Contents

[1. Core Architecture (Reusable Framework) 2](#_Toc206408473)

[1.1. Input layer 2](#_Toc206408474)

[1.2. RAG pipeline 2](#_Toc206408475)

[1.3. Output layer 2](#_Toc206408476)

[2. Dataset Swapping for Each Character 2](#_Toc206408477)

[3. Roadmap 3](#_Toc206408478)

[Phase 1 — Core MVP (2-3 weeks) 3](#_Toc206408479)

[1. Collect baseline dataset 3](#_Toc206408480)

[2. Data preprocessing 3](#_Toc206408481)

[3. Build RAG pipeline: 3](#_Toc206408482)

[4. Basic web UI 3](#_Toc206408483)

[Phase 2 — Add Speech I/O (2 weeks) 3](#_Toc206408484)

[5. Integrate STT model 3](#_Toc206408485)

[6. Integrate TTS model 3](#_Toc206408486)

[7. Update UI 3](#_Toc206408487)

[Phase 3 — Persona Tuning (2-3 weeks) 3](#_Toc206408488)

[8. Add style control in prompts 3](#_Toc206408489)

[9. Fine-tune TTS 3](#_Toc206408490)

[10. Test 3](#_Toc206408491)

[Phase 4 — Customized Version (2 weeks) 3](#_Toc206408492)

[11. Script for automatic vector DB rebuild 3](#_Toc206408493)

[12. Script for TTS voice model training 3](#_Toc206408494)

[Phase 5 — Optional Video Avatar (1-2 weeks) 4](#_Toc206408495)

[14. Integrate lip-sync video generation API 4](#_Toc206408496)

[15. Deploy 4](#_Toc206408497)

[4. Timeline 4](#_Toc206408498)

# Core Architecture (Reusable Framework)

## Input layer

* Text input → directly goes to RAG pipeline
* Voice input → **STT model** (e.g., OpenAI Whisper) → text → RAG

## RAG pipeline

* Preprocess + store the character’s text corpus
* Use embeddings (e.g., e5-large, OpenAI text-embedding-3-large)
* Retrieval + LLM answer generation with character’s tone/style
* Prompt engineering for persona consistency

## Output layer

* Text → TTS model (e.g., ElevenLabs, Coqui TTS, Bark) for character voice
* Optional: Sync with video avatar

# Dataset Swapping for Each Character

For each new character:

* Replace **text knowledge base** (books, speeches, transcripts, articles)
* Replace **voice samples** for TTS training/fine-tuning
* Replace **prompt persona**
* Replace **optional video avatar** resources

# Roadmap

## Phase 1 — Core MVP (2-3 weeks)

**Goal:** End-to-end pipeline with text in → text out (RAG only)

* + 1. Collect baseline dataset (speeches, interviews, books)
    2. Data preprocessing (cleaning, splitting, chunking, metadata tagging)
    3. Build RAG pipeline:
* Choose embedding model
* Store vectors in FAISS (or cloud alternative)
* Retrieval + LLM
  + 1. Basic web UI (text in / text out)

## Phase 2 — Add Speech I/O (2 weeks)

**Goal:** Voice in → voice out

* + 1. Integrate STT model
    2. Integrate TTS model (start with generic voice, later fine-tune with character’s voice)
    3. Update UI for microphone input + audio playback

## Phase 3 — Persona Tuning (2-3 weeks)

**Goal:** Make it *feel* like the character

* + 1. Add style control in prompts (tone, vocabulary, phrasing)
    2. Fine-tune TTS with **5–10 mins of real audio** for character voice
    3. Test with real audience for accuracy & tone

## Phase 4 — Customized Version (2 weeks)

**Goal:** Easily swap datasets for new characters

* + 1. Script for automatic vector DB rebuild from new text sources
    2. Script for TTS voice model training from uploaded audio
    3. Documentation + dashboard for dataset/voice updates

## Phase 5 — Optional Video Avatar (1-2 weeks)

**Goal:** Show a speaking face

* + 1. Integrate lip-sync video generation API (D-ID, Wav2Lip)
    2. Deploy with lightweight streaming

# Timeline

|  |  |
| --- | --- |
| **Phase** | **Estimated Duration** |
| Core MVP | 2-3 weeks |
| Add Speech I/O | 2 weeks |
| Persona Tuning | 2-3 weeks |
| Customized Version | 2 weeks |
| Video Avatar | 1-2 weeks |