# The USTC Systems for The NIST-2006 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 LPCC
- FrontEnd Processing for Pitch
- FrontEnd Processing with Wavelet



# FrontEnd Processing for MFCC & LPCC

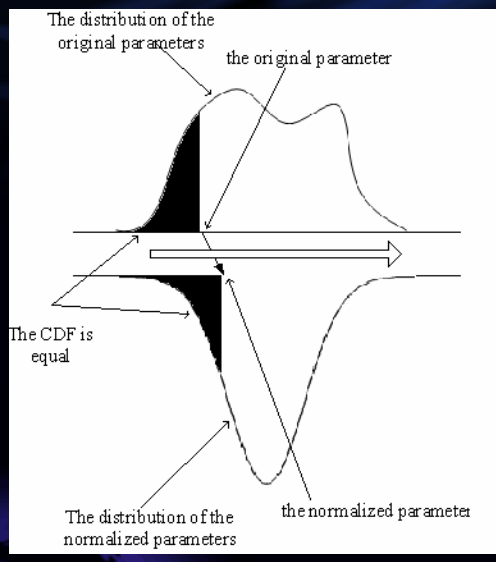
- Band-limited (300Hz – 3400Hz)
- MFCC+Delta(16+16) with the 0th removed
- RASTA
- CMS
- Remove Silence
- Feature Warping

# 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



# Feature Warping or Short-time Gaussianization




$$T(x) = \Phi^{-1}(F_X(x))$$

$$F_X(x) = p_X(X \le x) = \int_{-\infty}^{x} p_X(t)dt$$

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

# FrontEnd Processing for pitch

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

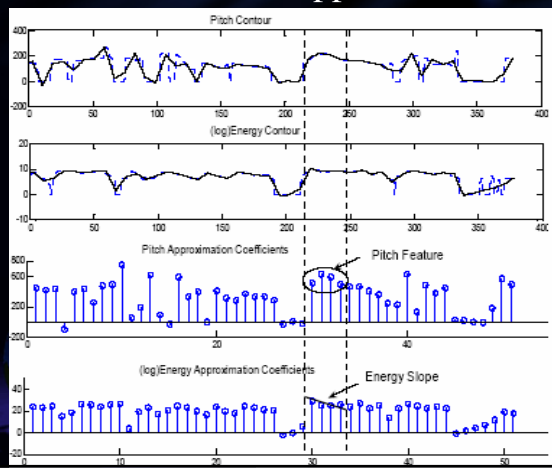
- log (mean_F0) averaged over a segment
- log (max_F0) of a segment
- log (min_F0) of a segment
- F0_slop of a segment

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

23系SSIP实验室

# 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'04&05 training and test data
- Training Algorithm
  - EM Algorithm



# Speaker Model Adaptation

- Model Type
  - Same as UBM
- Training data
  - Training data in NIST'06
- Training algorithm
  - MAP from UBM_M or UBM_F

# LLR Score Computation

- Log Likelihood Ratio

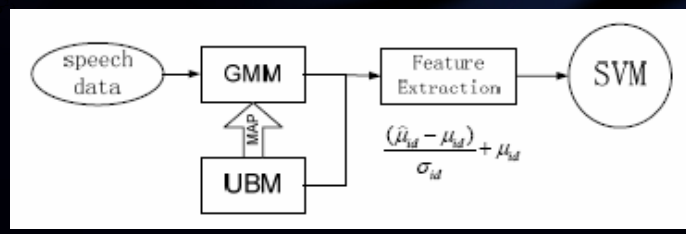
$$\Lambda(\mathbf{O}) = \frac{1}{T}\sum_{t=1}^{T}\left(\log p(\mathbf{O}_t \mid \lambda_{tar}) - \log p(\mathbf{O}_t \mid \lambda_{UBM})\right)$$

- 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.



# SVM system

- Feature: extracted by adapted GMM.
- RBF kernel



# 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.



# Fusion Step 1

- Clustering and training
  - Clustering the models in NIST'05 for each gender
  - Using the Kullback-Leibler distance and hierarchical agglomerative clustering
  - Each gender contain 4 clusters
  - A perceptron classifier is trained for each cluster and the threshold in each cluster is got, respectively

# Fusion Step 2

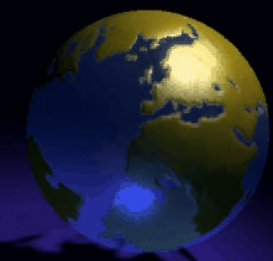
- Classifying and fusion
  - Classify each model in NIST'06 to 1 of the former clusters for each gender
  - Fusion the score of the sub-systems of each 06's model with the corresponding perceptron classifier and threshold.



# Making Decision

- Threshold is tested with NIST'05 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

23系SSIP实验室

# FrontEnd Processing

- Feature for 2-sp Segmentation
  - Band-limited(0Hz - 4000Hz)
  - MFCC(23)（without delta)



# 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'04&05 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 two clusters ( speaker 1, speaker 2)
  - Refine clustering (rescore each 1sec segment by new speaker model and discard some segments with low score)

# 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

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

$\alpha = 4$



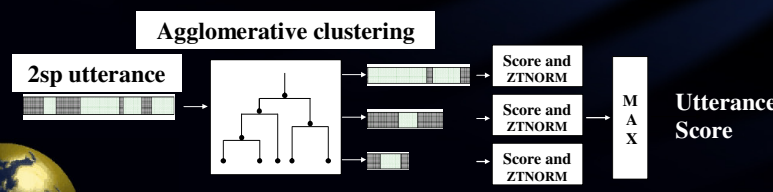
## Speaker Model Adaptation

- 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
  - NIST05 2-spk Evaluation Test Segments
  - Minimal DCF