Biometrics: Voice

Michael Stark

Outline

Outline
Fundamentals and Preprocessing
Pattern Matching Methods
Stochastic Models
Air Traffic Control System Presentation

- Fundamentals
- Features - Preprocessing
- Pattern Matching Methods
- Air Traffic Control System
- Conclusion

Outline

Fundamentals and Preprocessing
Speech Processing
Fundamentals
Vocal Apparatus
Problems in Speaker Recognition
Generic Speaker Verification
Features - Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation

Fundamentals and Preprocessing

Speech Processing

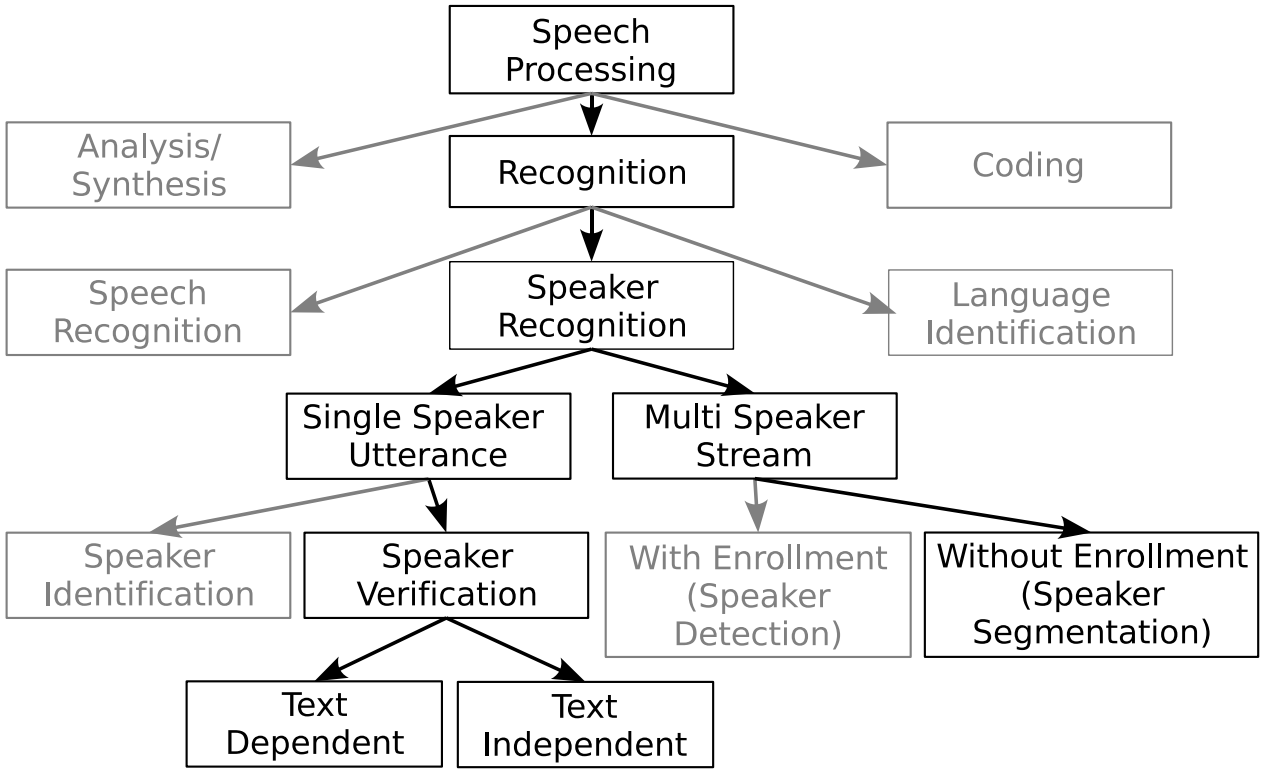
- Outline

Fundamentals and Preprocessing
- Speech Processing
- Fundamentals
- Vocal Apparatus
- Problems in Speaker Recognition
- Generic Speaker Verification
- Features - Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation



Fundamentals

Outline

Fundamentals and Preprocessing
- Speech Processing
- Fundamentals
- Vocal Apparatus
- Problems in Speaker Recognition
- Generic Speaker Verification
- Features - Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation

- Behavioral Biometrics - speakers identity can not be measured directly
- Speech carries 2 Informations:
  - Meaning of the message
  - Information about themselves as a person
- Speaker specific characteristics in signal
  - speaker's anatomy
  - physiology
  - linguistic
  - experience
  - mental state

Individuality in the sound system
- segmental component (e.g., mental lexicon, pronounced word)
- supra-segmental component (e.g., timing, stress pattern and intonation of a sequence)
- number and identity of segments used in the sound inventory

taken from [6]

