The PKU 2005 Speaker Recognition System

Gaoxiong Yi, Qiying Li

National Key Lab. on Machine Perception
Center for Information Science
Peking University

Outline

- Introduction
- Description of Submitted system
- Result analysis
- Conclusion

Introduction

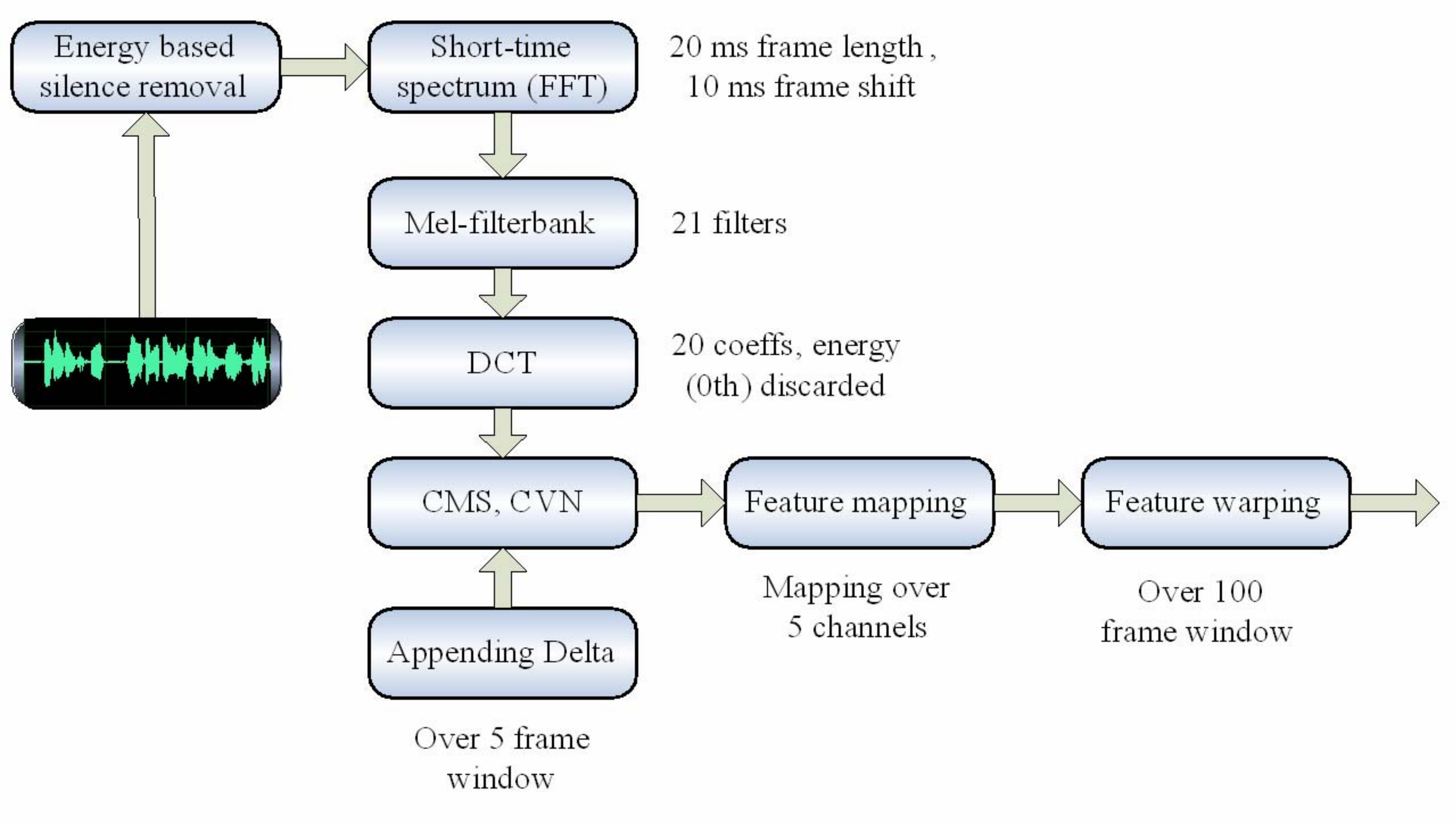
- National key lab. on machine perception
  - -Visual information processing group
  - -Auditory information processing group
    - Auditory computing
    - Spoken language processing
    - Natural language processing
    - Biometrics (voiceprint, face)
  - -Intelligent information processing group

