The USTC Systems for The NIST-2005 Speaker Recognition Evaluation

Beiqian Dai, Yanlu Xie, Xi zhou,
Zhiqiang Yao, Jixu Chen, Minghui Liu

Introduction

Participant Task:

		Test Segment Condition			
		10 sec 2-chan	1 conv 2-chan	1 conv summed-chan	1 conv aux mic
Training Condition	10 seconds 2-channel	○	○		
	1 conversation 2-channel	○	○		
	3 conversation 2-channel	○	○		
	8 conversation 2-channel	○	○		
	3 conversation summed-channel		○	○	

USTC SSIP Lab.

One Speaker System

Main Modules

- FrontEnd Processing
- Universal Background Model Training
- Speaker Model Adaptation
- LLR Score Computation
- Fusion
- Making Decision

FrontEnd Processing

- FrontEnd Processing for MFCC
- FrontEnd Processing for Pitch
- FrontEnd Processing with Wavelet

FrontEnd Processing for MFCC

- Band-limited (300Hz – 3400Hz)
- MFCC+Delta(16+16) with the 0th removed
- RASTA
- CMS
- Remove Silence
- Kurtosis Normalization

Silence Removal

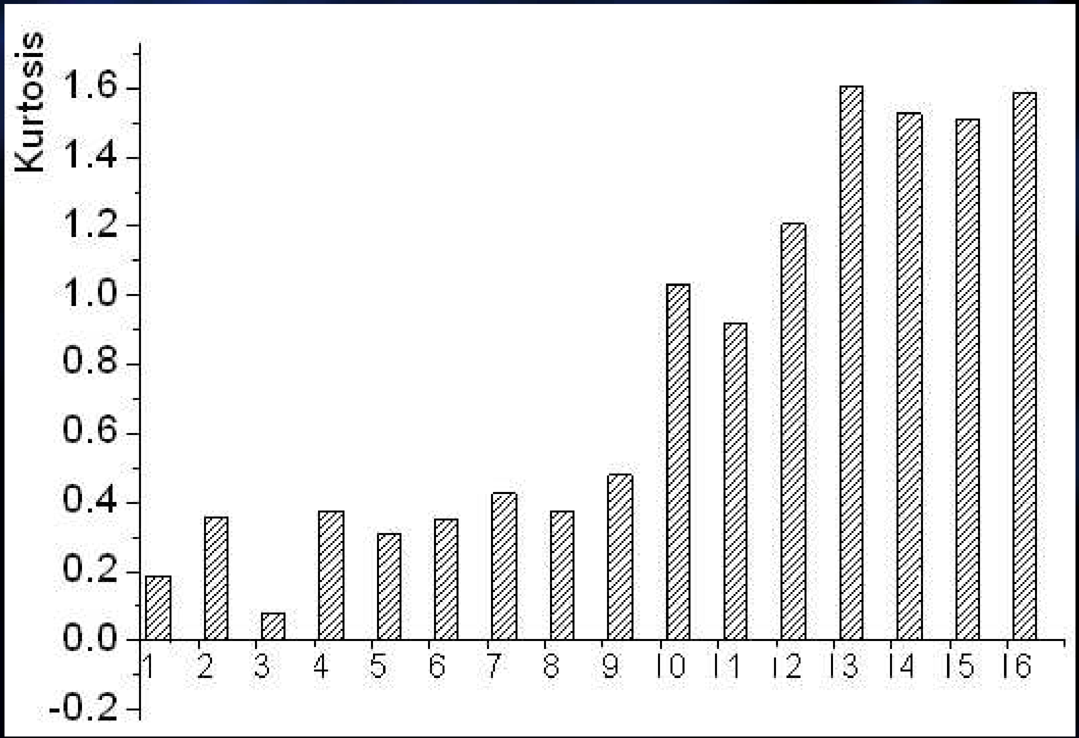
- Energy based threshold to remove long period silence

- Predictive Segment
  - H0 : current frame is a new segment first frame
  - H1: current frame is belong to previous segment
  - |Xt – Seedt-1| < |Xt – O| , choose H0,
  - Else, choose H1
- Energy & Duration based threshold to remove silence segment

