A Tutorial on Text-Independent Speaker Verification

F. Bimbot et al., 2004

Presented by Hassan A. Kingravi

Overview

- Introduction
- Methods for Parameterization and Modeling
- Normalization of Scores
- Evaluation
- Extensions
- Applications

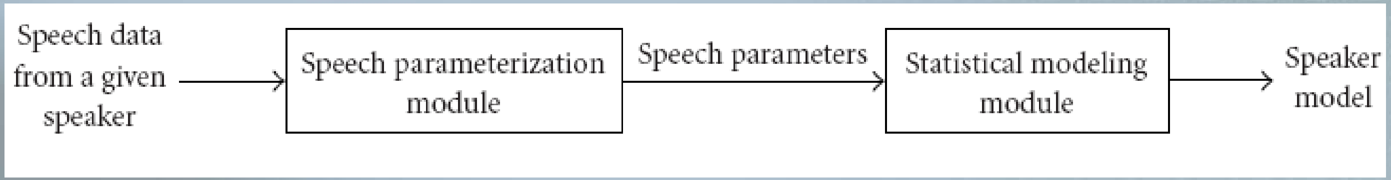
Introduction

- Speaker Verification?
  * Is this person who he/she claims to be? An example of biometrics
  * Natural source of data: considered to be less intrusive than other methods

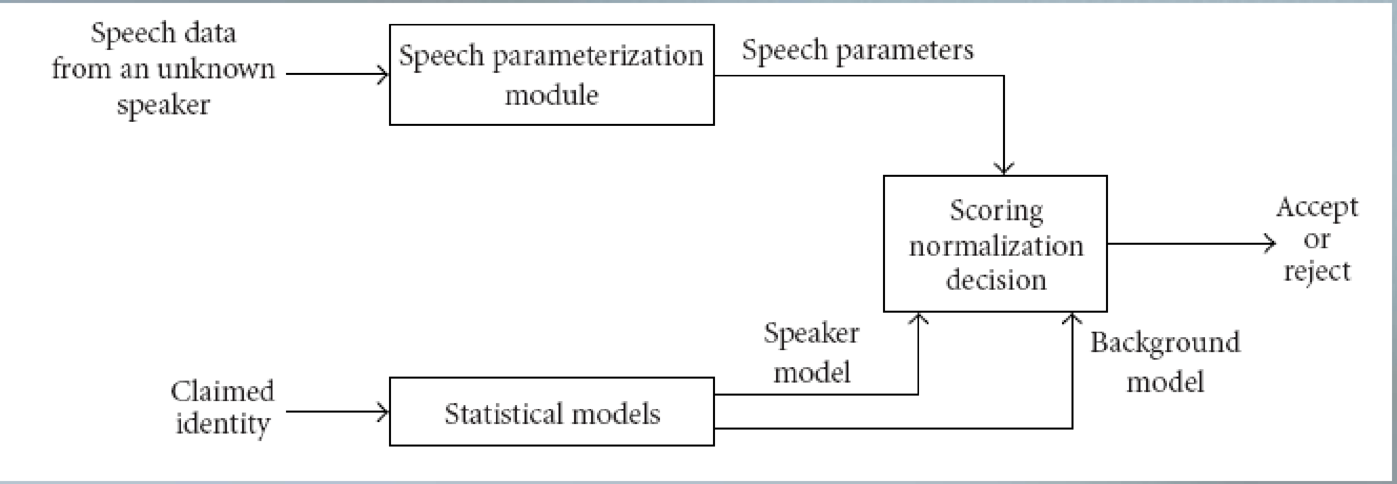
- Text-Independence?
  * Most applications based on digit recognition or fixed vocabulary
  * Text-independence implies operation independent of user cooperation and spoken utterance

