Speaker Recognition

An Outline of Neural Network-Based Approaches

Dmitriy Yakovlev

Department of Computer Science
Harvey Mudd College

November 18, 2008

Overview

What is Speaker Recognition?

Preliminary Principles

A detailed look at the process.

Neural Network Methods

Two Different Approaches

Overview

What is Speaker Recognition?

Preliminary Principles

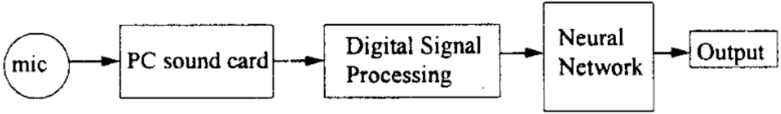
A detailed look at the process.

Neural Network Methods

Two Different Approaches

In General

- **Input**: Audio encoding of speech
- **Output**: Information that classifies speaker



Applications:

- User authentication
- Biometrics
- Caller ID
- ...

Paradigms

- **Speaker Recognition**
  Recognize that a member of a known population spoke
- **Speaker Verification**
  Verify that a given subject is who he claims to be
- **Speaker Identification**
  Detect a particular speaker from a known population
  - **Text Dependent**
  - **Text Independent**

Approaches

- **Acoustic Phonetic**
  Based on theory that speech can be broken down uniquely into phonemes that can then be easily characterized and used to categorize future inputs
- **Pattern Recognition**
  The 'brute force' approach: if enough samples are trained, common features will be accentuated and recognized
- **Artificial Intelligence**
  Combination of the above; systematic approach to emulate human recognition and extraction of signal features for pattern matching

Overview

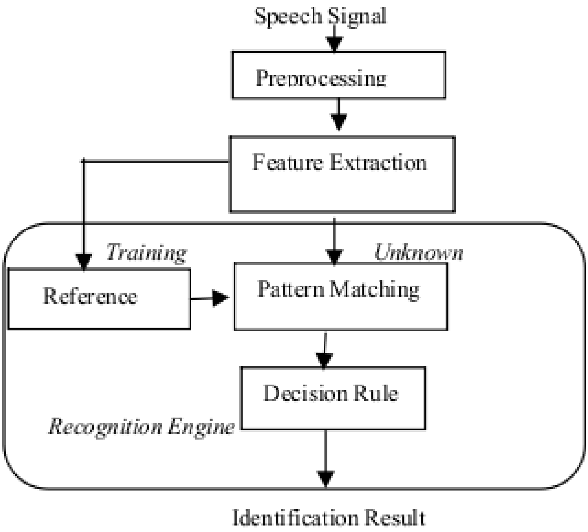
What is Speaker Recognition?

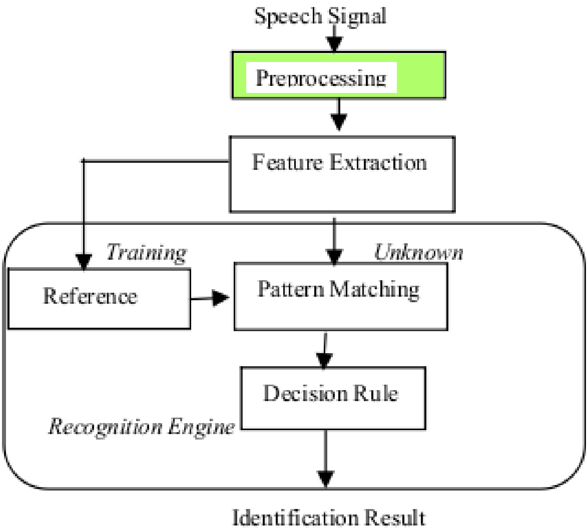
Preliminary Principles

A detailed look at the process.

Neural Network Methods

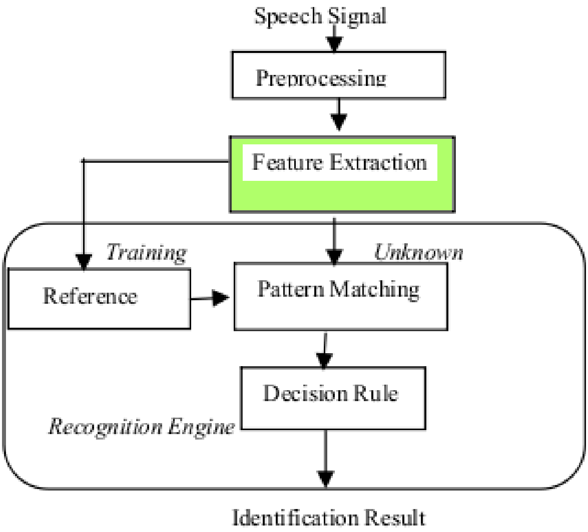
Two Different Approaches


Speech Signal
Preprocessing
Feature Extraction
Training
Unknown
Reference
Pattern Matching
Decision Rule
Recognition Engine
Identification Result


Speech Signal
Preprocessing
Feature Extraction
Training
Unknown
Reference
Pattern Matching
Decision Rule
Recognition Engine
Identification Result

Preprocessing

- Digital filtering
  - Noise Removal
  - Level Adjustment
  - Frequency Attenuation
- Endpoint detection


Speech Signal
Preprocessing
Feature Extraction
Training
Unknown
Reference
Pattern Matching
Decision Rule
Recognition Engine
Identification Result

Feature Extraction

We need to obtain characteristic values of the sound sample.
Why? How?