- Participated tasks:
  - 1 conv4w training – 1 conv4w testing
  - 1 conv4w training – 1 convmic testing
  - 8 conv4w training – 1 conv4w testing
  - 10 sec4w training – 10 sec4w testing
  - 3 conv2w training – 1 conv2w testing
- Same acoustic system for all tasks, except
  - Segmentation for 2-sp condition
  - Model size change for 10sec condition

Description of Submitted system

- Overview of our system :
  - -UBM-GMM structure
  - -Feature mapping and Feature warping on MFCC feature
  - -Cohort T-Norm score normalization
  - -Model-score based segmentation for 2-sp condition

Feature Extracting



Feature Mapping

- Channel UBM:
  - Male/Female for Elec, Carbon, GSM,CDMA, Cordless
    - Labeled data from Evaluation data of the past several years
      - Elec and Carbon data is from NIST 97
      - GSM data is from NIST 2001
      - CDMA data is from NIST 2003
      - Cordless data is from NIST 2004
    - Approximately 6 hours of training data for each Channel UBM

The UBM-GMM structure

- Gender-specific UBM trained using pooled data of feature mapping
  - -1024 components diagonal GMMs (change to 64 components in 10sec training condition)
  - -Channel balanced data
  - -Channel UBM were adapted from UBM with the same gender
- Speaker models constructed by mean-only adaptation
  - -Relevance factor is fixed at 16

Scoring

- For each trial, LLR is computed
  - -Only scoring the 3 best components
  - -Cohort T-Norm for score normalization
- We take the normalized LLR as final score
  - -The threshold is set to meet the minimum DCF on 2004 Evaluation

