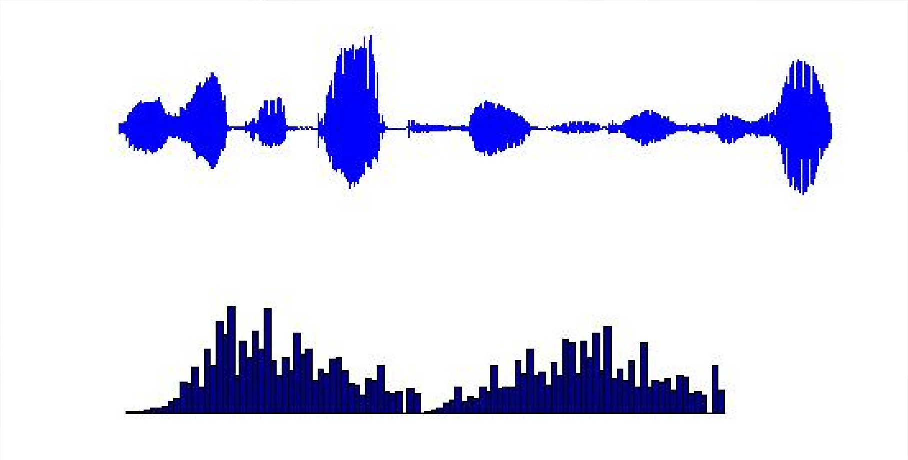
UCLA

EE 214A Project

Design and Implementation of Speaker Similarity Estimation System



Dept. of Electrical and Computer Engineering

Team members:

Yucong Wang Jingjing Zhang

Guanqun Yang Zhengtao Zhou

Contents

UCLA

UCLA

1

PART ONE

Problem Statements

1 / Backgounds

UCLA

Universally Used

It can be used as access control for physical facilities or computer networks and websites. The voice identification of rightful users can prevent the entrance of outsiders, which is more reliable than key or password. Another important application is transaction authentication. Voice authentication exhibited its superior compared to password or verification code since it is nearly impossible to copy.

Uniquely Identified

Different speakers can be identified through their speech because they have different vocal tract shapes, larynx sizes, and other parts of their voice production organs. Beside the physical differences, the manner of speaking of each speaker characterizes their speech, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on.

1 / Objective of Project

Database : 50 male speakers

1 / Related Works

1. Speaker recognition system based on GMM model
2. Mel-Frequency Cepstral Coefficients (MFCCs), LPCs, Pitch Frequency are widely used
3. Text-independent speaker recognition: From features to supervectors
4. Neural network and SVM classifier.

UCLA

2

PART TWO

Pre-processing

2 / Pre-processing

UCLA

▷ Remove silence and control noise

Based on two audio features:

Energy-Based:
•Filter out the intervals with relatively low energy.
•Work perfectly for high-quality recordings.
•Sensitive to noise.

Spectral centroid-Based:
•Spectral centroid is defined as the center of "gravity" of its spectrum.
•A measure of the spectral position, with high values corresponding to "brighter" sounds.
•Also help control noise.

Reference: Theodoros Giannakopoulos, "A method for silence removal and segmentation of speech signals", Department of Informatics and Telecommunications University of Athens, Greece

UCLA

▷ Attempt : remove babble noise

Wiener Noise Suppressor with TSNR&HRNR algorithms
Wiener filter based on tracking a priori SNR using Decision-Directed method

Two-step noise reduction (TSNR)
- Remove the reverberation effect while maintain the benefits of the decision-directed approach.
- But introduce harmonic distortion in the enhanced speech.

Harmonic regeneration noise reduction (HRNR)
- Refine the a priori SNR used to compute a spectral gain which is able to preserve the speech harmonics.

Problem : good by hearing but bad on results, might remove useful information in speech .

Reference :Plapous, C.; Marro, C.; Scalart, P., "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", IEEE Transactions on Audio, Speech, and Language Processing, Nov. 2006

3

PART THREE

Feature Selection and Extraction

3 / Feature Selection and Extraction

- **Two Categories of Speech Features**
  - High-level features
  - Low-level features

- **Low-level features are available**

- **Different types of low level features**
  - Pitch
  - Formants
  - Subglottal resonance frequency
  - Voice quality features
  - Others: MFCC, LPC, LPCC

Reference: Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. IEEE Signal processing magazine, 32(6), 74-99.