Vocal Apparatus

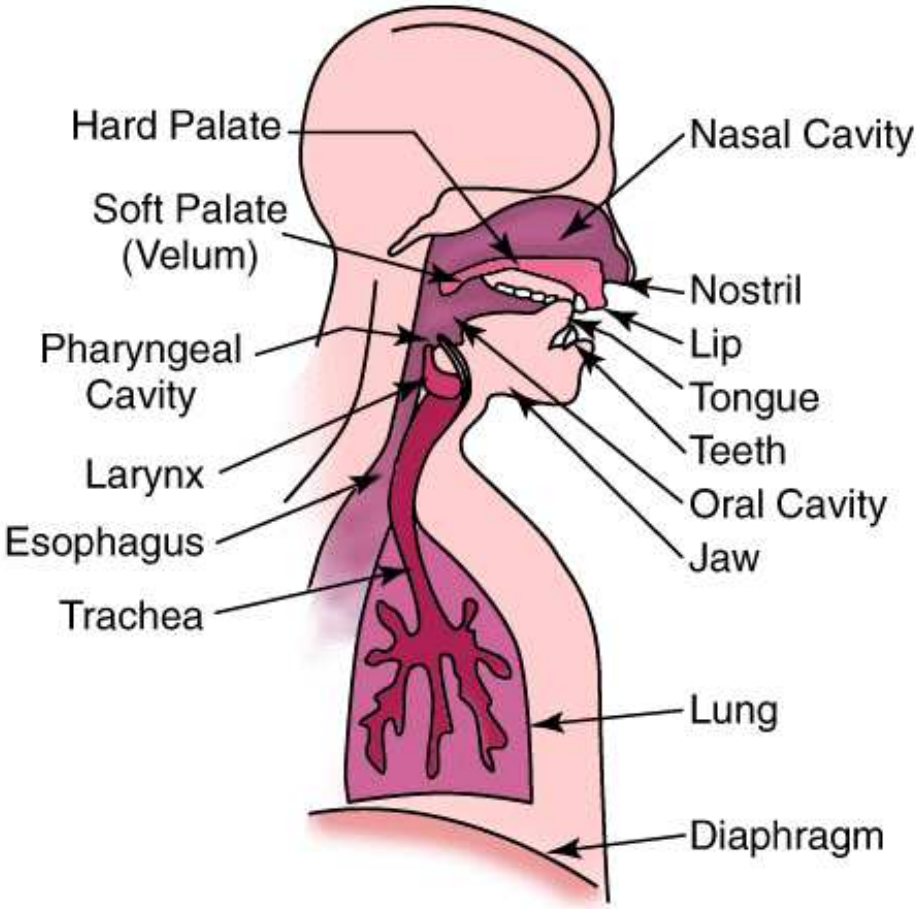
Outline

Fundamentals and Preprocessing
- Speech Processing
- Fundamentals
- Vocal Apparatus
- Problems in Speaker Recognition
- Generic Speaker Verification
- Features - Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation



adapted from [5]

Problems in Speaker Recognition

Outline

Fundamentals and Preprocessing
- Speech Processing
- Fundamentals
- Vocal Apparatus
- Problems in Speaker Recognition
- Generic Speaker Verification
- Features - Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation

- Misspoken or misread prompted phrases
- Extreme emotional states (e.g., stress or duress)
- Time varying (intra- or intersession) microphone placement
- Poor or inconsistent room acoustics (e.g., multipath and noise)
- Channel mismatch (e.g., using different microphones for enrollment and verification)
- Sickness (e.g., head colds can alter the vocal tract)
- Aging (the vocal tract can drift away from models with age)

taken from [5]

Generic Speaker Verification

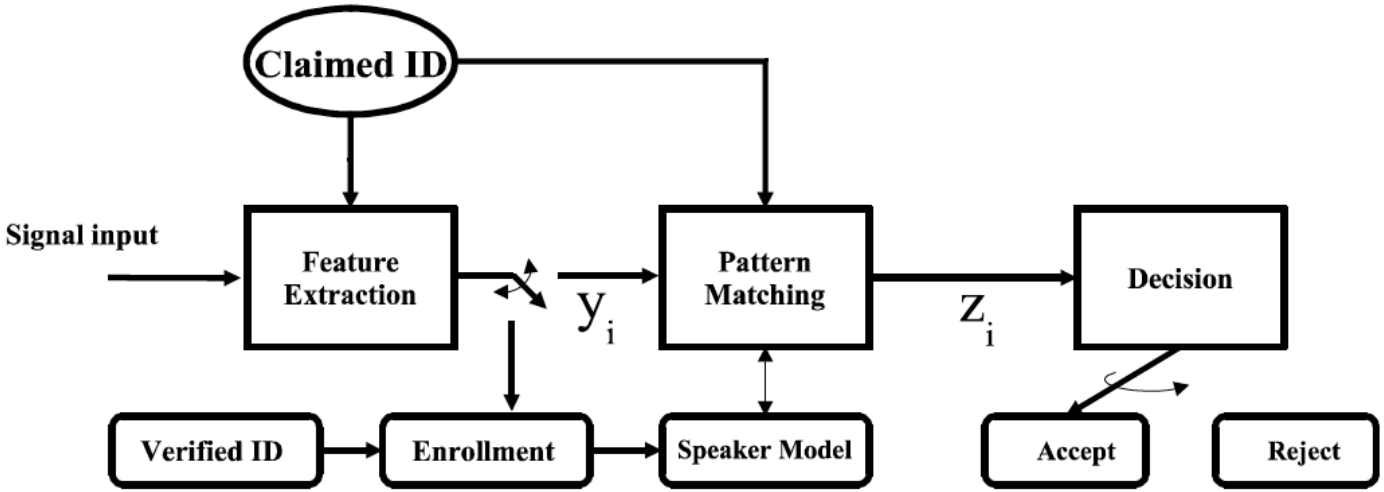
- Outline

Fundamentals and Preprocessing
- Speech Processing
- Fundamentals
- Vocal Apparatus
- Problems in Speaker Recognition
- Generic Speaker Verification
- Features - Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation



Features - Preprocessing

Outline

Fundamentals and Preprocessing
- Speech Processing
- Fundamentals
- Vocal Apparatus
- Problems in Speaker Recognition
- Generic Speaker Verification
- Features - Preprocessing

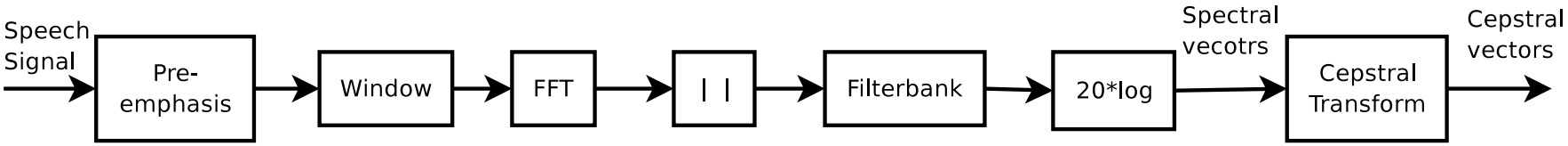
Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation

- Speech parameterization: Feature extraction from the speech signal
- Voice activity detection
- End point detection
- Feature normalization
- Dynamic information

Example Feature: Cepstral coefficients



Taken from [7]

Outline

Fundamentals and Preprocessing

Pattern Matching Methods
- Template Models
- Dynamic Time Warping
- Vector Quantization Source Modeling
- Nearest Neighbors
- Performance

Stochastic Models

Air Traffic Control System Presentation

Pattern Matching Methods

Template Models

Outline

Fundamentals and Preprocessing

Pattern Matching Methods
- Template Models
- Dynamic Time Warping
- Vector Quantization Source Modeling
- Nearest Neighbors
- Performance

Stochastic Models

Air Traffic Control System Presentation

Definition of template: $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}$, with N training vectors. Then a distance function can be defined as:

$$d(\boldsymbol{x}, \overline{\boldsymbol{x}}) = (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{W} (\boldsymbol{x} - \overline{\boldsymbol{x}}),$$

where $\boldsymbol{W}$ defines the chosen distance function.

Dynamic Time Warping

Outline

Fundamentals and Preprocessing

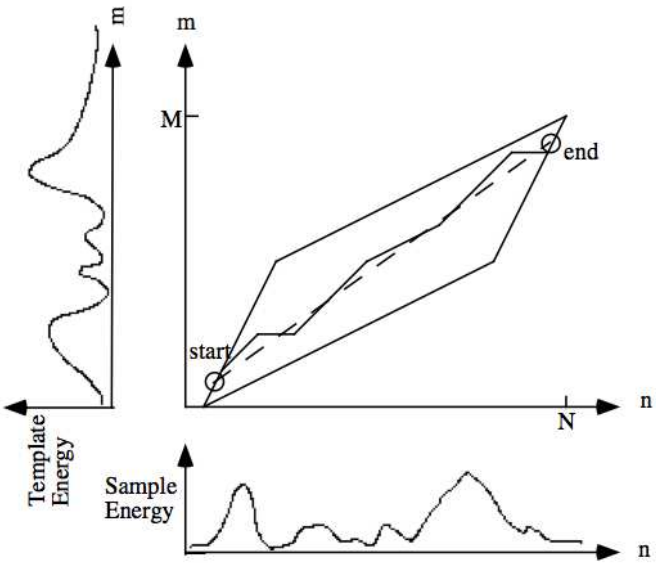
Pattern Matching Methods
- Template Models
- Dynamic Time Warping
- Vector Quantization Source Modeling
- Nearest Neighbors
- Performance

Stochastic Models

Air Traffic Control System Presentation

- Time-dependent methods
- Algorithm to compensate speaking rate variability
- Piece wise linear mapping of the time axis to align 2 signals and minimize z
- Text- dependent

The asymmetric match score z is given as: $z = \sum_{t=1}^{T} d(\boldsymbol{x}_t, \overline{\boldsymbol{x}}_{j(t)})$



Vector Quantization Source Modeling

Outline

Fundamentals and Preprocessing

Pattern Matching Methods
- Template Models
- Dynamic Time Warping
- Vector Quantization Source Modeling
- Nearest Neighbors
- Performance

Stochastic Models

Air Traffic Control System Presentation

- Time-independent
- Create a VQ code book as a collection of code words for each speaker by clustering
- No temporal information about the speaker used

