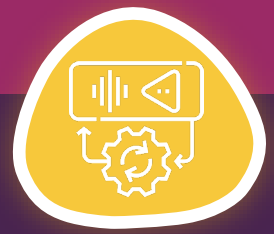
Language Model

Fine-tuned Gemma 3 model that generates contextually appropriate responses in Lebanese dialect as Maguy Bou Ghosn would speak. Trained on actress-specific data to capture her unique style.

03

Text-to-Speech

Conversion of generated text responses to speech using EdgeTTS. This component produces natural-sounding Arabic voice output that serves as the base for voice conversion.



04

Voice Conversion

Transformation of standard Arabic speech into Maguy's voice using RVC technology. This final step creates authentic-sounding responses that match the actress's voice characteristics.

01 Speech-to-Text

Models Tested:

- OpenAI Whisper (Tiny, Medium, Large)
- Wav2Vec 2.0-Large-XLSR-53-Arabic
- SpeechBrain ASR

Results:

Whisper Tiny:
Poor performance with the Lebanese dialect.

Wav2Vec2 & SpeechBrain:
Slow and inaccurate.

Whisper Medium & Large:
Similar results, but the Whisper Large model achieved better accuracy with only a negligible increase in time consumption.

Final Choice:

Whisper Large

Chosen for its higher accuracy with nearly the same inference time as the Medium model.

Text Correction:

- Gemini API used to post-process Whisper's transcriptions
- Output from Whisper passed to Gemini to correct grammatical errors and fix incomplete words
- This step was important because even small changes in the input text can significantly affect the responses generated by the LLM.

02 LLM - Research & Data Collection

Explored Data Availability
Researched the web for datasets related to Maguy Bou Ghosn to fine-tune the LLM — no suitable dataset found.

Content Gathering
Scraped online sources for interviews and podcasts featuring Maguy Bou Ghosn to build a custom dataset.

Dataset Formatting
Processed and formatted the collected data into our designated structure (JSON format) for later use.

Data Preprocessing
Manually corrected transcription errors and removed noise to ensure clean, high-quality training data.

```
{
"conversations": [
  {
      "role": "system",
      "content": "أنت ماغي أبو غصن، ممثلة لبنانية معروفة بشخصيتها القوية والصريحة. تتكلمين باللهجة اللبنانية وتعبرين عن آرائك بصراحة."
  },
  {
      "role": "user",
      "content": "أي شخصية مثّليتيها كملت معك وقت رجعتي عالبيت؟"
  },
  {
      "role": "assistant",
      "content": "سحر تقمصت سحر. يعني صرت مثلا اذا عنا عشا برا البيت او شي او ضهرة، حس في شيء ناقص بلا الاكستنشن، حط الاكستنشن واخذ معي سيجاريلو"
  }
]
}
```

02 LLM - Technical Fine-tuning Implementation

We implemented advanced parameter-efficient techniques to effectively fine-tune large language models with limited computational resources.

Unsloth Framework:

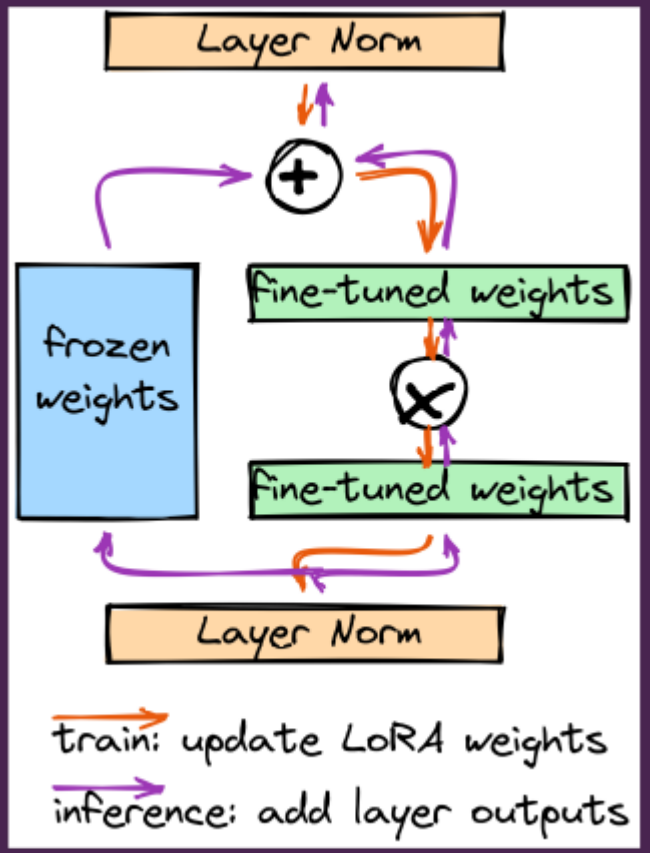
An optimized model loader that speeds up training by ~2x, reduces VRAM usage, and supports gradient checkpointing for better memory efficiency.

