DocuMind Voice - Complete Project Summary

DocuMind Voice - An AI-powered multimodal assistant that reads, understands, and speaks about documents.

----

1. Core Concept

Text Mode: User types query -> RAG system retrieves and answers using Llama-3.

Voice Mode: User speaks -> Whisper converts speech -> RAG processes -> Coqui generates spoken reply.

----

2. Tech Stack

Frontend: React / TailwindCSS

Backend: Flask (REST API)

STT: Faster-Whisper

TTS: Coqui TTS (VITS)

RAG Engine: MiniLM + Groq Llama-3.1-8B

Table Extraction: Camelot / Tabula

Diagram Understanding: Donut (NAVER AI) / LayoutLMv3

Vector DB: ChromaDB

Cache: Redis

Auth: Supabase / Flask-Login

Deployment: DigitalOcean + Netlify + Railway

CI/CD: CircleCI

----

3. Architecture

User (Text / Voice)

Frontend (React + Tailwind)

Flask Backend

/transcribe -> Whisper

/query -> RAG Engine

/speak -> Coqui

/upload -> PDF/Table/Diagram Processor

Vector DB (ChromaDB + Redis)

Response (Text + Audio)

----

4. Phased Development

Phase 1: Core Voice RAG MVP - 13 days

Phase 2: Table Understanding - 5 days

Phase 3: Diagram Understanding - 8 days

Phase 4: Multimodal Fusion - 5 days

Phase 5: Frontend Revamp - 8 days

Phase 6: Deployment & Scaling - 5 days

Phase 7: Productization - 3 days

Total: ~47 days (~7 weeks)

----

5. Key Features

- Voice Interface (Whisper + Coqui)
- Smart Retrieval (RAG with Llama-3)
- Table Parsing (Camelot)
- Diagram Reading (Donut, LayoutLMv3)
- Multi-Document Support
- Spotify-style UI with Waveform
- Redis Session Memory
- Cloud Deployment

----

6. Cost Breakdown (Monthly)

DigitalOcean Backend	$6
Railway DB + Redis	$5
Netlify / Vercel Frontend	Free
Hugging Face Model Hosting	$9
Domain + SSL	$1
Total Monthly	$16

----

7. UI Summary

- ChatGPT-style chat interface
- Spotify-style horizontal audio player
- Sidebar for uploaded files
- Light/Dark mode toggle
- Animated mic button

----

8. Target Users

- Students & Researchers
- Educators & Trainers
- Knowledge Workers
- Legal / Medical Professionals
- Enterprises

----

9. Monetization

Free: Text + Voice Q&A (Single Doc)

Pro: Table + Diagram Comprehension

Enterprise: API + Private Hosting

----

10. Outcomes

- Fully functional multimodal AI app

- Clean, responsive frontend
- Modular backend
- Deployed on scalable infrastructure

----

11. Timeline

Total Duration: 45-50 days (~7 weeks)

MVP: 2 weeks

Full Product: 6-7 weeks

----

12. Investment

One-time: ~$12 (Domain)

Monthly: ~$16 (Hosting + API)

Developer Time: ~45 days

----

13. Future Enhancements

- Real-time streaming (WebSocket)
- Multi-language voice support
- Mobile app (React Native)
- API monetization

Final Goal:

A portfolio-grade AI product that demonstrates full-stack, multimodal, and design excellence.