The match score is defined as:

$$z = \sum_{t=1}^{T} \min \; d(\boldsymbol{x}_t, \overline{\boldsymbol{x}})$$

Nearest Neighbors

Outline

Fundamentals and Preprocessing

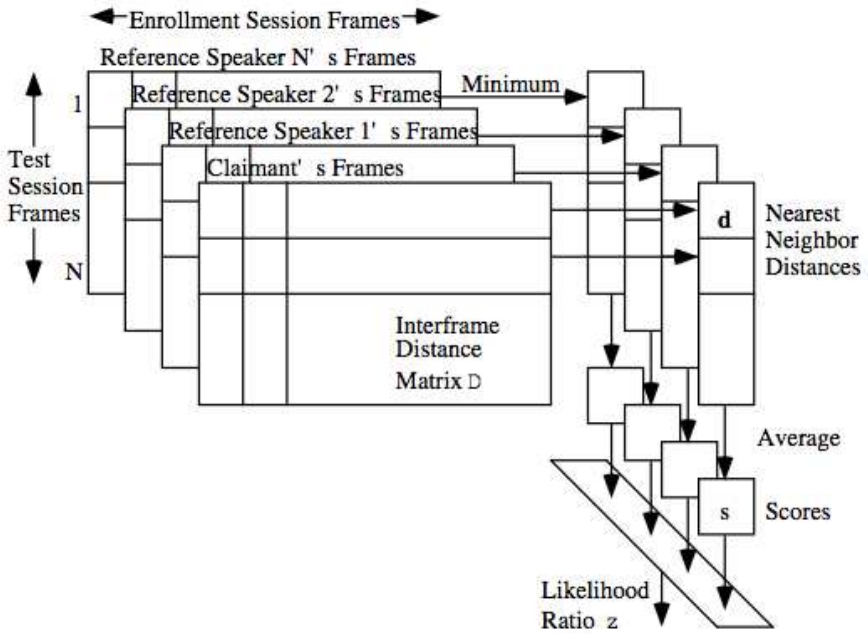
Pattern Matching Methods
- Template Models
- Dynamic Time Warping
- Vector Quantization Source Modeling
- Nearest Neighbors
- Performance

Stochastic Models

Air Traffic Control System Presentation

- Distance based classification by direct computation
- No models or data reduction by clustering
- Powerful method with high computational complexity

$$
\begin{aligned}
d(U,R) &= \frac{1}{|U|}\sum_{\boldsymbol{u}_i\in U}\min_{\boldsymbol{r}_j\in R}|\boldsymbol{u}_i-\boldsymbol{r}_j|^2+\frac{1}{|R|}\sum_{\boldsymbol{r}_j\in R}\min_{\boldsymbol{u}_i\in U}|\boldsymbol{u}_i-\boldsymbol{r}_j|^2 \\
&- \frac{1}{|U|}\sum_{\boldsymbol{u}_i\in U}\min_{\boldsymbol{u}_j\in U}|\boldsymbol{u}_i-\boldsymbol{u}_j|^2-\frac{1}{|R|}\sum_{\boldsymbol{r}_i\in R}\min_{\boldsymbol{r}_j\in R}|\boldsymbol{r}_i-\boldsymbol{r}_j|^2
\end{aligned}
$$



Performance

Outline

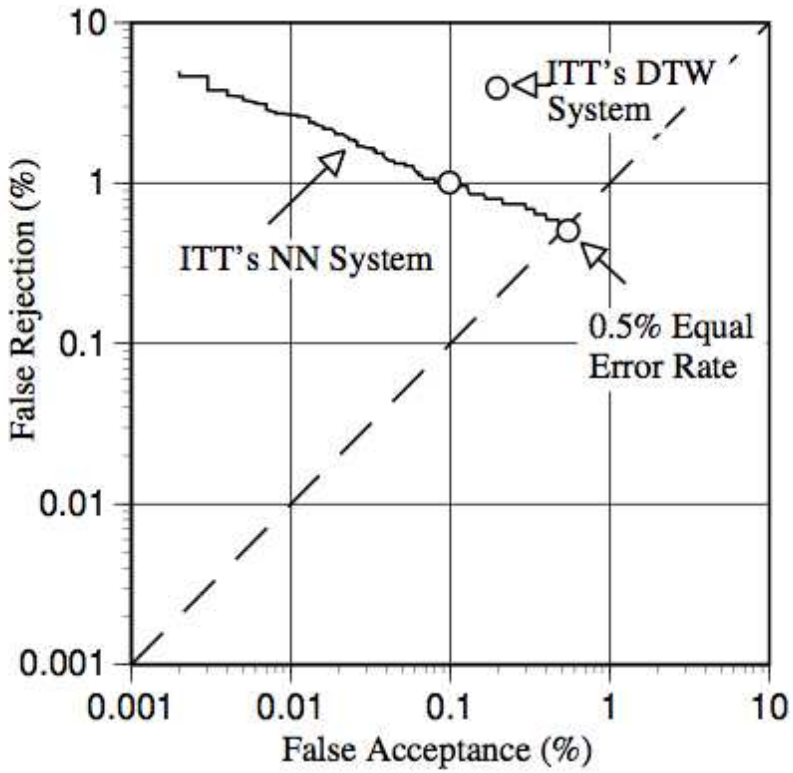
Fundamentals and Preprocessing

Pattern Matching Methods
- Template Models
- Dynamic Time Warping
- Vector Quantization Source Modeling
- Nearest Neighbors
- Performance

