Speaker Recognition

SRT project of Signal Processing

Xinyu Zhou, Yuxin Wu, Tiezheng Li

Department of Computer Science and Technology
Tsinghua University

January 6, 2016

Overview

Speaker Recognition

Identification of the person who is speaking by characteristics of their voice biometrics.

Primary Goal:

- **Accurate** and **Efficient Short** utterance speaker recognition.

Additional Goal:

- Scalability over large number of speakers.
- Low Latency Real-Time recognition
- A working Real-Time recognition system.

Content

VAD

Voice Activity Detection shall be applied for all signals as a pre-filter. We've tried 2 different approaches:

- Energy-Based:
  - Filter out the intervals with relatively low energy.
  - Work perfectly for high-quality recordings.
  - Sensitive to noise.
- Long-Term Spectral Divergence
  - Compare long-term spectral envelope with noise spectrum.
  - More robust to noise, used in our GUI.
  - *Efficient voice activity detection algorithms using long-term speech information, Ramırez, Javier, 2004*

VAD

Voice Activity Detection shall be applied for all signals as a pre-filter. We've tried 2 different approaches:

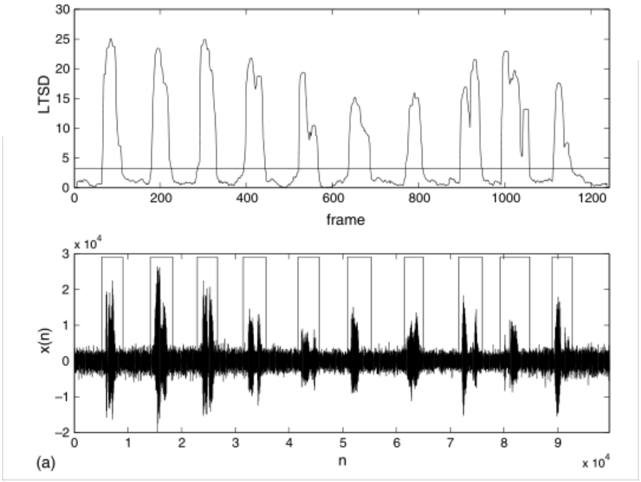
- Energy-Based:
  - Filter out the intervals with relatively low energy.
  - Work perfectly for high-quality recordings.
  - Sensitive to noise.
- Long-Term Spectral Divergence
  - Compare long-term spectral envelope with noise spectrum.
  - More robust to noise, used in our GUI.
  - *Efficient voice activity detection algorithms using long-term speech information, Ramırez, Javier, 2004*

VAD

Voice Activity Detection shall be applied for all signals as a pre-filter. We've tried 2 different approaches:

- Energy-Based:
  - Filter out the intervals with relatively low energy.
  - Work perfectly for high-quality recordings.
  - Sensitive to noise.
- Long-Term Spectral Divergence
  - Compare long-term spectral envelope with noise spectrum.
  - More robust to noise, used in our GUI.
  - *Efficient voice activity detection algorithms using long-term speech information, Ramırez, Javier, 2004*

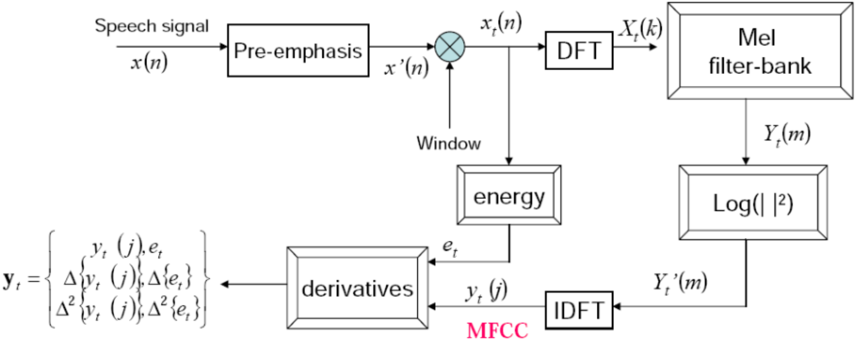
LTSD


30
25
20
15
10
5
0
LTSD
0
200
400
600
800
1000
1200
frame
x 10^4
3
2
1
0
−1
−2
x(n)
0
1
2
3
4
5
6
7
8
9
x 10^4
n
(a)