4

PART FOUR

Training and Testing

4 / Training and Testing

- **Similarity metrics and Classification algorithm**

- **Scheme 1**

4 / Training and Testing

Main idea:

Use the difference between two speech files and train the model to predict the output.

Steps:

- Extract features from each speech file, convert them into feature vectors
- Compute distance between two feature vectors
- Try different features and train different models
- Test the result and give conclusions

4 / Training and Testing

UCLA

Vector Distance

Using threshold method to test the effect of different distance on features

- Euclidean distance

Feature	MFCCs	LPCs	LPCCs	Pow Spec
FPR	65.9%	74.8%	74.5%	70%
FNR	88.1%	74.8%	74.8%	56%

- Cosine distance

Feature	MFCCs	LPCs	LPCCs	Pow Spec
FPR	59.3%	79.3%	67.8%	74.8%
FNR	83.7%	76.3%	83.7%	75.6%

4 / Training and Testing

Training model and testing result

- SVM

Feature	MFCCs	LPCs	LPCCs	Pow Spec
FPR	1.43%	2.25%	2.19%	2.12%
FNR	88.1%	86.7%	94.1%	96.3%

- Neural network

Feature	MFCCs	LPCs	LPCCs	Pow Spec
FPR	10%	3%	6.4%	4.1%
FNR	64.4%	82%	73.3%	90.3%

4 / Training and Testing

Conclusion

The best result using this method is extracting MFCCs ,then calculating cosine distance and applying Neural network, which achieves FPR = 5.6% and FNR = 62%.

Two main Problems:

- Calculating mean of features makes dimension alignment incorrect and more robust methods are required
- Training dataset has far more 0's than 1's, making it more likely for model to make false negative prediction and this imbalance should be addressed.

4/ Training and Testing

UCLA

- Similarity metrics and Classification algorithm

- **Scheme 2**

4 / Training and Testing

UCLA

Steps:

C D E F G

1. Remove silence and control noise
2. Feature extraction of speech signal
3. **Design similarity metrics**
4. **Classification algorithm**
5. Test and improvement

UCLA

Feature Extraction

Mel-Frequency Cepstral Coefficients / MFCC

Linear Prediction Coefficients / LPC

F_0 and First 3 Formants

Spectral Subband Centroids / SSC

Linear Prediction Cepstral Coefficients / LPCC

Power spectrum

MFCC + LPC perform best.

Reference: Speech feature extraction functions for ASR and speaker identification.
http://www.practicalcryptography.com/miscellaneous/machine-learning/matlab_speech_features-documentation/

4 / Training and Testing

▷ Similarity metrics

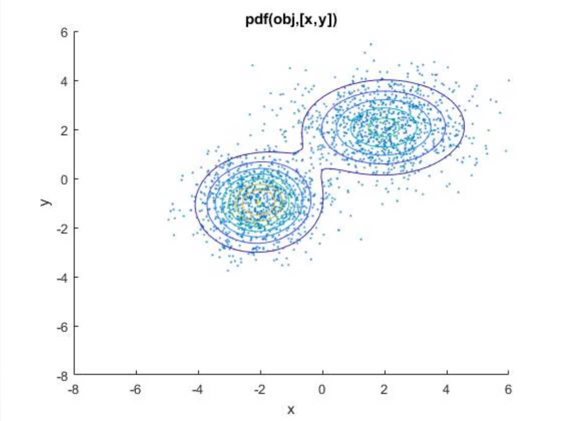
Likelihood
between GMM model and feature set

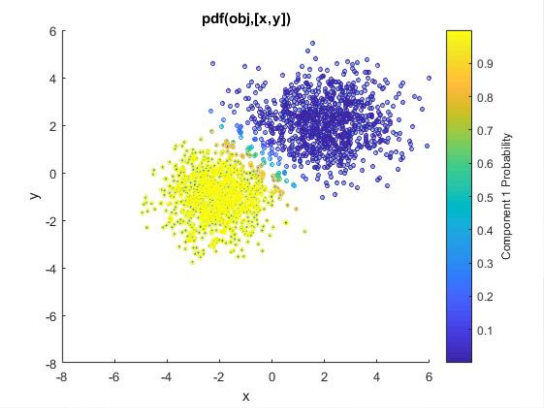
Gaussian Mixture Model / GMM
Model the feature distribution and fit original signal.

$$p(z|\lambda) = \sum_{i=1}^{M} \alpha_i N(z; \mu_i,, \Sigma_i)$$

z: feature vector
λ: speaker model
N: Gaussian function with mean vector μ and covariance matrix Σ
α_i: component density
M: the number of mixtures

Posterior
The posterior probabilities of each component in the Gaussian mixture distribution.