Kurtosis Normalization

The kurtosis of a random variable x is defined as

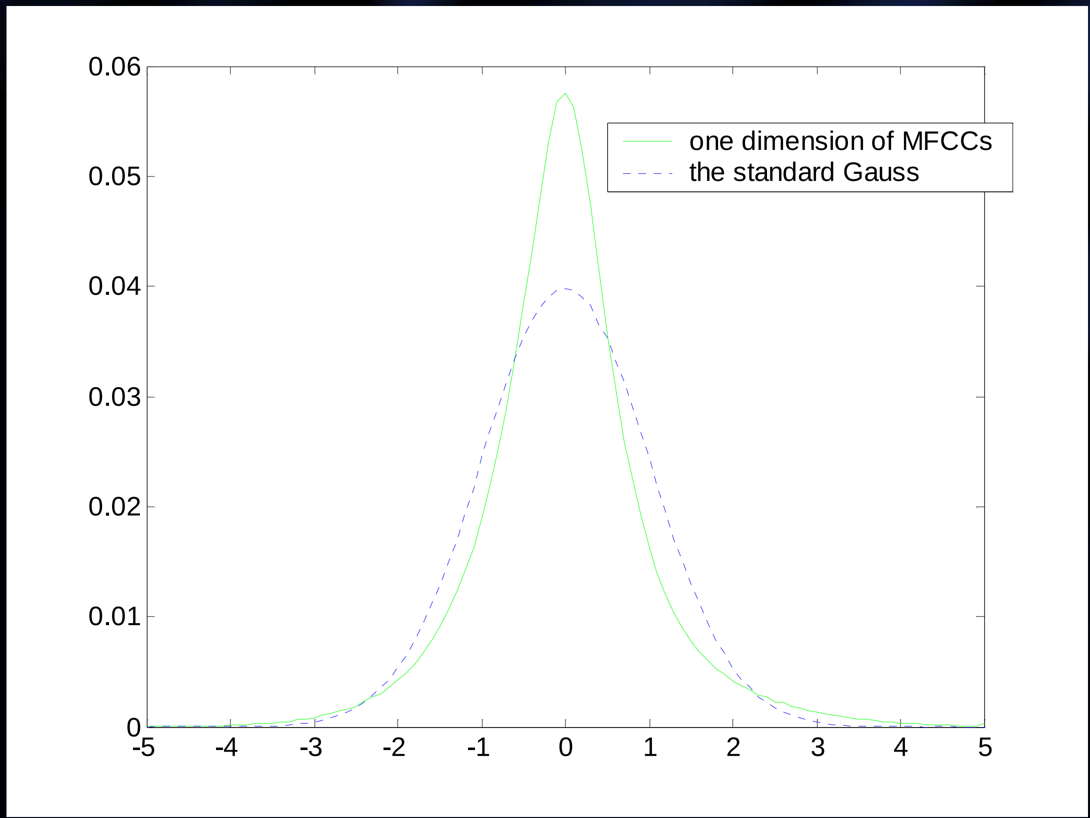
$$K(x) = \frac{E(x^4)}{E(x^2)^2} - 3$$



If a random variable has a kurtosis less than zero, it is termed platykurtic i.e. sub-Gauss. If it has kurtosis greater than zero, it is termed leptokurtic i.e. super-Gauss. Speech signals are generally leptokurtic, so are speech cepstral parameters.

Kurtosis Normalization

The comparison of pdfs between MFCCs and the standard normal.