- Discrete Fourier Transform
- Power Spectral Density
- Linear Predictive Coding
- Average Mean Distance function

All three methods give us *numbers*, which Neural Networks like for input. These numbers should be unique and characteristic of the sound samples they were generated from.

Feature Extraction

We need to obtain characteristic values of the sound sample. Why? How?

- Discrete Fourier Transform
- Power Spectral Density
- Linear Predictive Coding
- Average Mean Distance function

All three methods give us *numbers*, which Neural Networks like for input. These numbers should be unique and characteristic of the sound samples they were generated from.

Feature Extraction

We need to obtain characteristic values of the sound sample. Why? How?

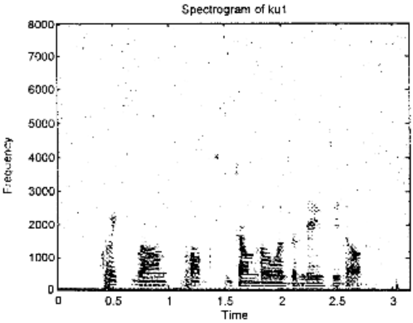
- Discrete Fourier Transform
- Power Spectral Density
- Linear Predictive Coding
- Average Mean Distance function

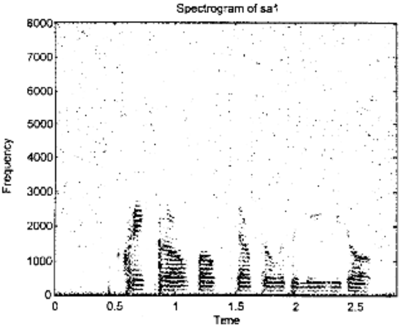
All three methods give us *numbers*, which Neural Networks like for input. These numbers should be unique and characteristic of the sound samples they were generated from.

What are they?

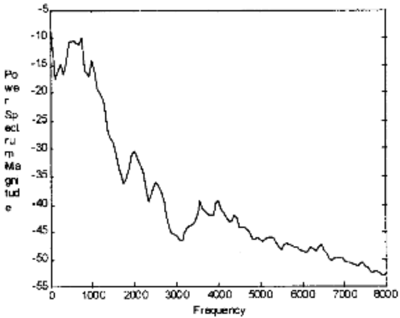
- **Discrete Fourier Transform**
  Algorithm to move from continuous signal to discrete frequency domain
- **Power Spectral Density**
  Computed from the DFT, unique spectral representation of a signal
- **Linear Predictive Coding**
  Used to represent the logarithmic power spectrum of a signal in compressed form
- **Average Mean Distance function**
  Used to find the fundamental frequency of a signal

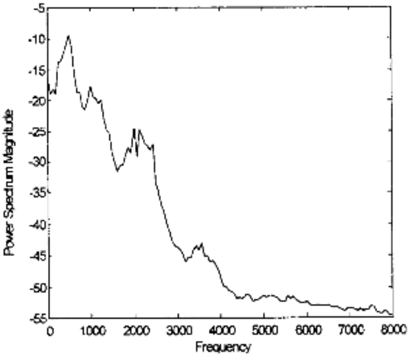
Example of DFT spectra





Example of PSD graphs

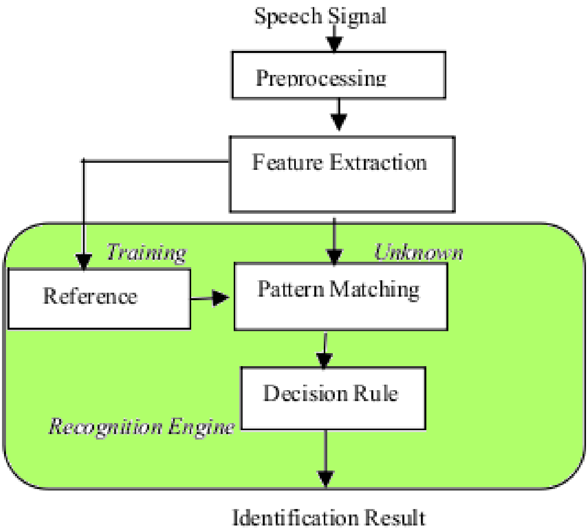




Overview
What is Speaker Recognition?