4 Training and Testing

▷ Classifier

SVM
Trains or cross-validates a support vector machine (SVM) model for two-class (binary) classification on a low- through moderate-dimensional predictor data set.
Train feature: likelihood matrix. Train target : labels (0 or 1)

Problem: High dimension data set

SVD
Singular value decomposition
perform a singular value decomposition of matrix, reduce dimensions.

Problem: Ignore feature importance and uniqueness.

Results tend to be almost all zero(5% and 90%), bad prediction.

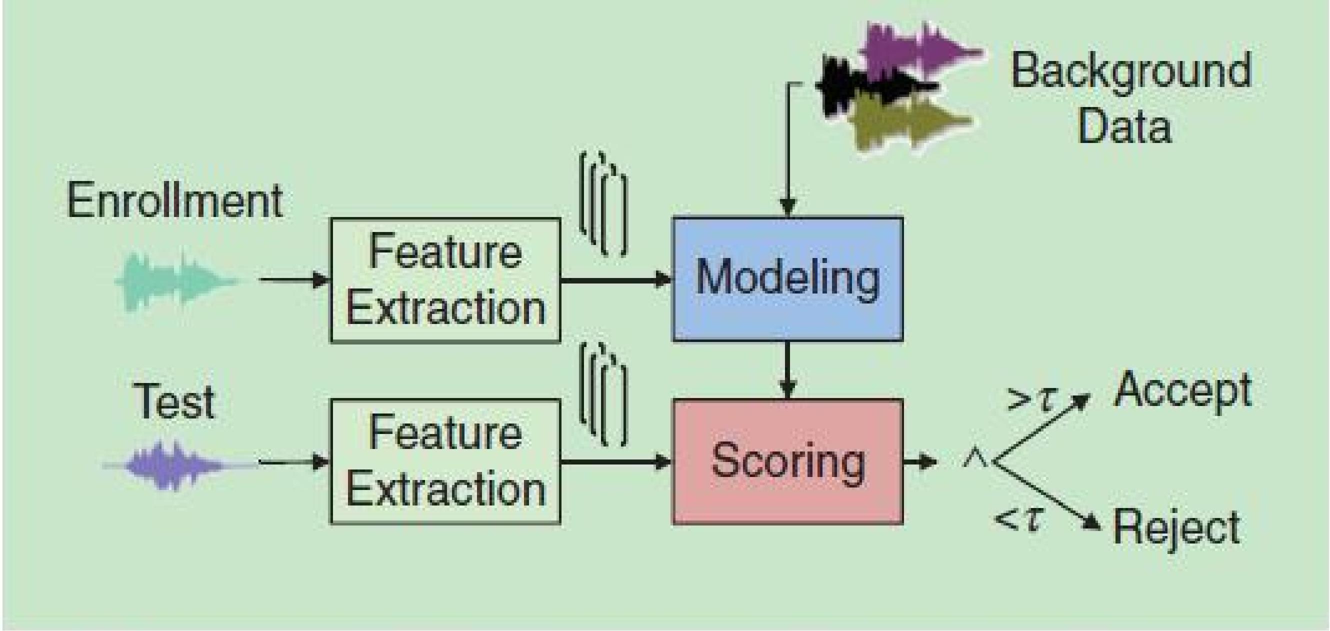
4 / Training and Testing

- **Similarity metrics and Classification algorithm**

  - **Scheme 3**

4 / Training and Testing

Automatic Speaker Verification (ASV)
Determine whether a given pair of utterances are from the same speaker or two different speakers (binary decision)