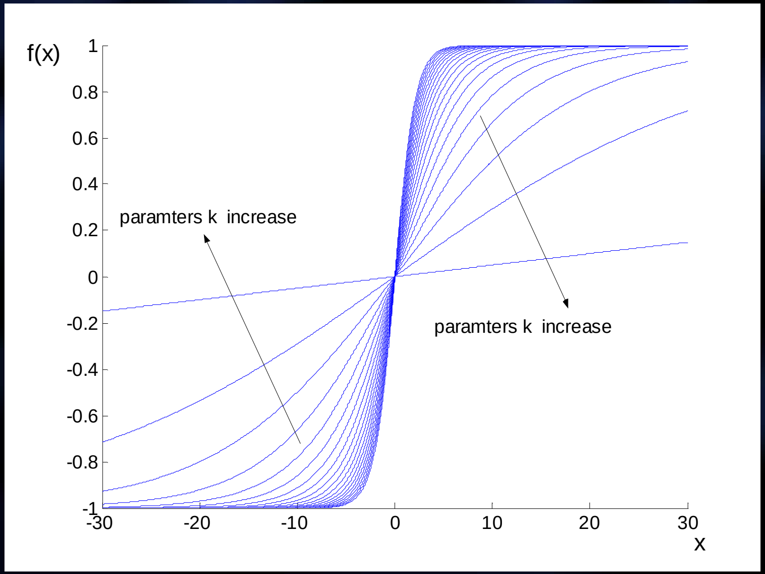


Kurtosis Normalization

the sigmoid functions

$$f(x) = \frac{a}{1+\exp(-kx)} - b$$

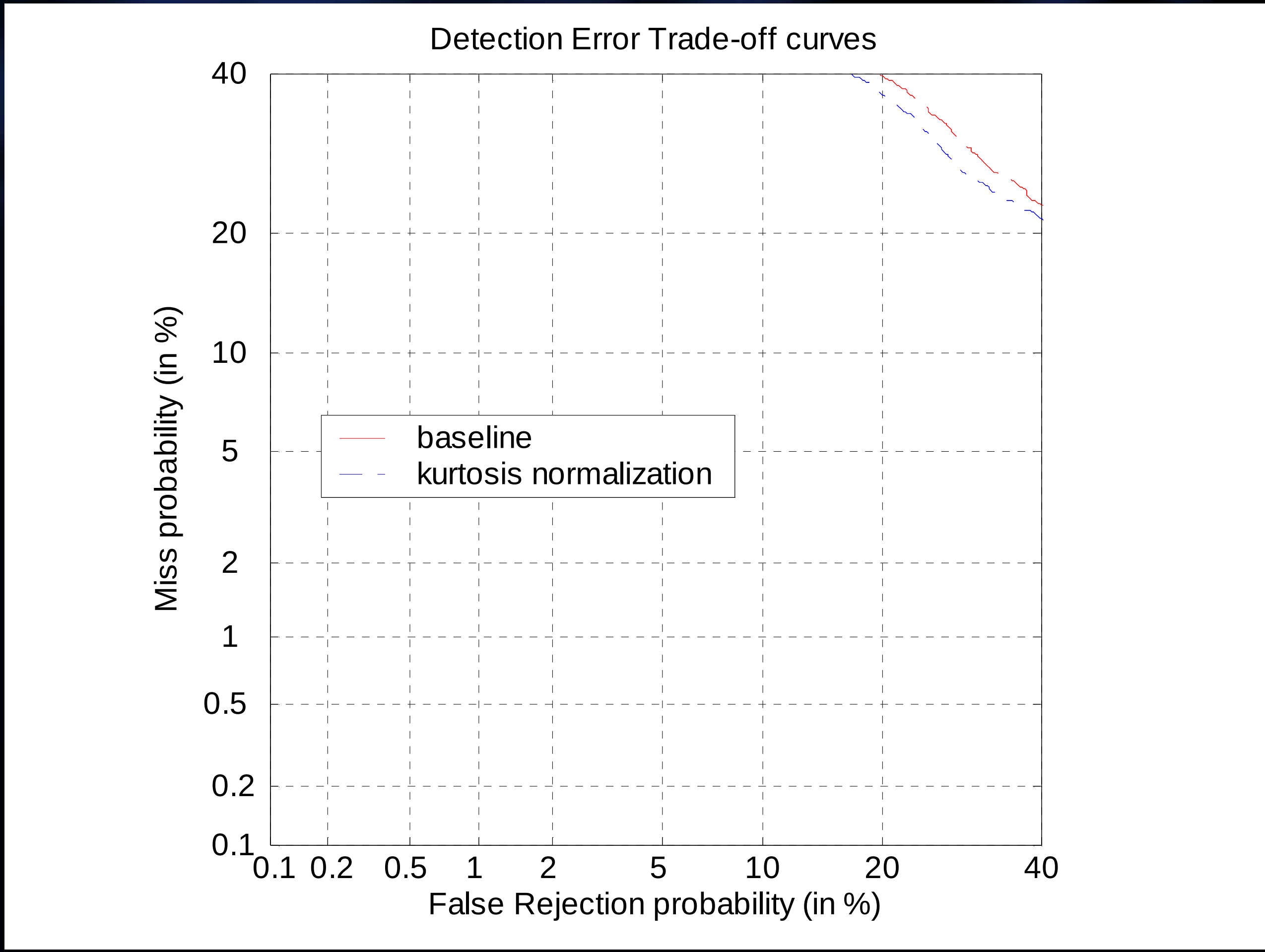
where a and b are constant coefficients, k>0. In order to keep the means of speech parameters invariable, coefficients a and b are chosen to be 2 and 1 respectively.



