AI Based Audio to Text and Speech Analysis / LLM

1. Product Overview

The AI Based Audio to Text and Speech Analysis system is an enterprise-ready AI-powered solution designed to convert spoken audio into accurate textual data and perform advanced speech analytics. It enables organizations to process voice recordings, live calls, and meetings to extract meaningful insights such as sentiment, emotion, intent, and keywords. The solution improves productivity, reduces manual listening effort, and enables data-driven decision-making from unstructured audio content.

2. Objective

- To build an AI-powered speech-to-text and speech analysis platform.
- To provide accurate transcription of audio and voice recordings.
- To analyze speech for sentiment, emotion, and intent.
- To reduce manual dependency on audio review processes.
- To enable scalable and secure audio analytics for enterprises.

3. Key Features

- Automatic Speech-to-Text (ASR) conversion
- Multi-language and accent support
- Speaker diarization and identification
- Sentiment and emotion analysis
- Keyword extraction and topic detection- Timestamped transcription output

4. Technical Stack

Layer	Technology
Frontend / UI	Web Portal (React)
Backend API	FastAPI (Python, Dockerized)
Speech Processing	ASR Models (Whisper / Speech APIs)
AI / NLP Layer	LLM / NLP Models (Transformers)
Vector Database	Amazon OpenSearch (Optional)
Audio Storage	Amazon S3
Authentication	IAM / OAuth / SSO
Monitoring	AWS CloudWatch

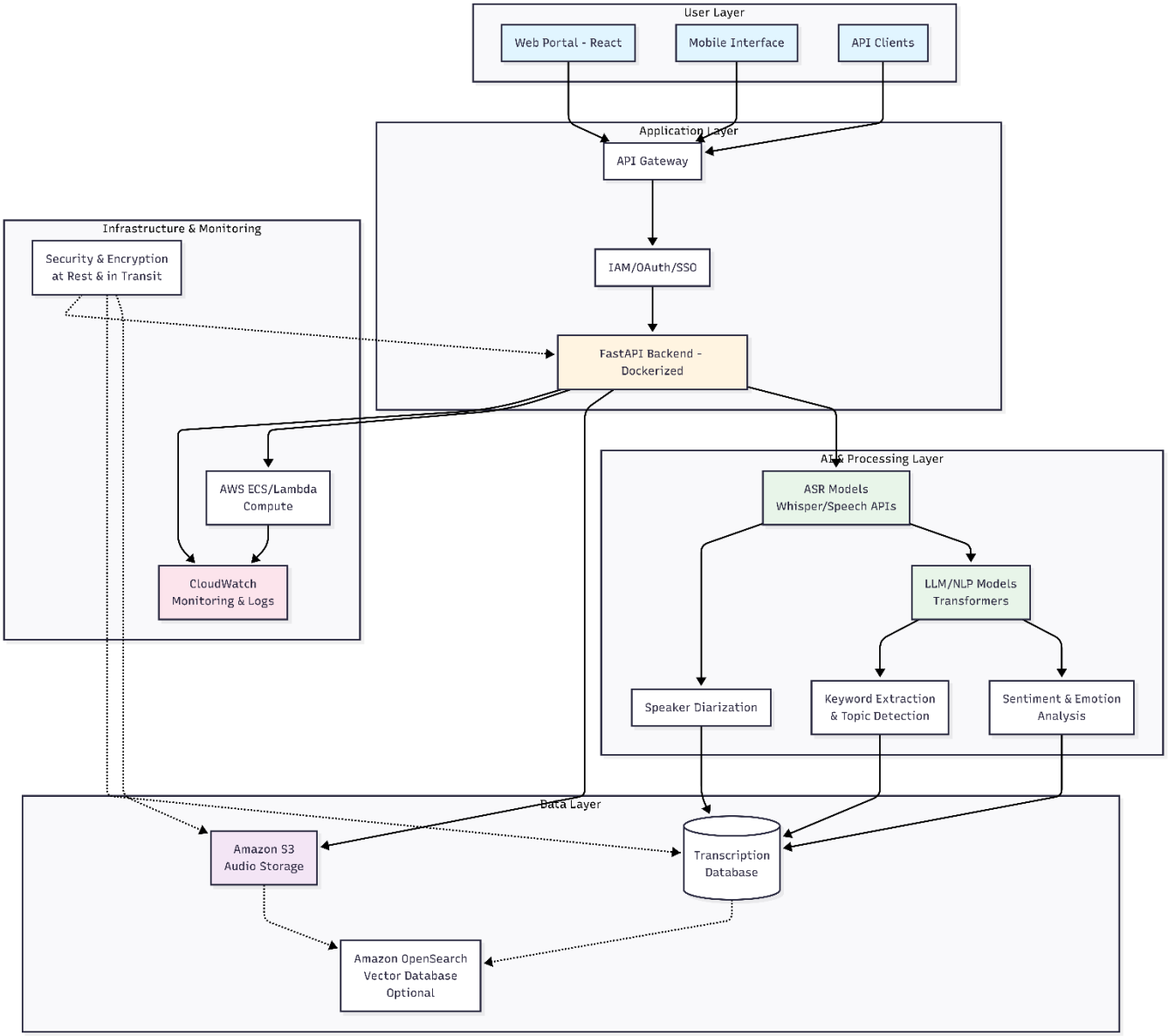
5. Security & Governance

- Internal and authorized access only
- Role-based access control
- Encryption of audio and transcription data at rest and in transit
- Audit logging of transcription and analysis requests- AI models restricted to approved data sources