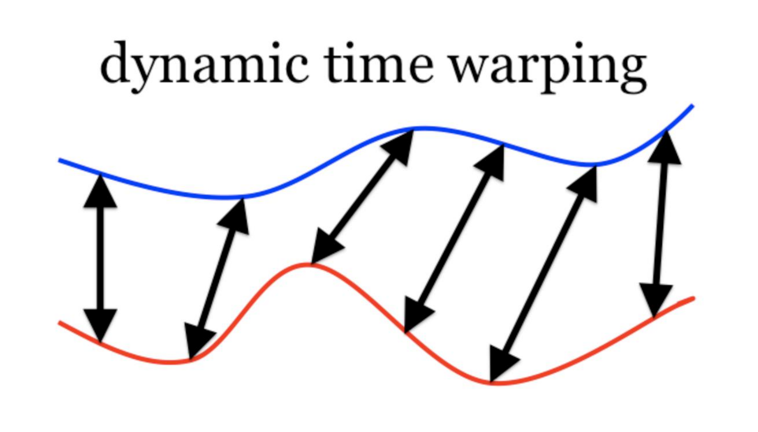
[FIG4] An overall block diagram of a basic speaker-verification system.

(Hansen and Hasan, 2015)

4 / Training and Testing

UCLA

▷ Similarity metrics



DTW distance:
Dynamic Time Warping

- Warped non-linearly in the time dimension
- Measures a distance-like quantity between two given sequences.
- Find the best mapping with the minimum distance.
- Good for speech differentiation

Features are vectors of time series and we need to compress or expand in time in order to find the best mapping.

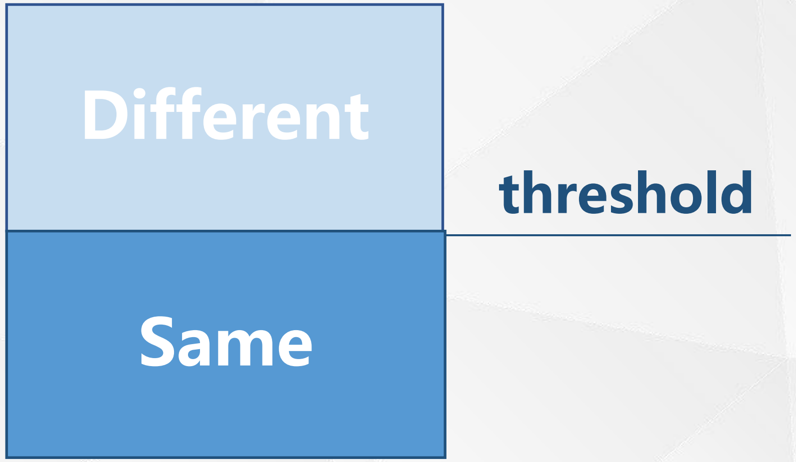
4/ Training and Testing

▷ Classifier

Threshold
------ naïve but performs well

Compute a threshold value as a boundary condition by equal error rate (EER).

Lower than/equal to the distance threshold → same speaker
Higher than the distance threshold → different speaker



5

PART FIVE

Result and Conclusion

5 / Results

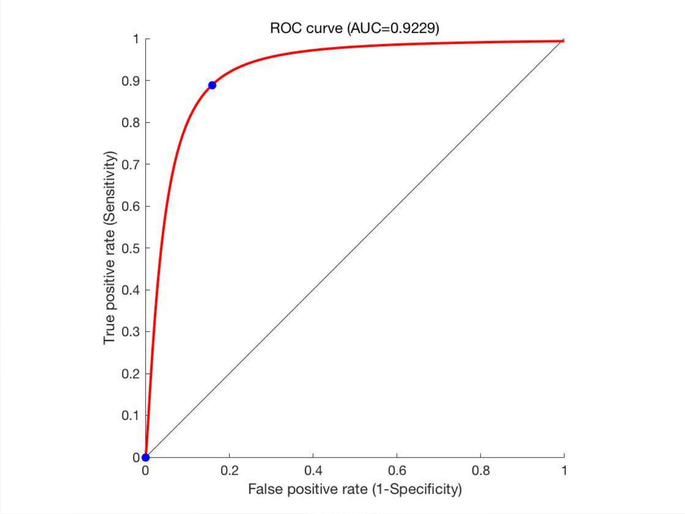
Train	Test	false positive rate	false negative rate
Clean	Clean	15%	11%
Clean	Babble	36%	31%
Multi	Clean	13%	12%
Multi	Babble	33%	36%