Stochastic Models

Air Traffic Control System Presentation

- YOHO database with 186 Subjects
- 9300 imposter trials
- DTW: 0.2% FA / 4 % FR; EER $\approx 1.5\%$
- NN: 0.1% FA / 1 % FR ; EER $\approx 0.5\%$



Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models
- Hidden Markov Models
- Gaussian Mixture Models
- GMM-UBM
- Support Vector Machines

Air Traffic Control System Presentation

Stochastic Models

Hidden Markov Models

Outline

Fundamentals and Preprocessing

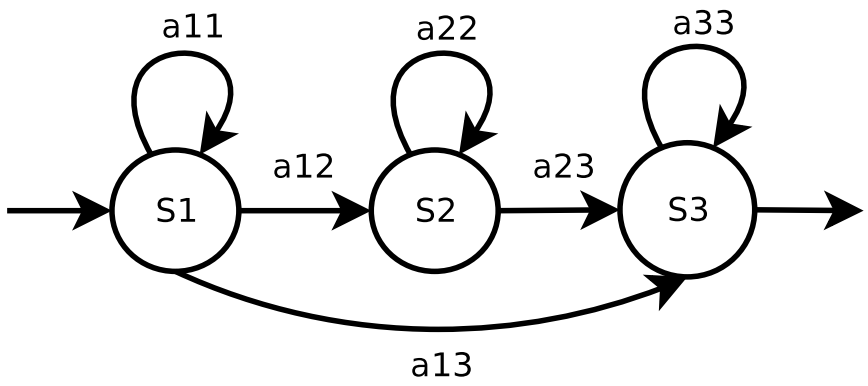
Pattern Matching Methods

Stochastic Models
Hidden Markov Models
Gaussian Mixture Models
GMM-UBM
Support Vector Machines

Air Traffic Control System Presentation

- Model represents a sequence of specific words
- Is a finite state machine, where each pdf $p(X|s_i)$ is associated with each state states are connected by a transition network with a given state transition probability $a_{ij} = p(s_i|s_j)$

$$p(\boldsymbol{x}|\lambda_i) = \sum_{\substack{\text{all state} \\ \text{sequences}}} \prod_{t=1}^{T} p(\boldsymbol{x}_t|s_t)\, p(s_t|s_{t-1})$$



EER $= 0.62\%$ @ $2.5s$ (YOHO, Che and Lin, 1995)

Gaussian Mixture Models

- Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models
- Hidden Markov Models
- Gaussian Mixture Models
- GMM-UBM
- Support Vector Machines

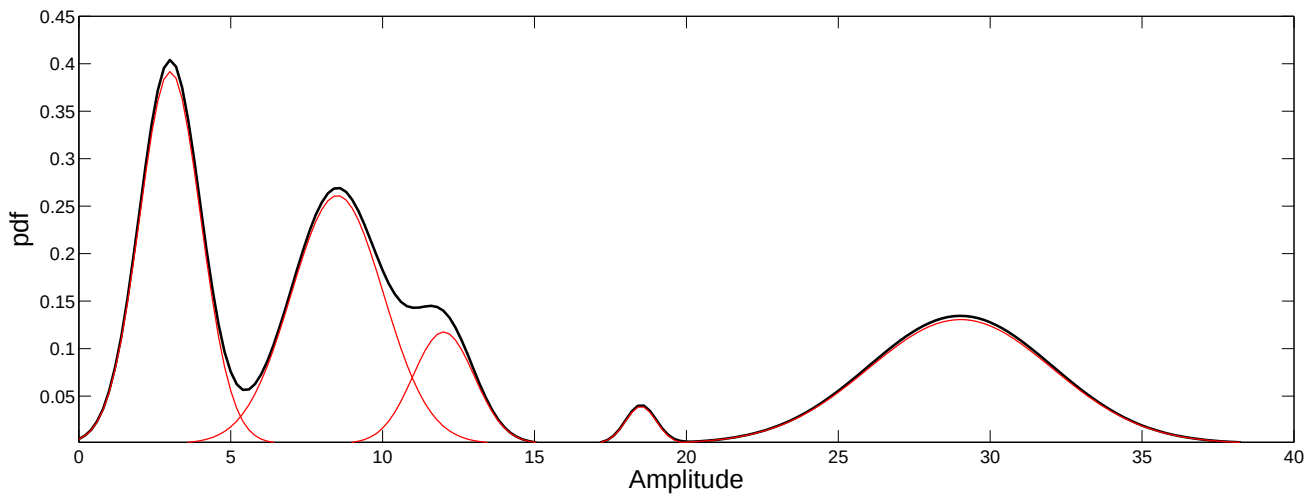
Air Traffic Control System Presentation

- Definition of a Gaussian Distribution

$$p_{\boldsymbol{x}}(\mu_c, \Sigma_c) = \frac{1}{(2\pi)^{D/2}|\Sigma_c|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_c})^T \Sigma_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu_c})\right]$$

- Weighted sum of C Gaussians to model target distribution

$$p(\boldsymbol{x}|\lambda) = \sum_{c=1}^{C} w_c \, p_{\boldsymbol{x}}(\mu_c, \Sigma_c)$$



GMM-UBM

Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models
- Hidden Markov Models
- Gaussian Mixture Models
- GMM-UBM
- Support Vector Machines