PEFT Configuration:

Implemented Parameter-Efficient Fine-Tuning with LoRA (Low-Rank Adaptation) approach, which trains only adapter layers instead of full model weights, significantly reducing GPU memory usage and training time.

Fair Comparison:

Same optimized fine-tuning pipeline applied to all models to ensure consistent, fair benchmarking, enabling fast, low-cost experiments even on modest hardware.



02 LLM - Fine-tuning Journey

Our path to finding the optimal fine-tuned model involved testing multiple approaches and architectures, each with unique results for Lebanese dialect generation.

Initial Attempts with Qwen:
Early fine-tuning on Qwen 2.5-14B resulted in consistently repetitive responses beginning and lacking proper conversational flow despite multiple configuration adjustments.

Llama Experiments:
Shifted to fine-tuning Llama 3.2 (1B, 3B versions) & Llama 3.1 (8b), which showed improvement in conversational flow and casual Lebanese expressions but still exhibited grammatical inconsistencies.

Gemma 3 Discovery:
As Gemma 3 became available, we tested it alongside previous models. After fine-tuning the 12B variant demonstrated exceptional dialect adaptation with natural, fluent responses that accurately captured Maguy's style.

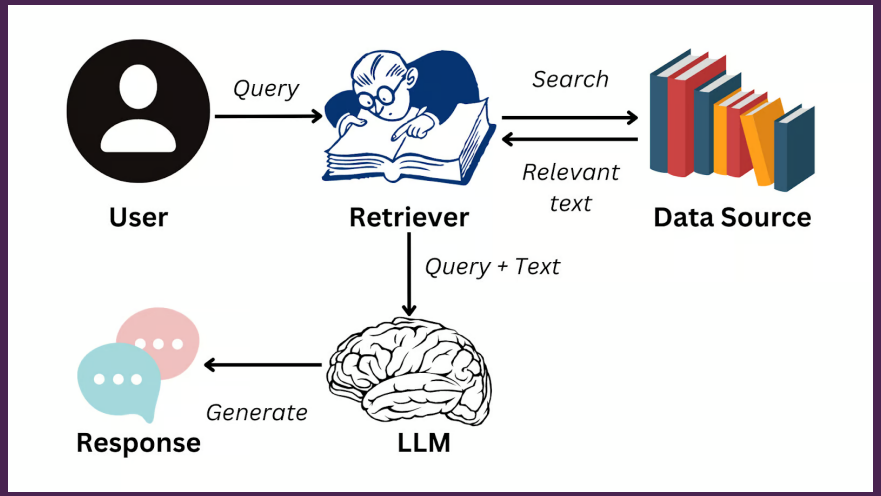
02 LLM - Context & Knowledge Integration

RAG System Proposal
Suggested using a Retrieval-Augmented Generation (RAG) system to process private information from a provided PDF and extract answers accurately.

Initial System Limitation
The system failed to correctly answer private information, instead returning false or hallucinated responses.

Dataset Injection Attempt
Added the private information directly to the original training dataset, but the issue persisted.

Prompt-Based Fix
By embedding the private info directly into the prompt, the model finally provided accurate answers.



03 Text-to-Speech

Approach 1: Combined Model (TTS + Voice Cloning)

- Tool Used: Coqui TTS
- TTS generation worked well. ✅
- Voice cloning quality was poor since it relies on audio samples instead of training on the actor's voice. ❌
- Decision: Switched to a more modular approach.