Introduction

- Training Phase

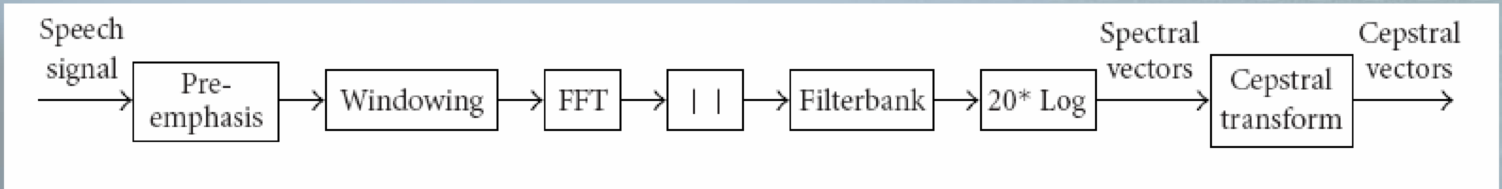


- Test Phase

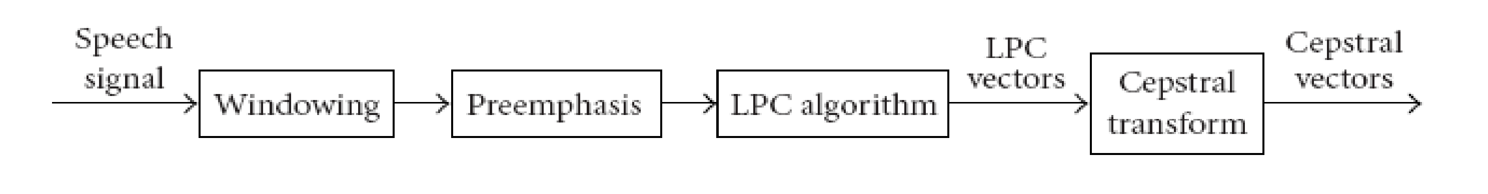


Parameterization

- Filterbank Cepstral Parameters



- LPC Cepstral Parameters



- Dynamic Information and log-Energy
- Discard Useless Information

Modeling

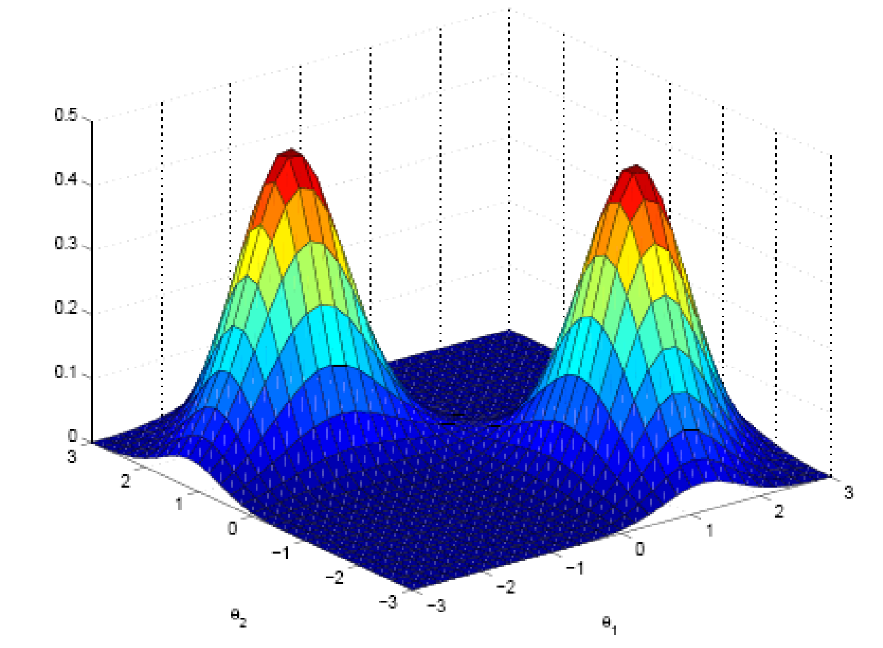
- Likelihood Ratio Detection
  Given segment of speech Y and speaker S, determine if Y was spoken by S. For this, we need two hypotheses; Y was spoken by S and Y was not spoken by S. To implement this, we train two models; if X represents parameterization of Y,
  p(X|alpha) = speaker model
  p(X|alpha_) = non-speaker model

  The ratio is the likelihood; if greater than threshold, accept, else reject

- Options for Non-speaker Model
  * Train multiple models
  * Train single model

Gaussian Mixture Models

- GMMs used to represent likelihood function p(X|alpha)
- Basically a weighted combination of M unimodal Gaussian densities of dimension D
- An example where D = 3 and M = 2



- Interpretation: each unimodal component represents a broad acoustic class

Gaussian Mixture Models

- Given a set of training vectors of dimension D, we select M and then train the model using the EM algorithm
- M = 512 on constrained, and 2048 on unconstrained speech
- The GMM is both parametric (has structure and parameters that can be tuned) and non-parametric (arbitrary density modeling)
- Advantages: computationally inexpensive, well-understood and insensitive to temporal aspects of speech
- The latter is actually a disadvantage in some cases; throws away information

Adapted GMM System

- Different methods of GMM training; one approach is to train background model first (using large M) and then train the speaker mode independently; often performs poorly

- Adaptation approach; train single GMM for background, and using training vectors for the speaker, create a new model for the speaker from the background GMM

- Method:
  Step 1) compute statistics from the new data such as weights, mean and variance
  Step 2) combine new information with the old s.t. mixtures with high counts of data from the speaker rely more on the new statistics and vice versa

Why Adaptation?