Air Traffic Control System Presentation

- Define a Universal Background Model (UBM)
- Perform speaker adaptation
- Tight coupling between SD and UBM model
- UBM also used as cohort model
- EER $\approx 10\%$ (2048 components)

Speaker adaptation methods:
- Weighted sum combining
- Maximum a posteriori combining (MAP)

MAP adaptation:

$$\begin{aligned}
\mathbf{c}_{k,spk_{Comb}} &= \left[\beta_k^c \mathbf{c}_{k,spk} + (1-\beta_k^c)\mathbf{c}_{k,ubm}\right]\epsilon \\
\mu_{k,spk_{Comb}} &= \beta_k^\mu \mu_{k,spk} + (1-\beta_k^\mu)\mu_{k,ubm} \\
\mathbf{\Sigma}_{k,spk_{Comb}} &= \beta_k^\Sigma \mathbf{\Sigma}_{k,spk} + (1-\beta_k^\Sigma)(\mathbf{\Sigma}_{k,ubm} + \mu_{k,ubm}^2) - \mu_{k,spk_{Comb}}^2,
\end{aligned}$$

with $\beta_k^\rho = \frac{\mathbf{c}_{k,spk}}{\mathbf{c}_{k,spk}+r^\rho}$

and r^ρ the relevance factor. taken from [7]

Support Vector Machines

Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models
- Hidden Markov Models
- Gaussian Mixture Models
- GMM-UBM
- Support Vector Machines

Air Traffic Control System Presentation

- Well suited for SV because of its binary nature of decision
- Construction of a boundary/hyperplane separating data sets
- Found optimum plane is a linear combination of a set of vectors (support vectors)
- For enrollment speaker and imposter data must be available
- Relaxation of linear separability condition to allow outliers
- Results in an EER : 0.59 % on the YOHO database

Performance for combined SVM-GMM system with non-linear kernel: EER $= 6.39\%$ (NIST 2006 SRE , 53966 tests, GMM-UBM baseline: 9.11%) [8]

Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation
- ATC System Presentation
- System Pattern Recognition Approach
- System Design
- Databases
- Results
- Conclusion
- References

Air Traffic Control System Presentation

ATC System Presentation

Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation

ATC System Presentation

System Pattern Recognition Approach

System Design

Databases

Results

Conclusion

References

- Technical Requirements
  - AM channel with poor quality → low SNR
  - *Narrow bandwidth in the region of 300 - 2500 Hz*
  - Real-time processing
- Speech Communication Specification
  - *Speaker turns on average only 5 seconds*
  - Hypothesized interval of uniform speaker through AIT
  - No offline speaker enrollment
  - By definition, start with reference speaker
  - Text-independent verification method used

System Pattern Recognition Approach

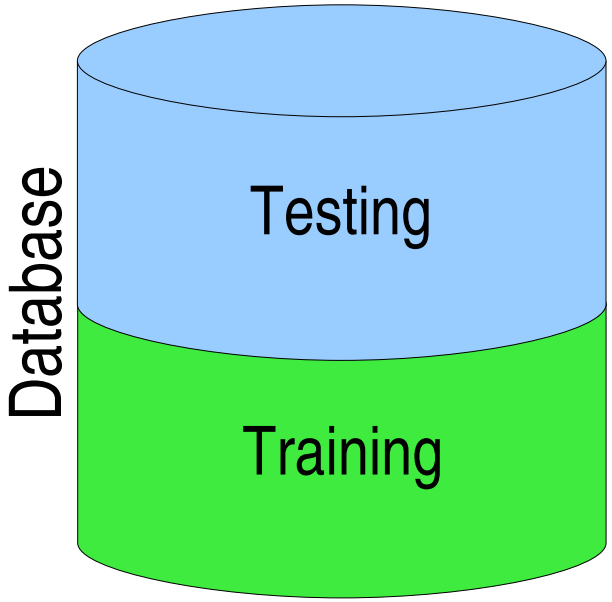
- Outline

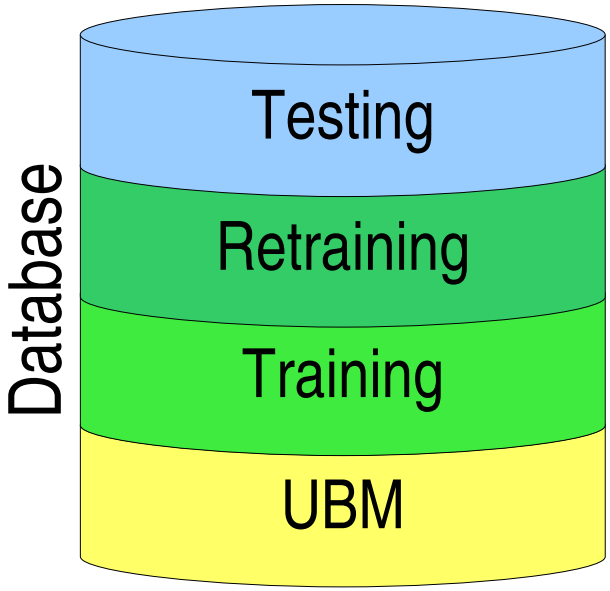
Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation
- ATC System Presentation
- System Pattern Recognition Approach
- System Design
- Databases
- Results
- Conclusion
- References