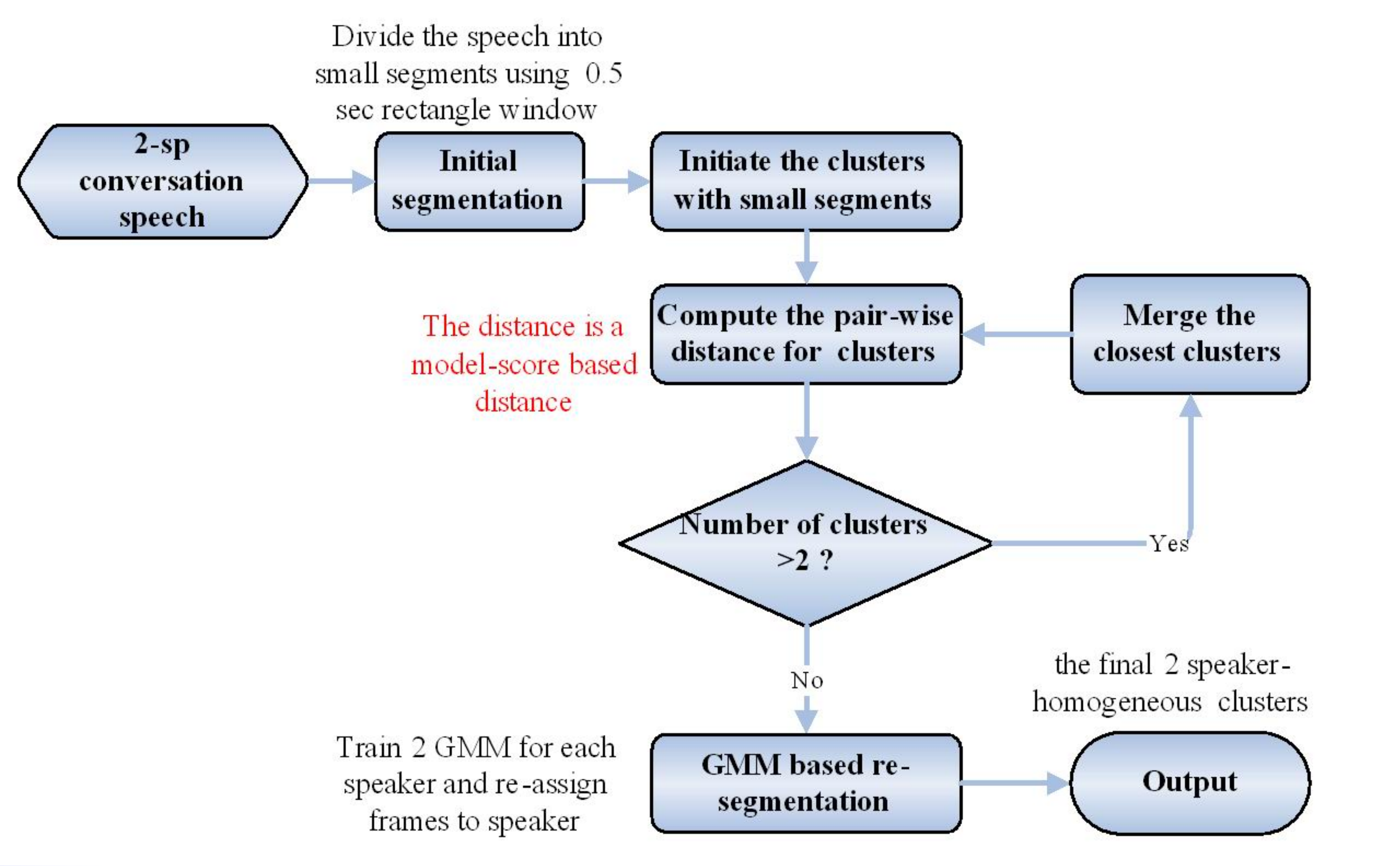
Cohort T-Norm

- Cohort T-Norm for score normalization
  - 750 female/male impostors pool from NIST 97, 2001, 2003 Evaluations
  - For each gender, only keep 250 models as T-Norm pool
    - Chosen models with the bigger average pair-wise distance with all other models, aiming at choosing more representative models
    - The distance between O1 and O2 is computed as:

$$D(O_1, O_2) = \sum_{i=1}^{C} w_i \sum_{j=1}^{F} [(u_{1ij} - u_{2ij})^2 / \sigma_{ij}{}^2]$$

  - 100 speaker-specific T-Norm models for each target speaker, by scoring 200 files to choose the most closest models

Segmentation



Divide the speech into small segments using 0.5 sec rectangle window
2-sp conversation speech
Initial segmentation
Initiate the clusters with small segments
The distance is a model-score based distance
Compute the pair-wise distance for clusters
Merge the closest clusters
Number of clusters >2 ?
Yes
No
the final 2 speaker-homogeneous clusters
Train 2 GMM for each speaker and re-assign frames to speaker
GMM based re-segmentation
Output

The distance measure between clusters

$$d(x,y)=\sum_{m=1}^{N} d(S_{x,m},S_{y,m})=\sum_{m=1}^{N}\frac{(S_{x,m}-S_{y,m})^2}{S_{x,m}{}^2+S_{y,m}{}^2}$$

x,y = cluster segments

$S_{x,m}$ = The LLR of one cluster computed against a GMM/UBM pair in speaker models pool

We assume that speech segments of the same speaker will get similar LLR on models, and the *d(x,y)* will be smaller