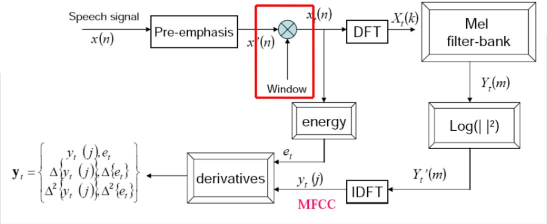
MFCC

Mel-Frequency Cepstral Coefficients

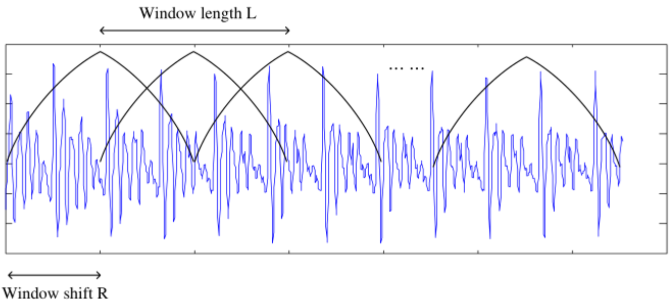
Cepstral feature which closely approximates human auditory system's response. Commonly used feature for Speech/Speaker Recognition.



Windowing

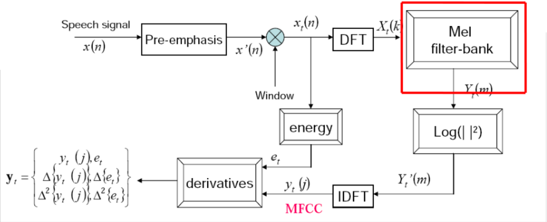


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
MFCC
energy
e_t
derivatives
$\mathbf{y}_t = \begin{Bmatrix} y_t(j), e_t \\ \Delta\{y_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t(j)\}, \Delta^2\{e_t\} \end{Bmatrix}$

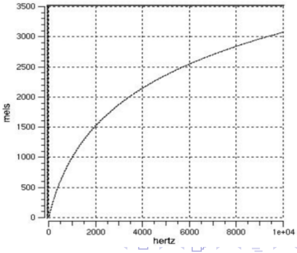


Window length L
… …
Window shift R

Mel-Scale



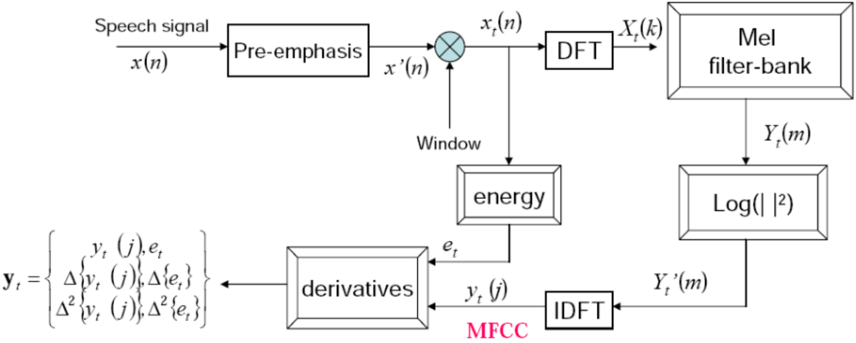
$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700})$$



MFCC

Mel-Frequency Cepstral Coefficients

Cepstral feature which closely approximates human auditory system's response. Commonly used feature for Speech/Speaker Recognition.



LPC

Linear Predictive Coding/Coefficients

Assumption

In a short period, the nth signal is a linear combination of previous p signals: $\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i)$

Minimize squared error $\mathrm{E}\left[\hat{x}(n) - x(n)\right]$ using Levinson-Durbin algorithm.

Use $a_1, \cdots, a_p$ as features.