System Design

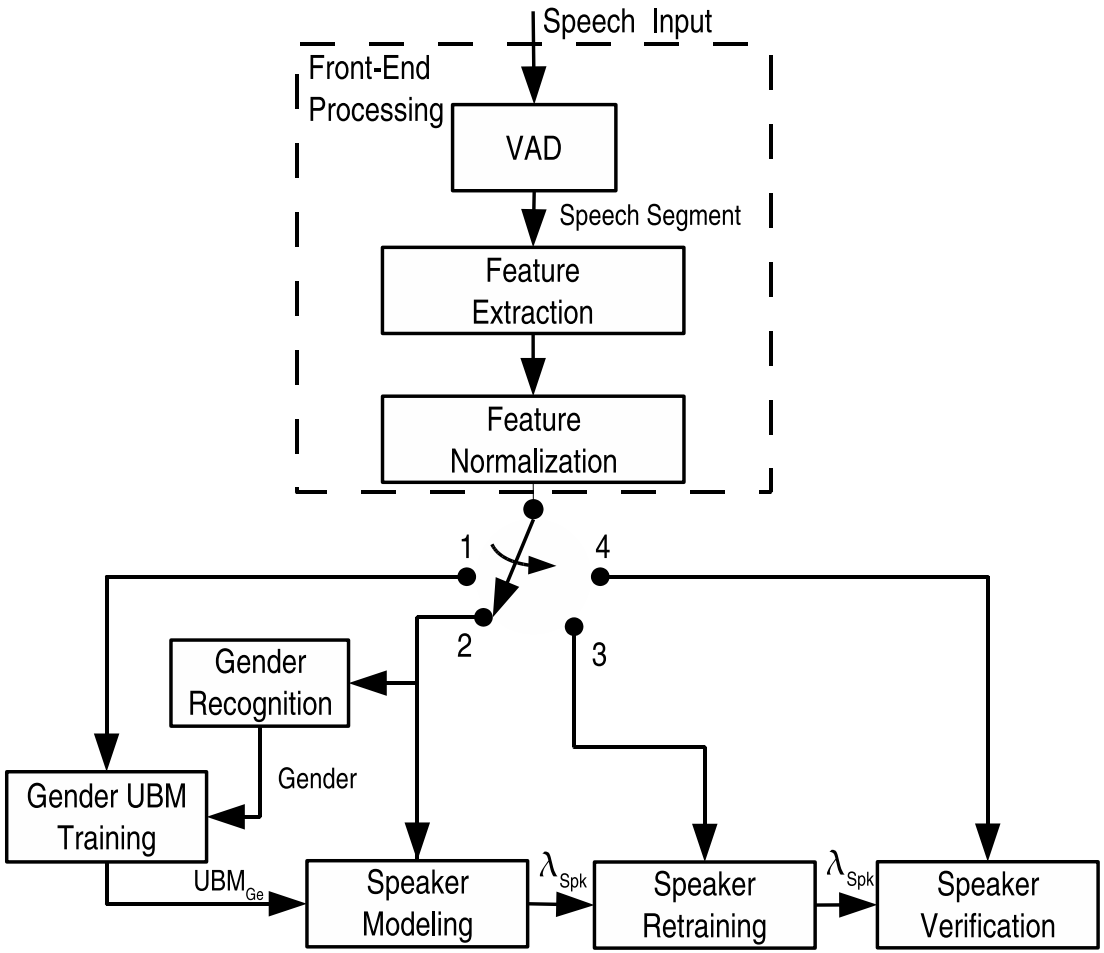
- Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation
- ATC System Presentation
- System Pattern Recognition Approach
- System Design
- Databases
- Results
- Conclusion
- References



Databases

Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation

ATC System Presentation

System Pattern Recognition Approach

System Design

Databases

Results

Conclusion

References

SPEECHDAT-AT: noisy telephone recordings

- Out of 100 speakers, 20 are marked as reference
- 6 utterances each are compared to the reference speaker
- $100*$ claimants $*6$ utterances each $*20$ reference $= 12000$ requests

WSJ0: almost clean database (Broadcast)

- All speakers produce the same utterances
- Out of 45 speakers, 24 are marked as reference
- 12 randomly selected utterances each are compared to the reference speaker
- $45*$ claimants $*12$ utterances each $*24$ reference $= 12960$ requests