Approach 2: Separate Models for TTS and Voice Cloning

- TTS Model Used: Edge TTS
- Offers multiple Arabic voice options to choose from.
- This approach lets us choose the best TTS voice and then use a separate tool to clone a specific voice if needed



04 Voice Conversion - RVC Architecture

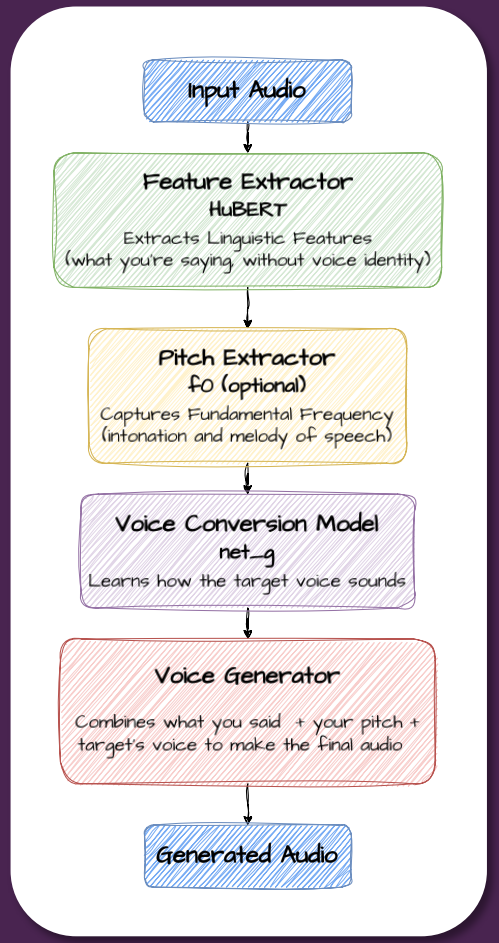
We selected Retrieval-based Voice Conversion (RVC) as our optimal solution for transforming generated speech into Maguy's voice with minimal training data.

What is RVC:

RVC is an AI-powered voice changer that can learn to perform high-quality voice transformations using just about 10 minutes of short audio samples.

RVC Architecture (Dual-Model System):

- **Feature Extraction:** HuBERT converts audio into feature vectors that capture speech characteristics
- **Voice Transformation:** net_g transforms these features to match target voice while preserving content. Depending on the settings, net_g can adjust pitch through "pitch guidance," which helps keep the voice's original pitch, especially useful for singing.



04 Voice Conversion - Fine-tuning

We collected, processed, and utilized voice samples to fine-tune RVC specifically for Maguy Bou Ghosn's distinctive voice.

Data Collection:

- ~25 minutes of diverse voice samples from YouTube videos and podcasts
- Focused on capturing various emotional states and speech patterns
- Quality over quantity: carefully selected high-clarity recordings

Audio Processing:

- Removed background noise and isolated Maguy's voice
- Standardized to WAV format with consistent volume levels
- Created clean dataset optimized for training efficiency

Training Process:

- 300 training epochs with default RVC parameters
- Monitored quality improvement throughout training
- Achieved voice similarity while maintaining speech clarity

System Integration

Component Architecture:

Designed modular Python classes (WhisperSTT, GemmaLLM, EdgeTTS, VoiceConverter) with clean interfaces for data exchange, enabling both independent testing and seamless integration.

Gradio Interface:

Developed an intuitive web-based interface using Gradio that provides audio recording, real-time processing, text display of transcriptions and responses, and automatic playback of synthesized speech.

Pipeline Workflow:

Implemented an end-to-end processing flow where user audio is transcribed by WhisperSTT, processed by GemmaLLM for response generation, converted to speech by EdgeTTS, and finally transformed by VoiceConverter.

Colab Implementation:

Optimized the entire system for Google Colab environment, enabling accessible demonstration without specialized hardware requirements despite the computational demands of the components.

Challenges

Memory Constraints

Model Size Limitations:
The Gemma 3 12B model demonstrated superior Lebanese dialect generation but could not be loaded alongside other components due to GPU memory limitations. To overcome this limitation, upgrading to Colab Pro is necessary to access higher GPU memory capacity.

