AI Voice Intelligence System: Gender Recognition, Speech-to-Text & Text Summarization Using NLP

Submitted by:
Rachuri Ramesh, Imarticus Learning,PGA-46, Hyderabad.

1. Tools Used:
Python, Librosa, Scikit-learn, TensorFlow.

---

2. Abstract / Executive Summary

This project focuses on building an **AI-based Voice Intelligence System** capable of: Recognizing **speaker gender**

Converting speech into **text transcription**

Generating **summarized insights** from the spoken content

Using audio signal processing, machine learning, and NLP techniques, the system processes raw audio files, extracts relevant features, performs automatic speech recognition using pre-trained models, and generates concise summaries using Transformer architectures.

The prototype demonstrates how organizations can automate call analysis, improve customer handling, and enhance voice-based analytics.

---

3. Introduction

Background

Modern industries generate massive volumes of voice data, including customer calls, meeting recordings, and assistant commands. Manual analysis is slow, inaccurate, and not scalable.

Need for AI-Driven Voice Intelligence

Traditional transcription tools cannot:

Detect user attributes (e.g., gender)

Extract insights automatically

Summarize content intelligently

There is a need for a **modular and scalable AI pipeline** integrating speech processing and NLP.

Objective

Develop an end-to-end automated system that provides:

Gender classification

High-quality speech-to-text

Coherent summarization of long text

This system aims to demonstrate real-time voice understanding using open-source tools.

---

4. Problem Statement

Issues Identified

Manual transcription is time-consuming.

Extracting insights from audio data requires multiple tools.

No unified solution for gender detection + ASR + summarization.

Challenges

Background noise variation

Accent diversity

Limited labeled gender datasets

High computation for ASR models

Goal

Build a consolidated Voice Intelligence System that converts speech → text → summary with gender interpretation.

---

5. Data Sources

Primary Data

Custom recorded audio samples for testing gender recognition and ASR.

Secondary Data

Mozilla Common Voice (speech & gender data)

LibriSpeech Corpus (ASR training dataset)

CNN/Daily Mail Dataset (summarization model fine-tuning)

Attributes Extracted

Audio signal

Sample rate

MFCCs

Pitch & spectral features

Transcribed text tokens

---

6. Methodology

Step 1: Audio Preprocessing

Silence removal

Noise reduction

Sampling consistency

Feature extraction (MFCC, pitch, zero-crossing rate)

Step 2: Gender Classification Model

Feature extraction using Librosa

CNN / RNN architecture for classification

Train–test split and model evaluation

Accuracy, precision, and recall analysis

Step 3: Speech-to-Text (ASR)

Using: vosc model small –offline, SpeechRecognition API

Process:
Audio → Tokenization → Acoustic model → Language model → Text output

Step 4: Text Summarization Module

Tokenize text

Preprocess (stopword removal, lowercasing)

Apply T5/BART summarizer

Generate concise summaries

Step 5: Integration & Deployment

Build a simple user interface with Streamlit

Input: Audio file

Output: Gender + Transcription + Summary

---

7. Implementation

(Add screenshots: MFCC plot, CNN model summary, Wav2Vec2 output, summarizer output.)

Implementation highlights include:

Python scripts for audio cleaning

CNN model for gender recognition integration

Transformer summarizer testing

Streamlit interactive UI

(Capstone 1 example had Python, SQL, and Power BI screenshots. Similarly, add your notebook and UI screenshots here.)

---

8. Results & Analysis

1. Gender Recognition

Achieved high accuracy on trained samples. Distinct MFCC patterns were visible for male vs female voices.

2. Speech-to-Text

Wav2Vec2 produced high-quality, noise-resilient transcripts.

3. Text Summarization

Summaries accurately captured core meaning with reduced sentence complexity.

4. End-to-End Output

Input Audio → Gender → Full Transcript → Summary
Validated with multiple users' recordings.

---

9. Discussion

Benefits

Automates audio intelligence pipeline

Reduces manual transcription efforts

Summaries help organizations extract actionable insights

Tool Comparison

Librosa for feature engineering

TensorFlow/PyTorch for modeling

HuggingFace for ASR & NLP

Streamlit for interactive UI

This creates a complete AI-powered voice analysis system.

---

10. Risks, Challenges, and Limitations

Accent diversity affecting ASR accuracy

Limited labeled datasets for gender classification

Transformer summarizers require GPU for optimal performance

Ethical concerns: privacy, consent, and misuse of voice data.

---

11. Conclusion

The project successfully integrates **speech processing + NLP** into a single pipeline. It demonstrates how gender recognition, transcription, and summarization can be automated using deep learning models.

The final system serves as a foundation for advanced applications in customer analytics, virtual assistants, accessibility tools, and call center intelligence.

---

12. Future Work

Same format as Capstone 1 future enhancements:

1. Deploy as a Cloud API

Host on AWS or Azure

Provide REST endpoints for organizations

2. Real-Time Streaming Input

Detect gender & transcribe during live calls

3. Add Emotion Recognition

Happy, angry, neutral tone classification

4. Add Topic Extraction & Keyword Detection

Provide auto-generated tags

5. Multi-Language Support

Extend ASR and summarizer for Indian languages

---

13. References

Mozilla Common Voice, LibriSpeech Corpus,

CNN/DailyMail Dataset, HuggingFace Transformers,

Librosa, Pandas, NumPy, TensorFlow, PyTorch

Streamlit, Capstone 2 Project Proposal