We have proved that the optimization the parameter k of the sigmoid functions can make the kurtosis be zero for speech parameters.

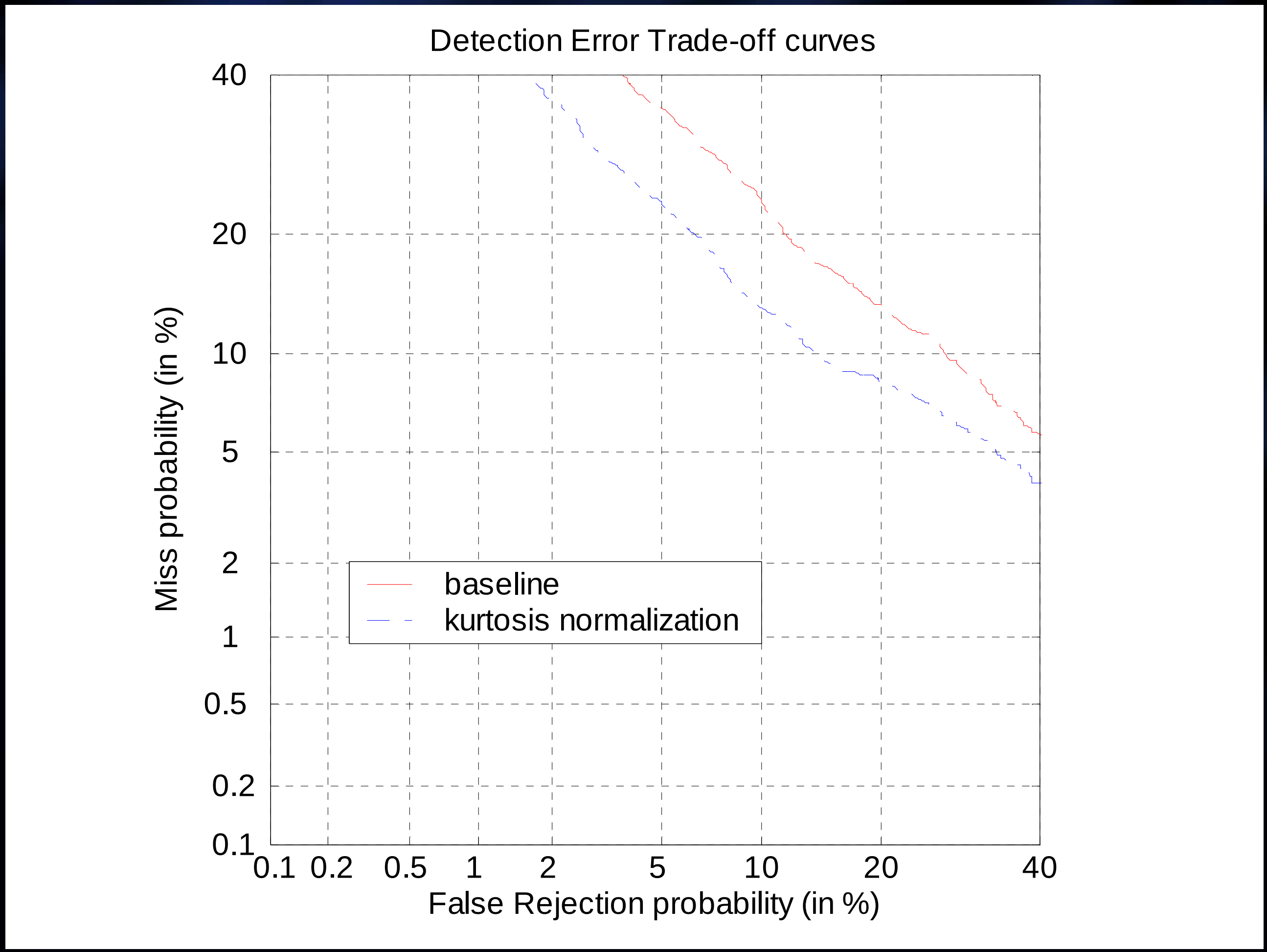
Kurtosis Normalization

Experiment on NIST'04 10seconds-10seconds male database



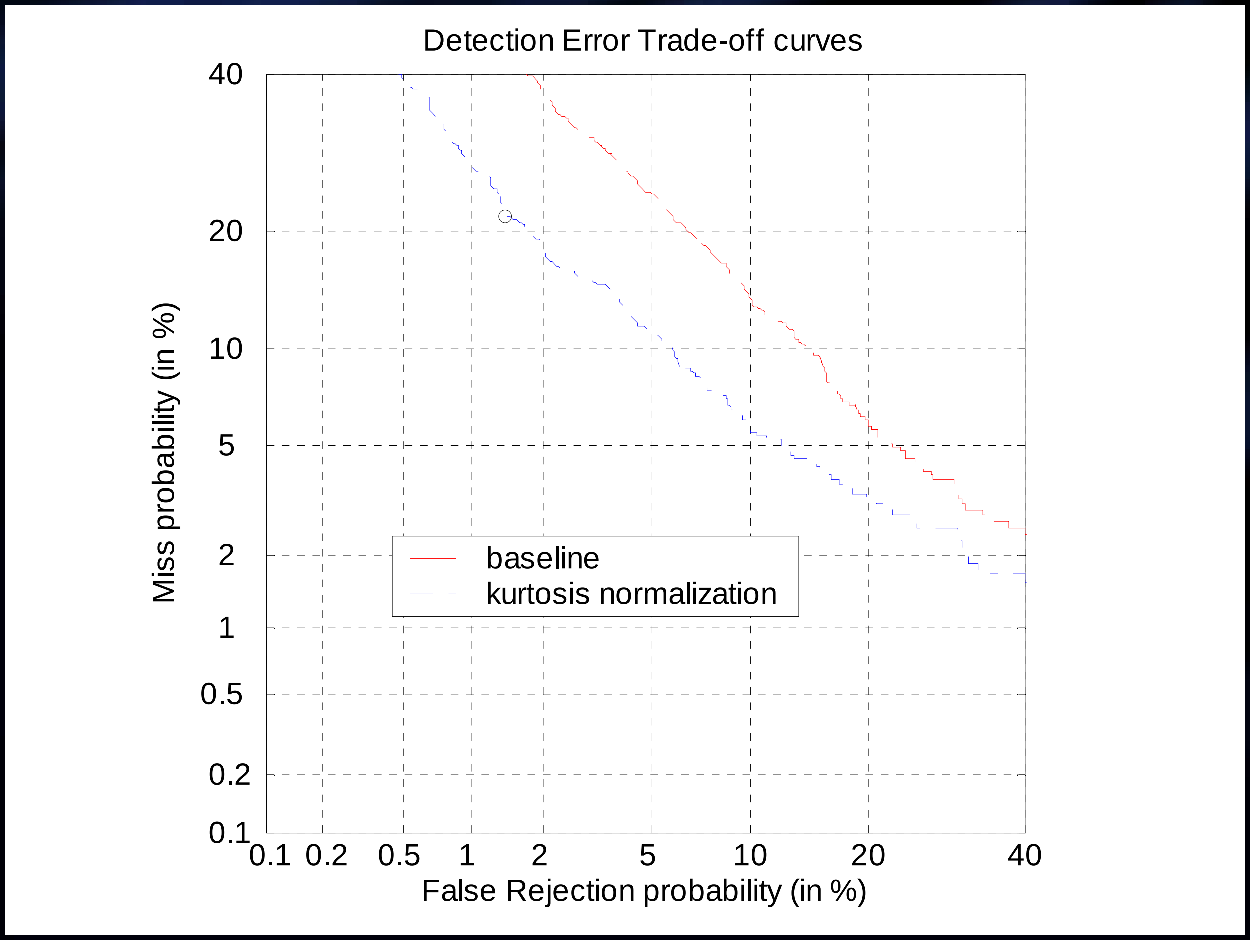
Kurtosis Normalization

Experiment on NIST'04 1conv-1conv male database



Kurtosis Normalization

Experiment on NIST'04 8conv-1conv male database



Kurtosis Normalization

Maybe the more speech is used, the performance of the system is improved further with kurtosis normalization method.