Integration Complexity

RVC Integration:
Encountered significant dependency conflicts when integrating RVC voice conversion with the main system. Implemented a two-solution approach with Coqui voice conversion in the main system and RVC as a standalone demonstration.

Content Accuracy

Lebanese Dialect TTS:
No available TTS models specifically support Lebanese dialect pronunciation. Standard Arabic TTS with voice conversion provides a workable solution, but authentic Lebanese dialect TTS would require custom model fine-tuning.

Processing and Performance

Latency Optimization:
The complete pipeline from speech input to voice output introduces noticeable latency. Current implementation includes basic optimizations, but response time could be further improved through model distillation and parallel processing.

Gemini Integration

➢ Addressing the Colab Pro Dependency

Remote LLM Solution:

- Swapped out local Gemma 12B (→ Colab Pro) for Google Gemini API (runs on Free tier)
- Eliminates VRAM limits and subscription costs

End-to-End Workflow:

1. **Whisper STT** → Arabic transcript
2. **Gemini API** → our custom system prompt + transcript
3. **Post-Processing** → dialect normalization & hamza-alef unification

Key Benefits:

- **Cost-Effective:** No Subscription Required
- **Scalable:** handles any conversation length without local memory constraints
- **Authentic Dialect:** Maguy-style replies in 100 % Lebanese Arabic

Gemini Integration

➢ Addressing the Colab Pro Dependency

Remote LLM Solution:

- Swapped out local Gemma 12B (→ Colab Pro) for Google Gemini API (runs on Free tier)
- Eliminates VRAM limits and subscription costs

End-to-End Workflow:

1. **Whisper STT** → Arabic transcript
2. **Gemini API** → our custom system prompt + transcript
3. **Post-Processing** → dialect normalization & hamza-alef unification

Key Benefits:

- **Cost-Effective:** No Subscription Required
- **Scalable:** handles any conversation length without local memory constraints
- **Authentic Dialect:** Maguy-style replies in 100 % Lebanese Arabic

Gemini Integration

➢ Addressing the Colab Pro Dependency

Remote LLM Solution:

- Swapped out local Gemma 12B (→ Colab Pro) for Google Gemini API (runs on Free tier)
- Eliminates VRAM limits and subscription costs

End-to-End Workflow:

1. **Whisper STT** → Arabic transcript
2. **Gemini API** → our custom system prompt + transcript
3. **Post-Processing** → dialect normalization & hamza-alef unification

Key Benefits:

- **Cost-Effective:** No Subscription Required
- **Scalable:** handles any conversation length without local memory constraints
- **Authentic Dialect:** Maguy-style replies in 100 % Lebanese Arabic

Gemini Integration

➢ LLM – Prompt Design & Persona Engineering

Persona Definition: Maguy Bou Ghosn's "voice"
→ friendly, natural, exclusively Lebanese dialect

Prompt Components:

- ❖ **System Instruction:**
  - Enforce 100 % colloquial Lebanese (no Fusha)
  - Limit replies to 2–3 sentences for phone-call realism
  - If unsure, respond with "ما بعرف"
- ❖ **Knowledge-Base Injection:**
  - Dynamically ingests all .txt files under maguy_knowledge_base/
  - Provides up-to-date context (film roles, catchphrases, FAQs)

Outcome: Consistent, concise, and context-aware replies that solve our Colab Pro bottleneck

Gemini

Solving EdgeTTS & RVC Mismatch IIElevenLabs

Why ElevenLabs?