Results

Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation

ATC System Presentation

System Pattern Recognition Approach

System Design

Databases

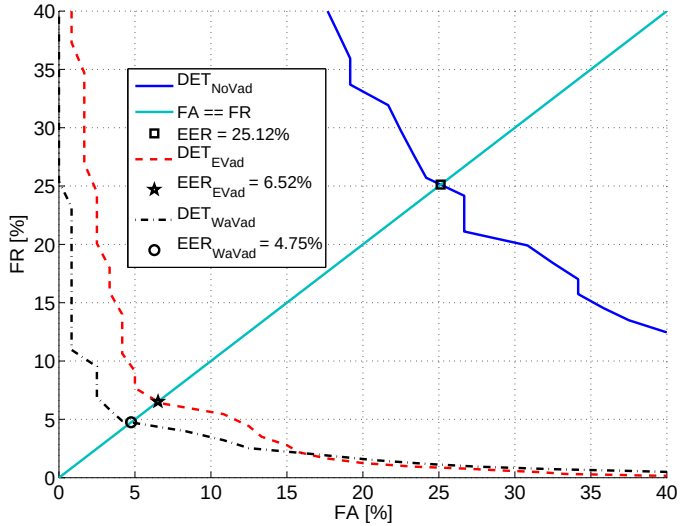
Results

Conclusion

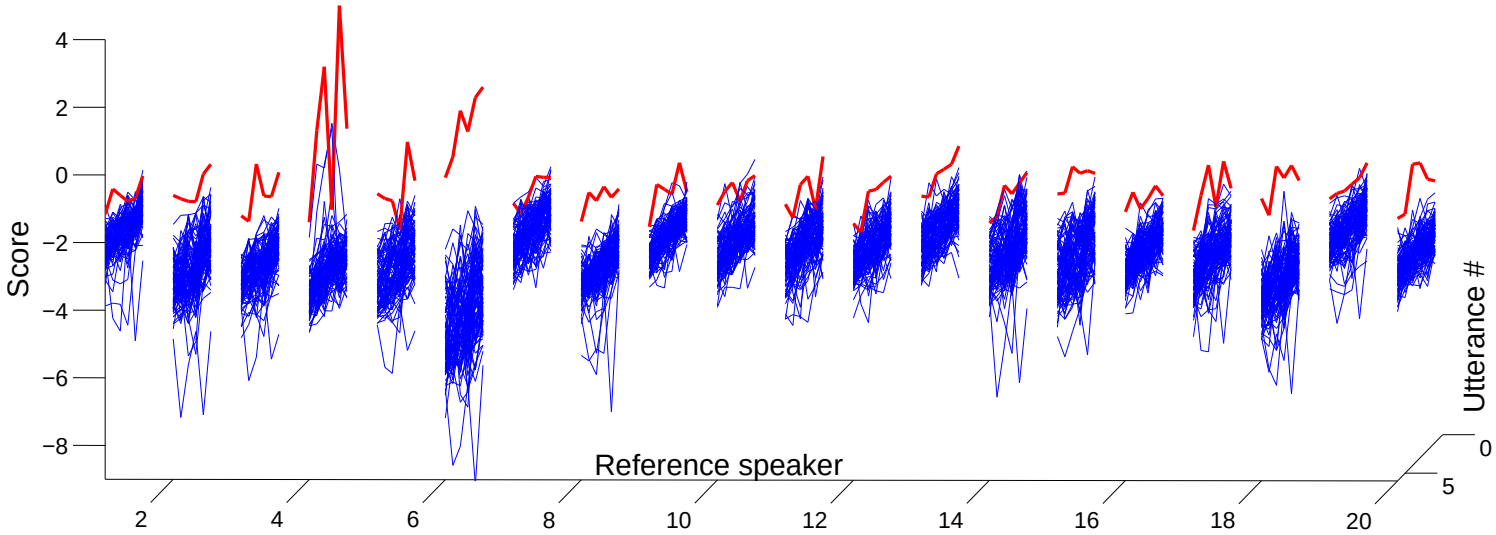
References

DET. . . Detection error tradeoff curve
FA. . . False acceptance rate
FR. . . False rejection rate

EER . . . Equal error rate (FA == FR)



Speaker Score Distribution



Conclusion

Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation

- ATC System Presentation
- System Pattern Recognition Approach
- System Design
- Databases
- Results
- Conclusion
- References

- System to choose is application dependent
- EER depends on test (database) condition
- Most systems assume known end points
- Text-idependent systems are still a challenge

References

Outline

Fundamentals and Preprocessing

Pattern Matching Methods

Stochastic Models

Air Traffic Control System Presentation

ATC System Presentation

System Pattern Recognition Approach

System Design

Databases

Results

Conclusion

References

[1] D.A. Reynolds,“Automatic speaker recognition: Current approaches and future trend” Proc. IEEE AutoID 2002, pp. 103-108, 2002.

[2] P.S. Aleksic and A.K. Katsaggelos,“Audio-Visual biometric”, Proceedings of the IEEE, 94(11), 2025-2044, 2006.

[3] J.P. Campbell, “Speaker recognition: A tutorial”, Proceedings of the IEEE, 85(9), pp. 1437-1462, 1997.

[4] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn,“Speaker Verificaiton Using Adapted Gaussian Mixture Models” Digital Signal Processing, 10, pp. 19-41, 2000.

[5] J.P. Campbell and F. Meade, “Speaker Recognition”, In A.K. Jain, R.M. Bolle, and S. Pankanti, editors, Biometrics: Personal Identification in Networked Society, pages 165-190, Kluwer Academic Press, Boston, 1999.

[6] Dellwo, V., Huckvale, M. and Ashby, M. “How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification”, Speaker Classification I, 2007, pp. 1-20

[7] Bimbot, F., Bonastre, J., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D. & Reynolds, D. A., “ Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing, 2000, pp. 19-41

[8] R. Dehak, N. Dehak, P. Kenny, P. Dumouchel,“Linear and Non Linear Kernel GMM SuperVector Machines for Speaker Verification”, Interspeech 2007, pp. 302-305