-The speaker models pool

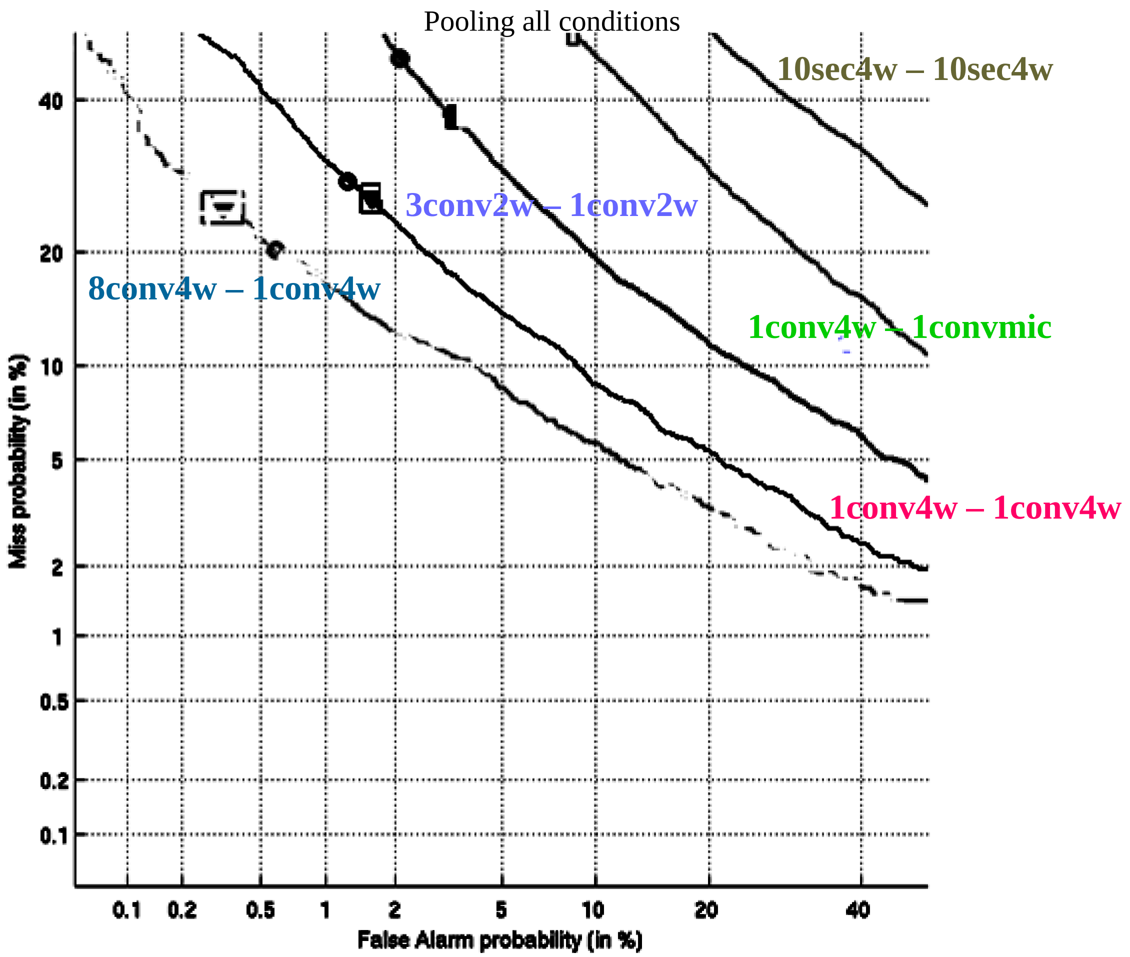
- contains 150 speakers
- chosen from Switchboard II Part1 corpus
- adapted from a 2048 component gender-independent UBM

A roughly segmentation result

-Using 1.5 hours of 2-sp conversation speech in NIST 2003 SRE data

Speaker gender condition	Average speech purity in output clusters
Female & Female conversation	83.6%
Female & Male conversation	95.3%
Male & Male conversation	94.7%

Result analysis



- Serious system performance degradation from 1conv4w-1conv4w to 1conv4w-1convmic
  - Performance loss due to serious channel mismatch, no special processing for microphone data
  - If a Mic channel was added when performing feature mapping, will there be improvement? how much?
- For the 10sec4w-10sec4w condition
  - The same set of T-Norm models was used as in 1 conv4w training condition.
  - There was difference between the amount of training data for target model and T-Norm model. How did it effect the performance?

Conclusion

- More work need to be done on mitigating channel effect
- More effective modeling need to be considered in condition of less training data
- Combing other information sources need to be considered. Sites which made use of high-level features have shown significant system improvement

Thank you!