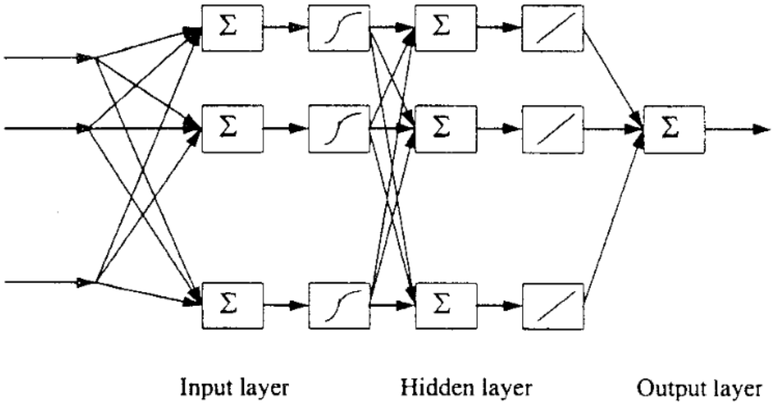
Preliminary Principles
A detailed look at the process.

Neural Network Methods
Two Different Approaches


Speech Signal
Preprocessing
Feature Extraction
Training
Unknown
Reference
Pattern Matching
Decision Rule
Recognition Engine
Identification Result

Multi-Layer Perceptrons

Basic three-layer feed-forward network



Multi-Layer Perceptrons

Basic three-layer feed-forward network

Positives:

- Simple to train
- Simple to understand
- Good accuracy on trained samples; up to 100%

Negatives:

- Possibility of overfitting data
- Bad results on untrained or new data
- Better suited to text-dependent recognition

Self Organizing Maps

- **Much more robust than MLP**
- Use the SOM as a map of codewords

Self Organizing Maps

- **Much more robust than MLP**
- Use the SOM as a map of codewords

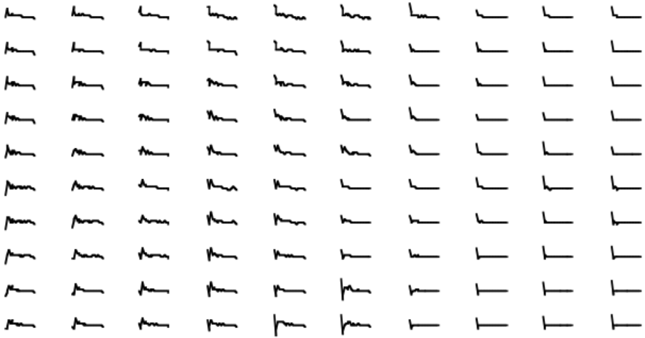


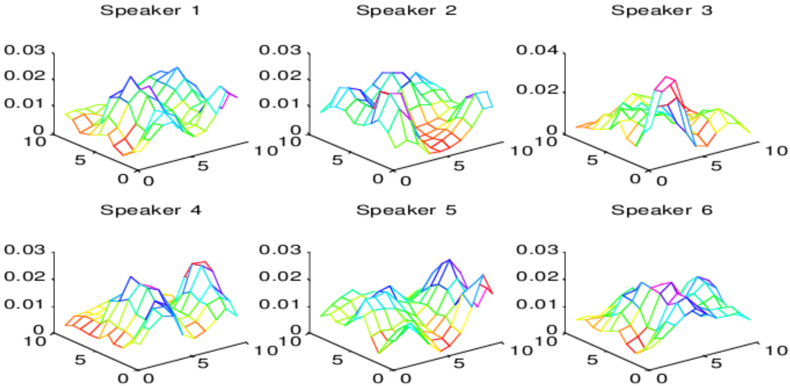
Figure 1: The topological property of the SOM: neighboring units on the SOM are associated with neighboring codewords.

Self Organizing Maps

- **Much more robust than MLP**
- Use the SOM as a map of codewords
- Similar words will be neighbors in the map
- Sample input can then be made into an occupancy histogram

Self Organizing Maps

- **Much more robust than MLP**
- Use the SOM as a map of codewords
- Similar words will be neighbors in the map
- Sample input can then be made into an occupancy histogram


Speaker 1
Speaker 2
Speaker 3
Speaker 4
Speaker 5
Speaker 6
0.03
0.02
0.01
0
0.04
10
5
0

SOM Experiment Results

SNR(DB)	clean	30 dB	20 dB	10 dB
LPC/SOM	98,2	95,0	55,0	8,0
MFCC/SOM	100	95,5	53,0	19,5
LPC/AHSM	97,3	85,14	36,5	5,6
MFCC/AHSM	98,6	96,4	68,9	27,6

- 100 speakers, 100 speech samples per speaker
- 25x25 Self-Organizing Map

Summary

- Speaker/Voice recognition can be accurately done by clever usage of a neural network
- Multi-Layer Perceptrons and Self Organizing Maps are two network types that have been adapted for this purpose