FrontEnd Processing for pitch

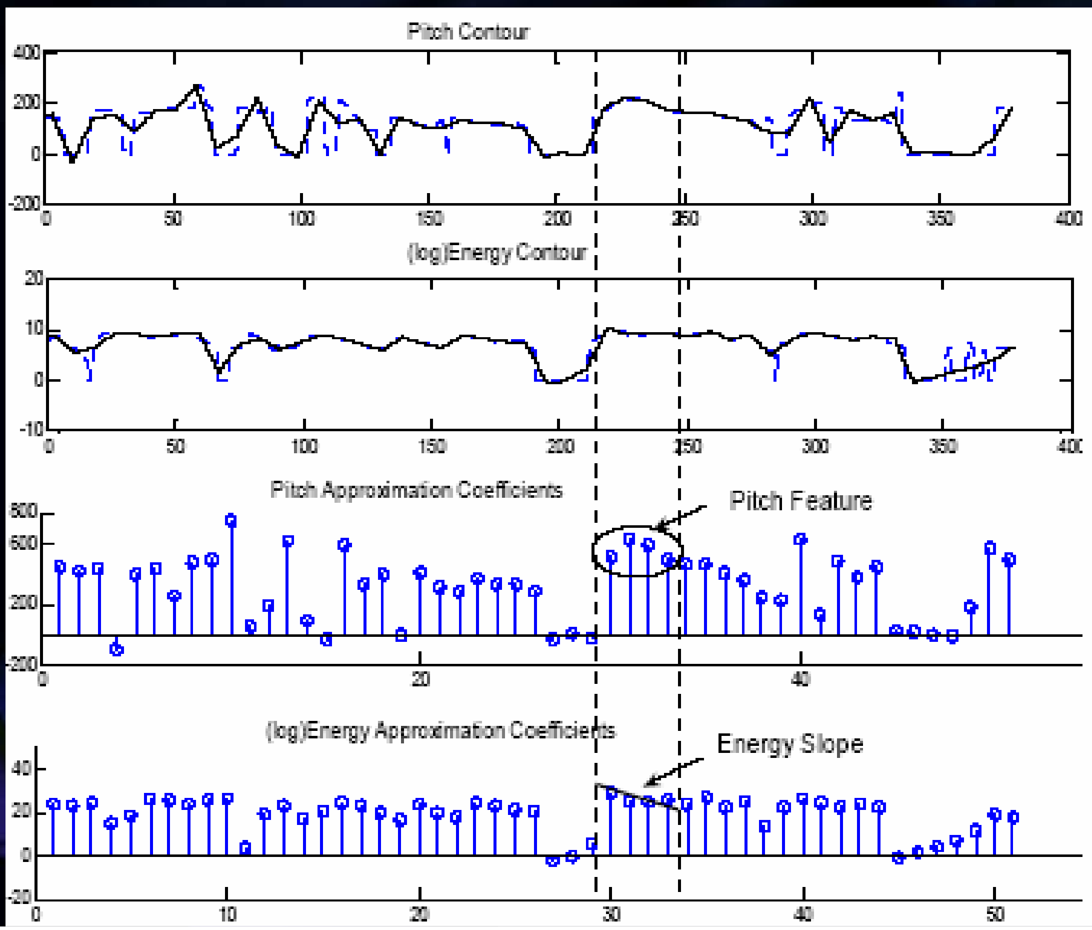
We firstly split pitch and energy contours into segment with 7 frames length. 4 parameters related to pitch were extracted:

- log (mean_F0) averaged over a segment
- log (max_F0) of a segment
- log (min_F0) of a segment
- F0_slop of a segment

Another 4 parameters related to energy are extracted as above. Total 8 parameters of a segment comprise an 8-dimension vector.

FrontEnd Processing with wavelet

We made wavelet analysis of the f0 and energy contour. Subsequently, the prosodic features were extracted only from the 3rd level approximation coefficients



Prosodic Feature:

[cA1 cA2 cA3 cA4 ESlope]

23系SSIP实验室

Universal Background Model

- Model Type
  - GMM consist of 2048 mixtures (1conv)
  - GMM consist of 512 mixtures (10seconds)
  - UBM_F for female and UBM_M for male
- Training data
  - Selected from NIST'03&04 training and test data
- Training Algorithm
  - EM Algorithm

Speaker Model Adaptation

- Model Type
  - Same as UBM
- Training data
  - Training data in NIST’05
- Training algorithm
  - MAP from UBM_M or UBM_F

LLR Score Computation

- Log Likelihood Ratio

$$\Lambda(\mathbf{O}) = \frac{1}{T}\sum_{t=1}^{T}\left(\log p\left(\mathbf{O}_t \mid \lambda_{tar}\right) - \log p\left(\mathbf{O}_t \mid \lambda_{UBM}\right)\right)$$

- TNORM
  - A speaker-specific T-norm selection
  - The closest set of P cohort models are used to Tnorm during run time where P is chosen to be 50.

Fusion

- The scores from the sub-systems are fused with a perceptron classifier. The number of input nodes of the perceptron is the same as the number of sub-systems applied. There is no hidden layers and only one output node.

