Probabilistic Modeling of Spoken Digits: A GMM-Bayesian Framework

Wanghley Soares Martins

2024-12-04

This project explores the application of Gaussian Mixture Models (GMMs) in spoken digit analysis. The aim is to evaluate the efficiency of GMMs in distinguishing audio features for accurate classification.

Executive Summary

This project presents a comprehensive analysis of spoken digits using Gaussian Mixture Models (GMM) and Mel-frequency cepstral coefficients (MFCCs). The study focuses on developing a robust system for analyzing and classifying spoken digits through advanced statistical modeling techniques.

Introduction

Background and Motivation

Speech recognition technology has become a cornerstone of human-computer interaction, powering applications from virtual assistants to automated customer service systems. This project focuses on the specific challenge of spoken digit recognition, which serves as an excellent starting point for more complex speech recognition tasks.

Project Objectives

The primary objectives of this analysis are:

- Examine and characterize the distribution of MFCC features extracted from spoken digits

- Implement and evaluate GMM clustering for modeling different pronunciations
- Analyze likelihood distributions across different digits
- Create visualizations to understand the underlying patterns in speech data

Scope and Limitations

This study focuses on:

- Single-speaker digit recognition
- Isolated digit pronunciation (not continuous speech)
- Clean audio recordings without background noise

Theoretical Framework

Mel-frequency Cepstral Coefficients (MFCCs)

Overview

MFCCs are coefficients that collectively represent the short-term power spectrum of a sound. They are derived from a type of cepstral representation of the audio clip.

Significance in Speech Recognition

MFCCs are particularly valuable because they:

- Approximate the human auditory system's response
- Provide a compact representation of the speech signal
- Capture important phonetic characteristics

Gaussian Mixture Models

Mathematical Foundation

GMMs are probabilistic models that assume all data points are generated from a mixture of a finite number of Gaussian distributions. The model is defined by:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where: - K is the number of components - π_k are the mixture weights - μ_k and Σ_k are the mean and covariance of each Gaussian component

Application to Speech Analysis

GMMs are particularly well-suited for speech analysis because they can:

- Model complex distributions
- Capture multiple modes in the feature space
- Provide probabilistic assignments of data points

Dataset Characteristics

Characteristic	Value
Total Tokens	8,800
Unique Digits	10 (0-9)
Speakers	88 (44 female, 44 male)
MFCC Features	13 coefficients
Frames per Token	~35-40

Arabic Digit Pronunciations

Digit	Arabic	Phonetic
0		sifir
1		wahad
2		ithnayn
3		thalatha
4		araba'a
5		khamsa
6		sittah
7		seb'a
8		thamanieh
9		tis'ah

Methodology

Data Processing Pipeline

[Continue with additional methodology sections]

Results and Analysis

Performance Metrics

Overall Accuracy

[Space for overall accuracy visualization]

Confusion Matrix

[Space for confusion matrix visualization]

Per-Digit Analysis

[Space for per-digit performance analysis]

Appendices

A. Implementation Details

Parameter	Value	Justification
GMM Components	K	Based on phoneme count
Convergence Threshold		Empirically determined
Frame Window	N	Optimal temporal coverage

Project Timeline