Architecture Diagram -

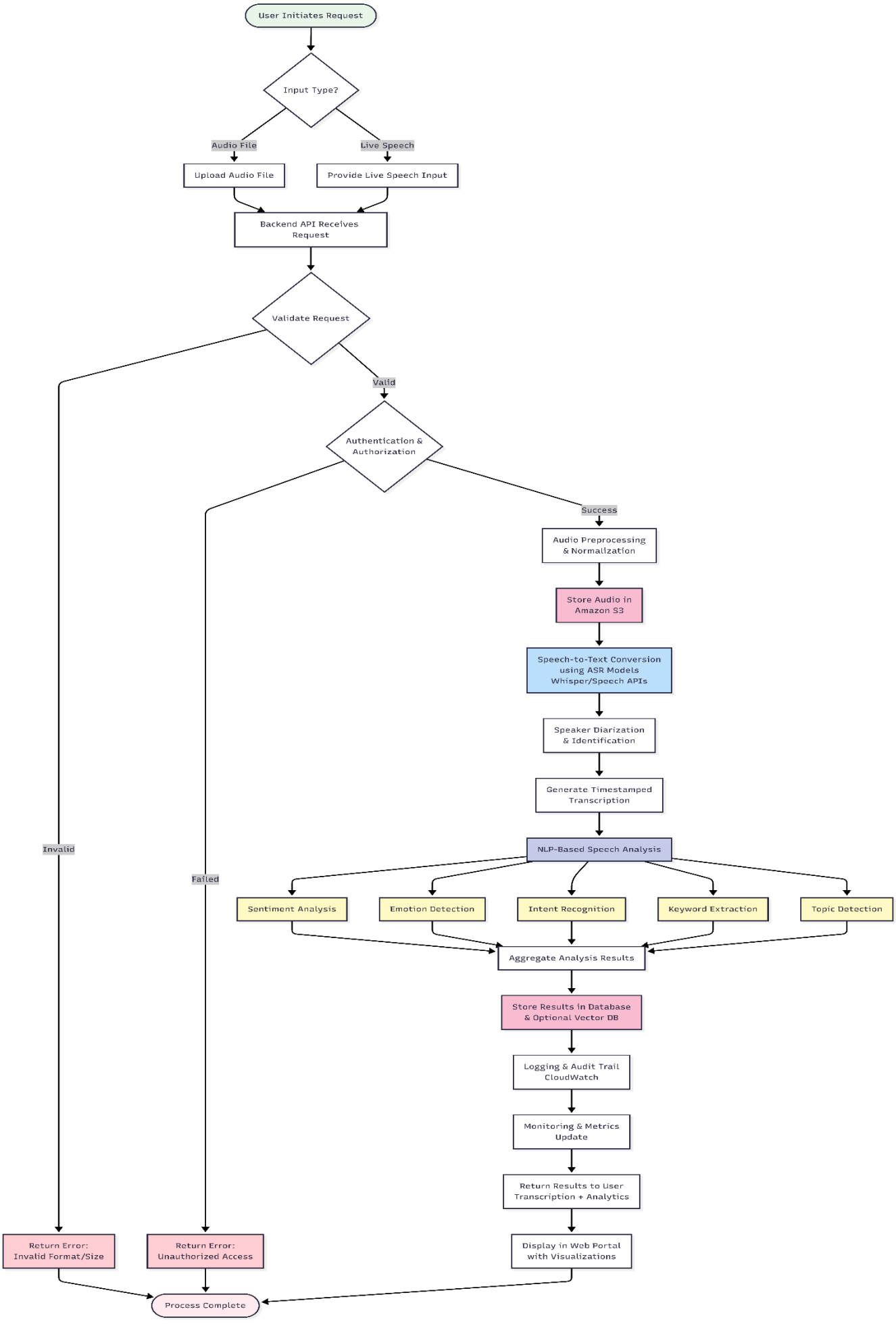
The system architecture follows a layered approach consisting of User Layer, Application Layer, AI & Processing Layer, Data Layer, and Infrastructure & Monitoring Layer. Users interact through a web interface, audio is

processed via backend APIs, converted into text using ASR models, analyzed using NLP models, and securely stored.



Flowchart -

1. User uploads audio / provides live speech input
2. Backend API receives and validates request
3. Audio preprocessing and normalization
4. Speech-to-Text conversion using ASR
5. NLP-based speech analysis
6. Results stored and returned to user
7. Logging and monitoring


User Initiates Request
Input Type?
Audio File
Live Speech
Upload Audio File
Provide Live Speech Input
Backend API Receives Request
Validate Request
Valid
Invalid
Authentication & Authorization
Failed
Success
Audio Preprocessing & Normalization
Store Audio in Amazon S3
Speech-to-Text Conversion using ASR Models Whisper/Speech APIs
Speaker Diarization & Identification
Generate Timestamped Transcription
NLP-Based Speech Analysis
Sentiment Analysis
Emotion Detection
Intent Recognition
Keyword Extraction
Topic Detection
Aggregate Analysis Results
Store Results in Database & Optional Vector DB
Logging & Audit Trail CloudWatch
Monitoring & Metrics Update
Return Results to User Transcription + Analytics
Display in Web Portal with Visualizations
Return Error: Invalid Format/Size
Return Error: Unauthorized Access
Process Complete

AWS Costing

AWS Services & Purpose

AWS Services	Purpose
Amazon S3	Storage of audio files and transcripts
Speech-to-Text API	Audio transcription
Embedding / NLP Models	Speech and sentiment analysis
Amazon OpenSearch	Semantic indexing (optional)
AWS ECS / Lambda	Backend processing
API Gateway	Secure API routing
CloudWatch	Monitoring and audit logs

Estimated Cost per 1,000 Audio Requests

Component	Estimated Cost (USD)
ASR / LLM Usage	$5.00
Embeddings / NLP	$0.80
Vector DB	$0.40
Compute	$1.00
Storage	$0.10
Monitoring & Security	$0.30
Total	~$7–8