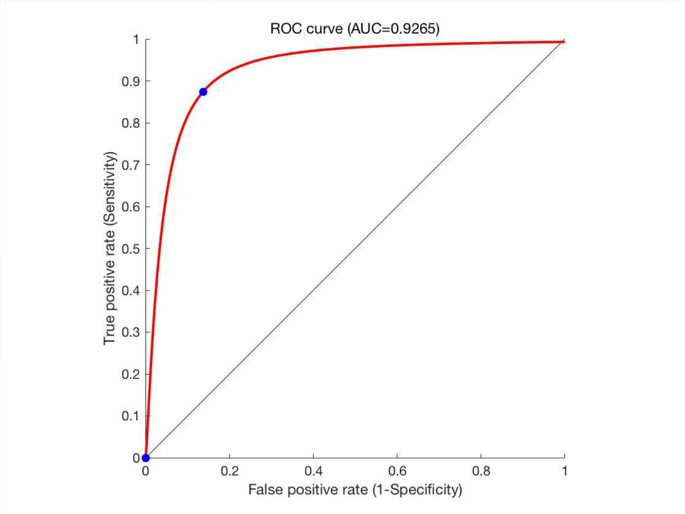
The method performs well on clean data, but less satisfying on noise data.
- The suppression of noise is not perfect.
- Features are not robust on noise condition.

5 / Results

ROC curves

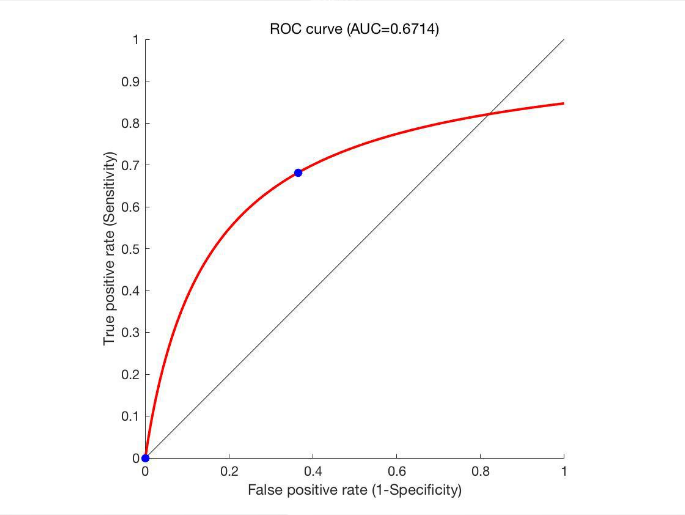
Train: clean
Test : clean

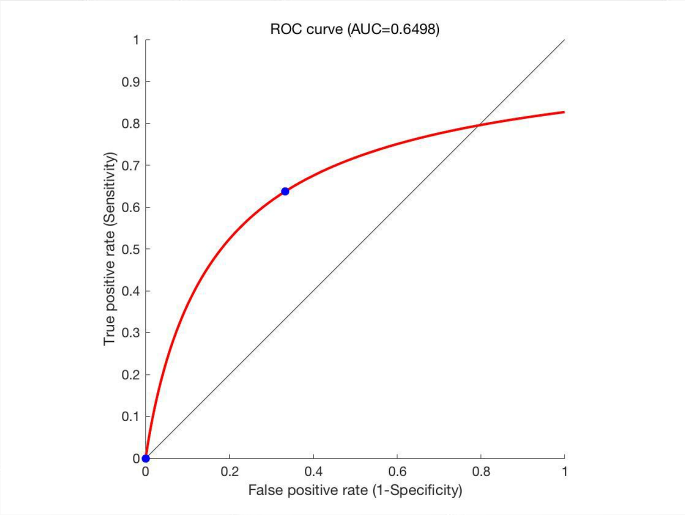




Train: multi
Test : clean

Train: clean
Test : babble





Train: Multi
Test : babble

5 / Conclusion

Existing problems

Noise condition
Feature robustness
Not good stability on classifier

Future works

Noise suppression
Speaker modeling (GMM / UBM / Identity Vector)
Dimension reduction method on feature vector

THANKS