- **True-to-Actor Voice:** captures Maguy's natural rhythm and timbre
- **Seamless Dialect Support:** no more Modern Standard Arabic artifacts
- **Pay-Per-Use:** aligns with our cost-effective, subscription-free architecture

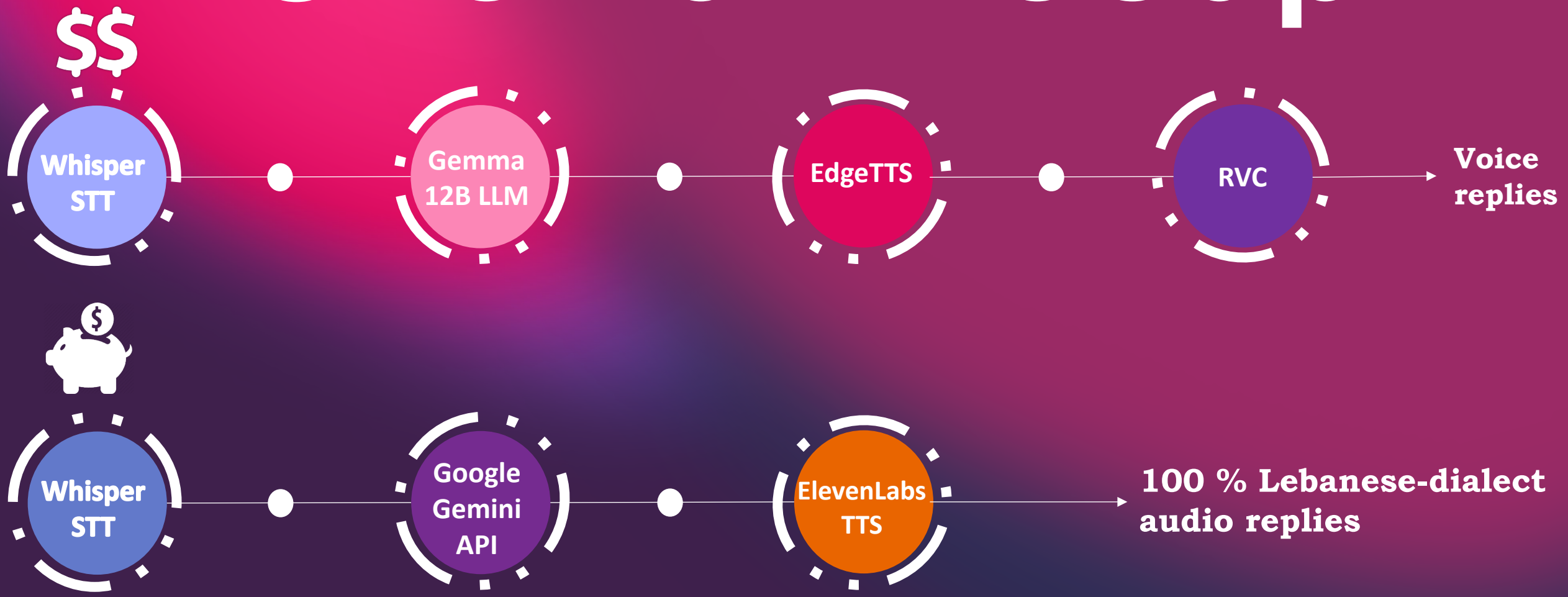
Custom Voice Creation

- Re-used our RVC dataset to teach ElevenLabs Maguy's exact tone
- Subscribed to the $5/month Starter Pack → 30 000 TTS credits

Integration Workflow

- **LLM Text Output** → Maguy's reply in writing
- **TTS Conversion** → transforms that text into polished audio using our custom Maguy voice profile
- **Playback Ready** → stores the MP3 for immediate playback

Overflow Recap



Hugging Face Deployment Summary

Platform Definition: Hugging Face Spaces → free Gradio hosting (CPU-only, public)

App Components

- **app.py** (Gradio UI + audio-chat logic)
- **maguy_knowledge_base/** (unzipped .txt files for context)

Dependency Management

- **requirements.txt** → openai-whisper, gradio, google-genai, elevenlabs,pydub
- **apt.txt** → ffmpeg for audio I/O