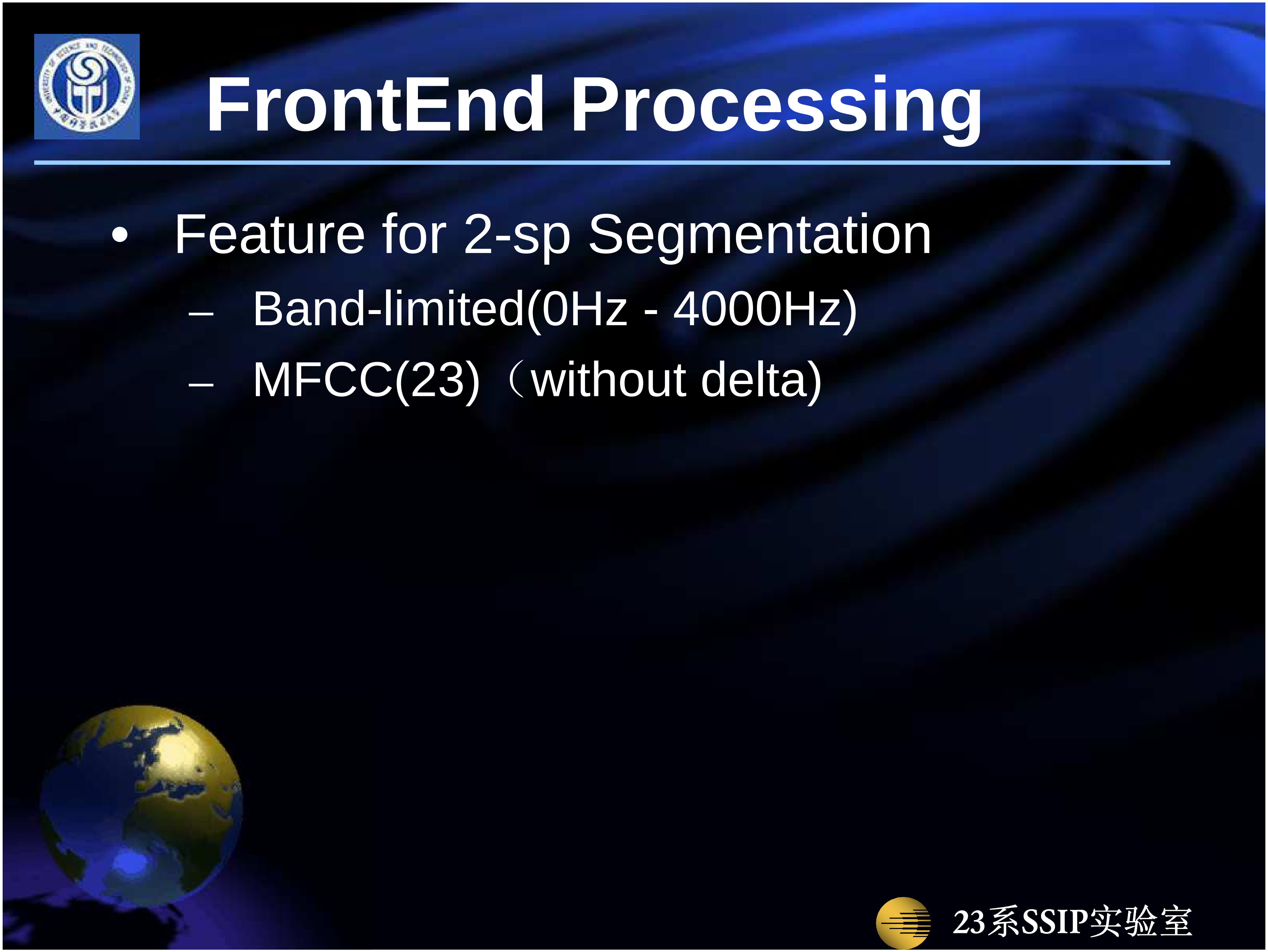
Making Decision

- Threshold is tested with NIST'04 test utterances when the minimal DCF is reached.