- Results indicate better performance
- The background models an acoustic space; tuning the existing one for speaker model leads to less surprises; likelihood ratio unaffected by “unseen” acoustic events
- Fast-Scoring:
  Step 1) For each feature vector, determine C top-scoring mixtures in background model; compute likelihood using only these
  Step 2) Score the vector against the top C mixtures in the speaker model
- Alternative Methods:
  * ANNs
  * SVMs

Normalization

- The problem: once the likelihood is calculated, compare with a threshold to make decision. How do we calculate the threshold?
- Score variability a major issue; speaker may be tired, in poor health etc, or there might be environmental issues (like background noise). Composition of training set for background also affects scores.
- Normalization of score variability makes decision threshold tuning easier:

$$\tilde{L}_\lambda(X) = \frac{L_\lambda - \mu_\lambda}{\sigma_\lambda}$$

Normalization Methods

- World-model and cohort-based normalizations

$$\tilde{L}_{\lambda}(X) = \frac{L_{\lambda}(X)}{L_{\bar{\lambda}}(X)}$$

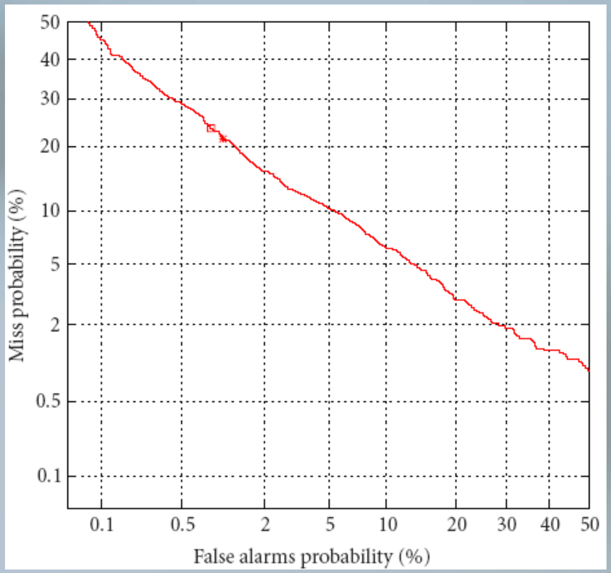
- Centered/Reduced Imposter Distribution
  Most commonly used (derived from equation on previous slide)

- The Norms
  Znorm, Hnorm, Tnorm, Htnorm, Cnorm, Dnorm

- WMAP
  World-model Maximum A Posteriori normalization. Produces a meaningful score in probability space

Evaluation

- 2 forms of errors; false negatives and false positives. Depends on application as to which is more serious.
- Performance measure: DET Curve



- Factors affecting performance: environmental issues, speaker “performance”, “goats and lambs”, training set size and diversity

Extensions

- Multiple Speaker Detection
- Speaker Tracking
- Segmentation

Applications: General

- On-Site
  In a given facility, voice recognition required for access to certain features or places
- Remote Applications
  Secure access to remote databases or services
- Information Structuring
  Automatic annotation of audio archives, speaker indexing, speaker change in subtitles etc.
- Games
  Personalized toys (seemingly humanity’s most pressing need)

Applications: Forensic

- Forensics
  Refers to criminal investigation, i.e. voice identification of a suspect.

- Difficulties
  Situation for recognition more difficult; more noise, variability etc.

- Controversy
  A “voice print” is not the same thing as a finger-print; not physiological because of the psychological factors involved. Because of possible errors, the concept of nonzero errors creates difficulties in judicial process.

- Systems
  Semiautomatic systems require expert input; “supervised selection of acoustic phonetic events”. Automatic systems exist, and are based on the preceding discussion.

Future Work

- Robustness Issues
  Channel variability and mismatched conditions, especially in microphones, play havoc with acoustic feature extraction. These need to be addressed, especially in a real-world, and not a laboratory setting.

- Exploitation of Higher Levels of Information
  Word usage, prosodic (manner of speech) measures etc.

- Emphasis on Unconstrained Tasks
  No prior assumptions on the state of the environment, for a given value of “no.”

References

- Bimbot et al. “A Tutorial on Text-Independent Speaker Verification.”
- Frank Dellart, “The Expectation Maximization Algorithm” (for GMM picture.)