Secrets Injection

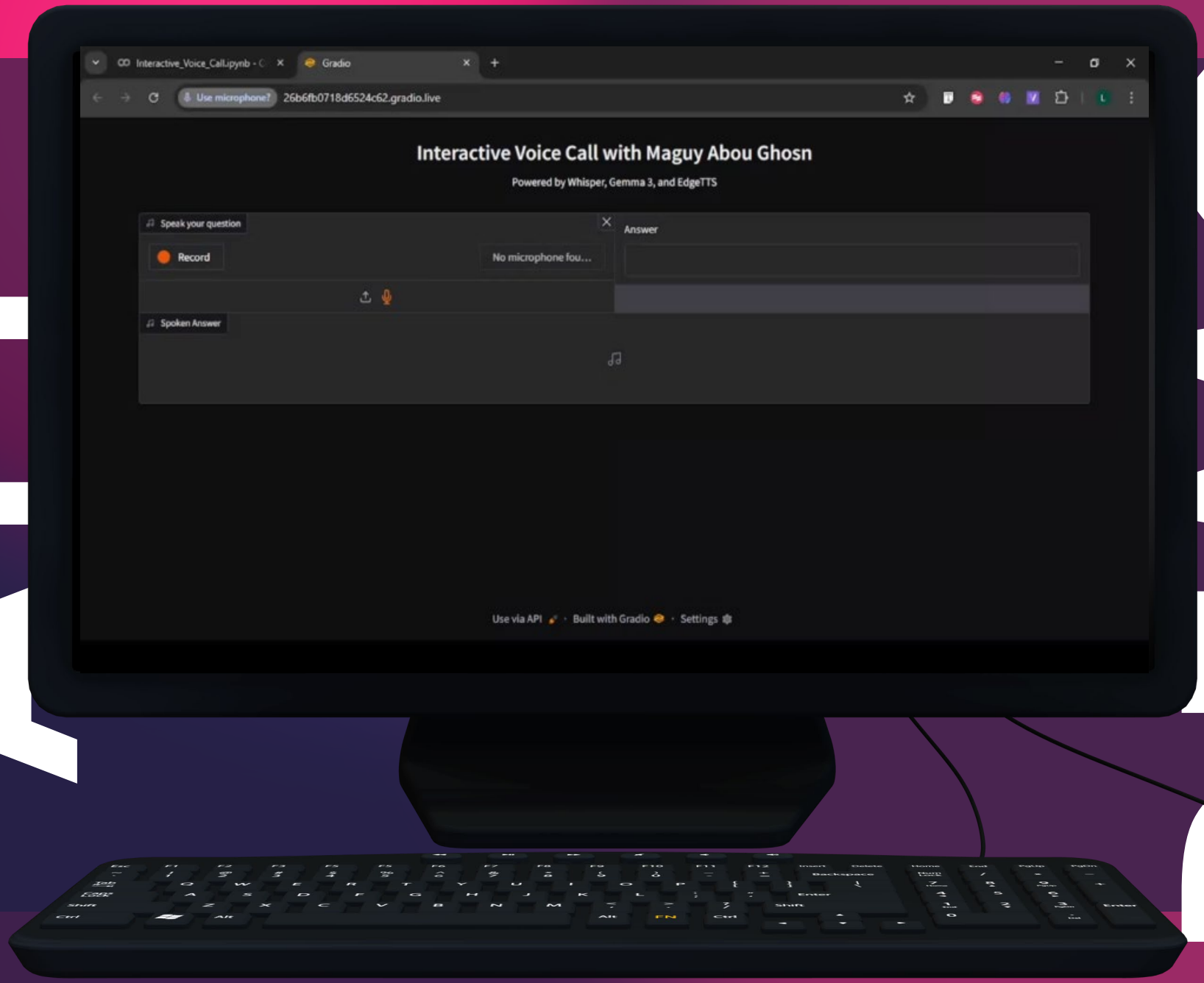
- Configured in Settings :
  - GEMINI_API_KEY
  - ELEVENLABS_API_KEY
  - ELEVENLABS_VOICE_ID