LPC

Linear Predictive Coding/Coefficients

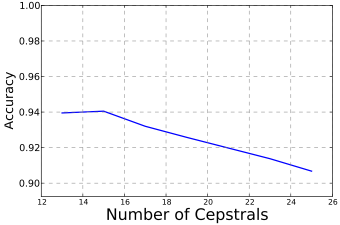
Assumption

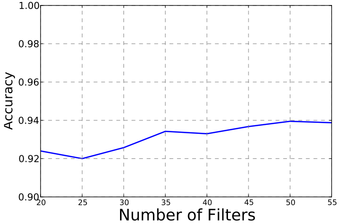
In a short period, the nth signal is a linear combination of previous p signals: $\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i)$

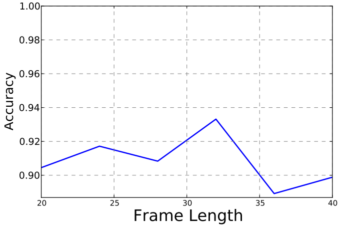
Minimize squared error $\mathrm{E}\left[\hat{x}(n) - x(n)\right]$ using Levinson-Durbin algorithm.

Use $a_1, \cdots, a_p$ as features.

MFCC Params

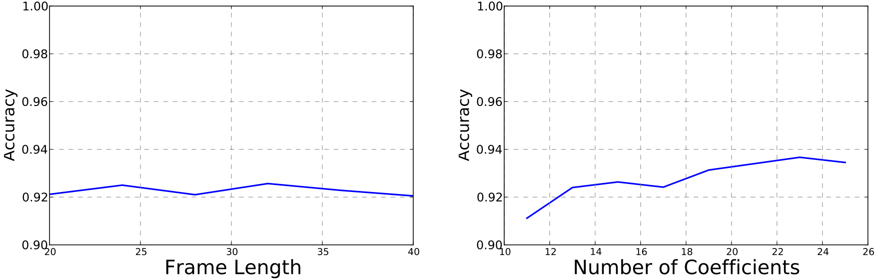






Best parameters in our cases:
Number of cepstrals: 15
Number of filters: 55
Frame length: 32ms

LPC Params

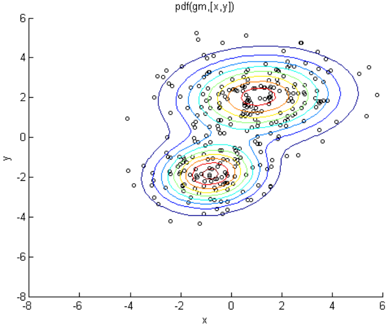


Best parameter in our cases:
Number of coefficients: 23
Frame length: 32ms

GMM

Gaussian Mixture Model is commonly used to model human's acoustic feature.

$$p(\theta) = \sum_{i=1}^{K} w_i \mathcal{N}(\mu_i, \Sigma_i)$$

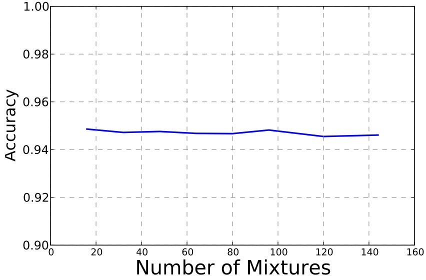


We use $K = 32$

GMM

Gaussian Mixture Model is commonly used to model human's acoustic feature.

$$p(\theta) = \sum_{i=1}^{K} w_i \mathcal{N}(\mu_i, \Sigma_i)$$



We use $K = 32$

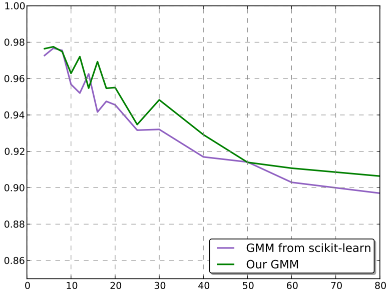
Optimized GMM

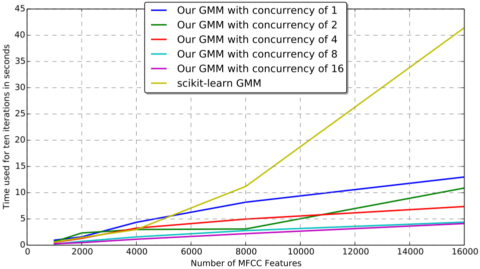
- Basic GMM training: random initialize, estimate parameters with EM.
- Improvment: initialize with a parallel KMeansII.
- Improvment: parallel training implementation in C++.
- Compared to GMM from scikit-learn:

Arthur, David, Sergei, 2007, k-means++: The advantages of careful seeding.
Bahmani, et. al, 2012, Scalable K-means++

Optimized GMM

- Basic GMM training: random initialize, estimate parameters with EM.
- Improvment: initialize with a parallel KMeansII.
- Improvment: parallel training implementation in C++.
- Compared to GMM from scikit-learn:





Arthur, David, Sergei, 2007, k-means++: The advantages of careful seeding.
Bahmani, et. al, 2012, Scalable K-means++

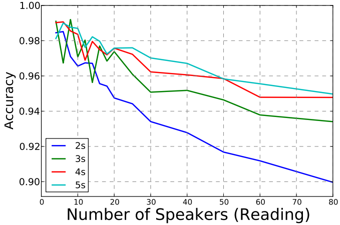
UBM

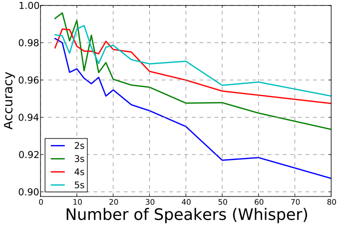
Universal Background Model is a GMM trained on giant datasets.
UBM can be used to:

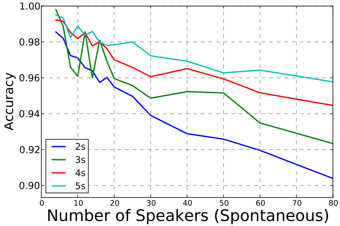
- Describe general acoustic feature of human.
- Reject the decision of GMM.
- Train adaptive GMM.

Reynolds, Douglas, et al, 2000,
Speaker verification using adapted Gaussian mixture models

GMM Results







Train duration: 20s

Random selected test utterance: 50

Each value in the graph is an average of 20 independent experiments.

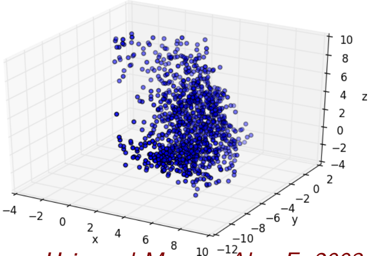
CRBM

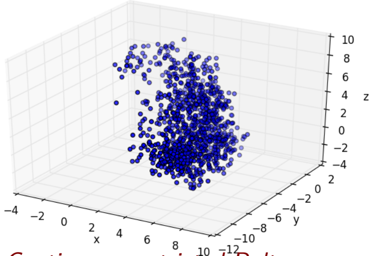
- **Restricted Boltzmann Machine** is a generative stochastic two-layer neural network.
- **Continuous RBM** extends RBM to real-valued inputs.
- RBM has the ability to reconstruct a layer similar to input layer. The difference between the two layers can be a used to measure the fitness of an input to the model.
- Therefore, RBM can be a substituion to GMM.

Chen, Hsin and Murray, Alan F, 2003, Continuous restricted Boltzmann machine with an implementable training algorithm

CRBM

- **Restricted Boltzmann Machine** is a generative stochastic two-layer neural network.
- **Continuous RBM** extends RBM to real-valued inputs.
- RBM has the ability to reconstruct a layer similar to input layer. The difference between the two layers can be a used to measure the fitness of an input to the model.
- Therefore, RBM can be a substituion to GMM.

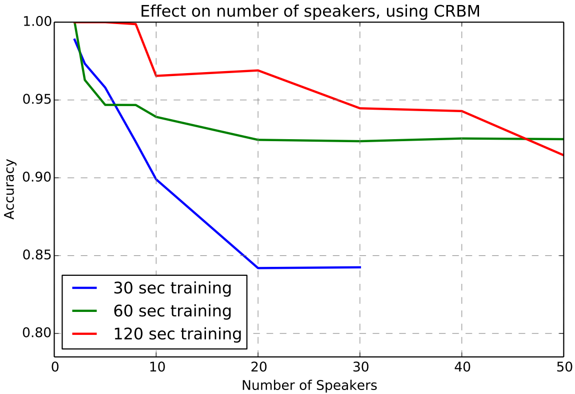




Chen, Hsin and Murray, Alan F, 2003, Continuous restricted Boltzmann machine with an implementable training algorithm

RBM Results

Results of CRBM, tested with 5 secs of utterance.



GUI Demo

Conclusion

- We implemented a faster GMM, also with better performance.
- Accuracy is kept even under short training and testing utterance.
- Our system is highly accurate, can almost response in real-time.
- 97% accuracy for 20$\sim$30 speakers, 95% for 70$\sim$80 speakers.

Thanks!