USTC 2-sp System

Main Modules

- FrontEnd Processing
- Universal Background Model Training
- Segmentation
- Speaker Model Adaptation
- LLR Score Computation
- Making Decision


FrontEnd Processing
• Feature for 2-sp Segmentation
– Band-limited(0Hz - 4000Hz)
– MFCC(23)（without delta)
23系SSIP实验室

FrontEnd Processing

- Feature for Speaker Verification
  - Band-limited(300Hz - 3400Hz)
  - MFCC + Delta(16 + 16)
  - RASTA
  - CMS
  - Remove Silence
  - Kurtosis Normalization

Universal Background Model

- UBM-F training
- UBM-M training
- Gender Independent UBM training

Gender Dependent UBM training (UBM-F and UBM-M)

- Setting
  - 2048 x 1
- Training Data:
  - NIST'03&04 Dev Training Data (IDs are selected)
- Training Algorithm:
  - EM algorithm

Gender Independent UBM training

- Setting
  - 4096 x 1
- Training Algorithm
  - Merge from UBM-F and UBM-M

Unsupervised Speaker Segmentation

- Hierarchical agglomerative clustering
  - Divide the speech into 1sec segments as initial clusters.
  - Merge two clusters which have minimum pair distance.
  - Until obtain three clusters (speaker 1, speaker 2, overlap of two speakers)

Pair-wise Distance Computing

- Likelihood Ratio Score for Segment

$$L(x:\theta_x)=\prod_{j=1}^{r}\sum_{k=1}^{K}g_k(x)N_k(v_j)$$

- Likelihood Ratio

$$\lambda_L = \frac{L(z:\theta_z)}{L(x:\theta_x)L(y:\theta_y)}$$

Pair-wise Distance Computing

- Transition Probability

$$f(n) \equiv \Pr[S_{i+n} = S_i] = \frac{1+(2p-1)^n}{2}$$

- Duration time bias

$$\lambda_D = \frac{\prod_i^C f(n_i)}{\prod_i^C (1-f(n_i))}$$

Pair-wise Distance Computing

$$d(x, y) = -\log(\lambda_L) - \boldsymbol{\alpha}\log(\lambda_D)$$

α =4

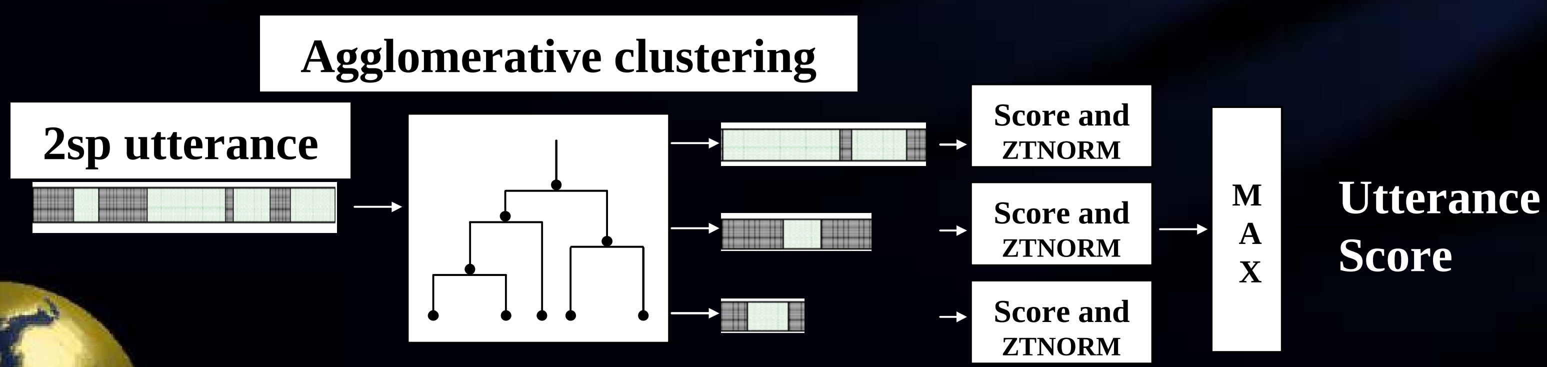
Speaker Model Adaptation

- Setting
  - Same as UBM
- Training data
  - 3 of the 9 Clusters are selected
    - Select most similar 3 clusters from 9 clusters.
- Training algorithm
  - MAP from UBM

LR Score Computation

- Likelihood Ratio Score

$$\Lambda(\mathbf{O}) = \frac{1}{T}\sum_{t=1}^{T}\left(\log p(\mathbf{O}_t \mid \lambda_{tar}) - \log p(\mathbf{O}_t \mid \lambda_{UBM})\right)$$



Making Decision

- Threshold Selecting
  - NIST04 2-spk Evaluation Test Segments
  - Minimal DCF