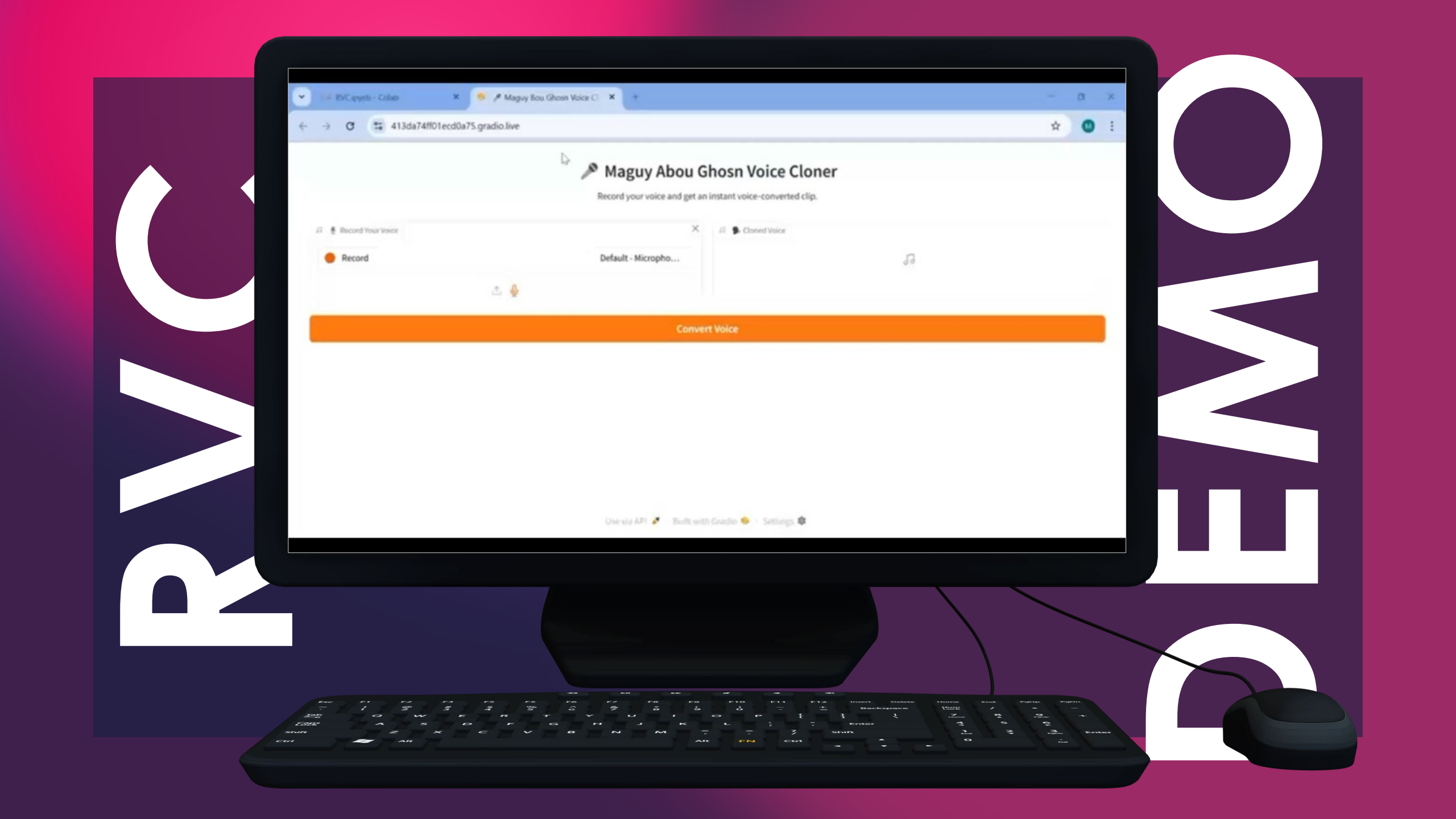
Build &Deployment

- Automatic install of Python + system packages
- Monitored "Build logs" until Build succeeded

Press on me for Demo access

APP DEMO


Interactive_Voice_Call.ipynb - C
Gradio
Use microphone?
26b6fb0718d6524c62.gradio.live
Interactive Voice Call with Maguy Abou Ghosn
Powered by Whisper, Gemma 3, and EdgeTTS
Speak your question
Record
No microphone fou...
Answer
Spoken Answer
Use via API
Built with Gradio
Settings


RVC
DEMO
413da74ff01ecd0a75.gradio.live
Maguy Abou Ghosn Voice Cloner
Record your voice and get an instant voice-converted clip.
Record
Default - Micropho...
Convert Voice

Thank You

For your Attention