Speaker Recognition Research by the Recent Graduates in CSLT

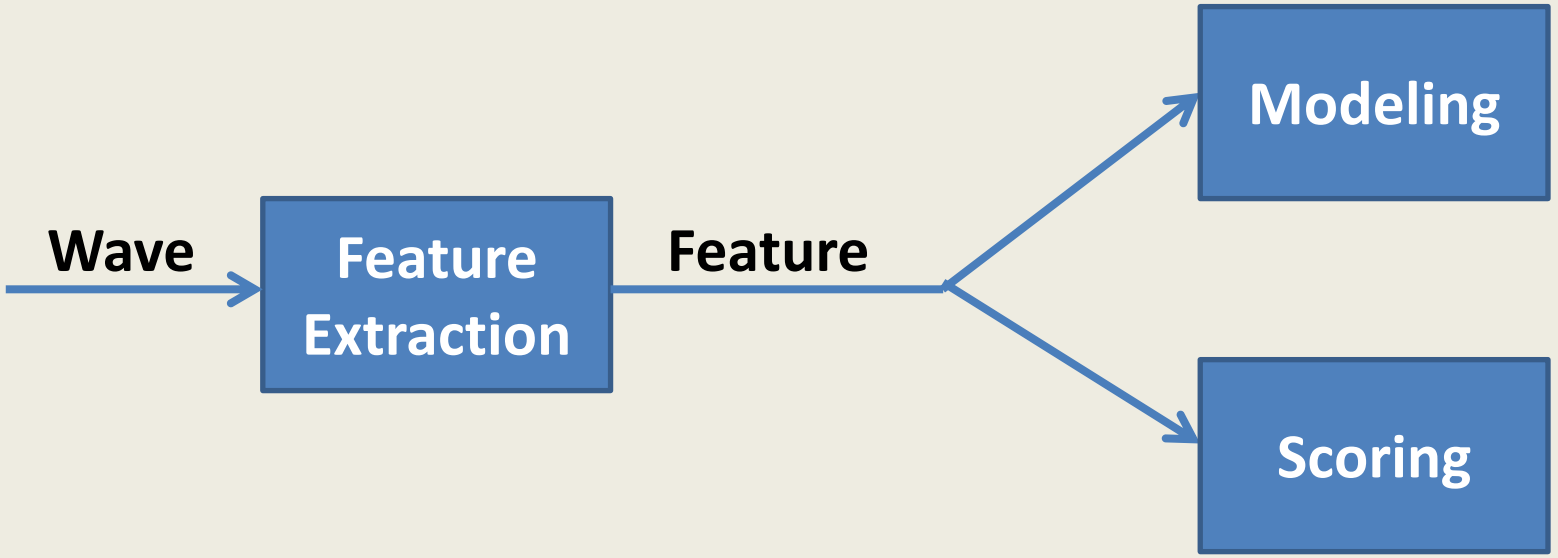
WU, Xiaojun

2012-11

Persons

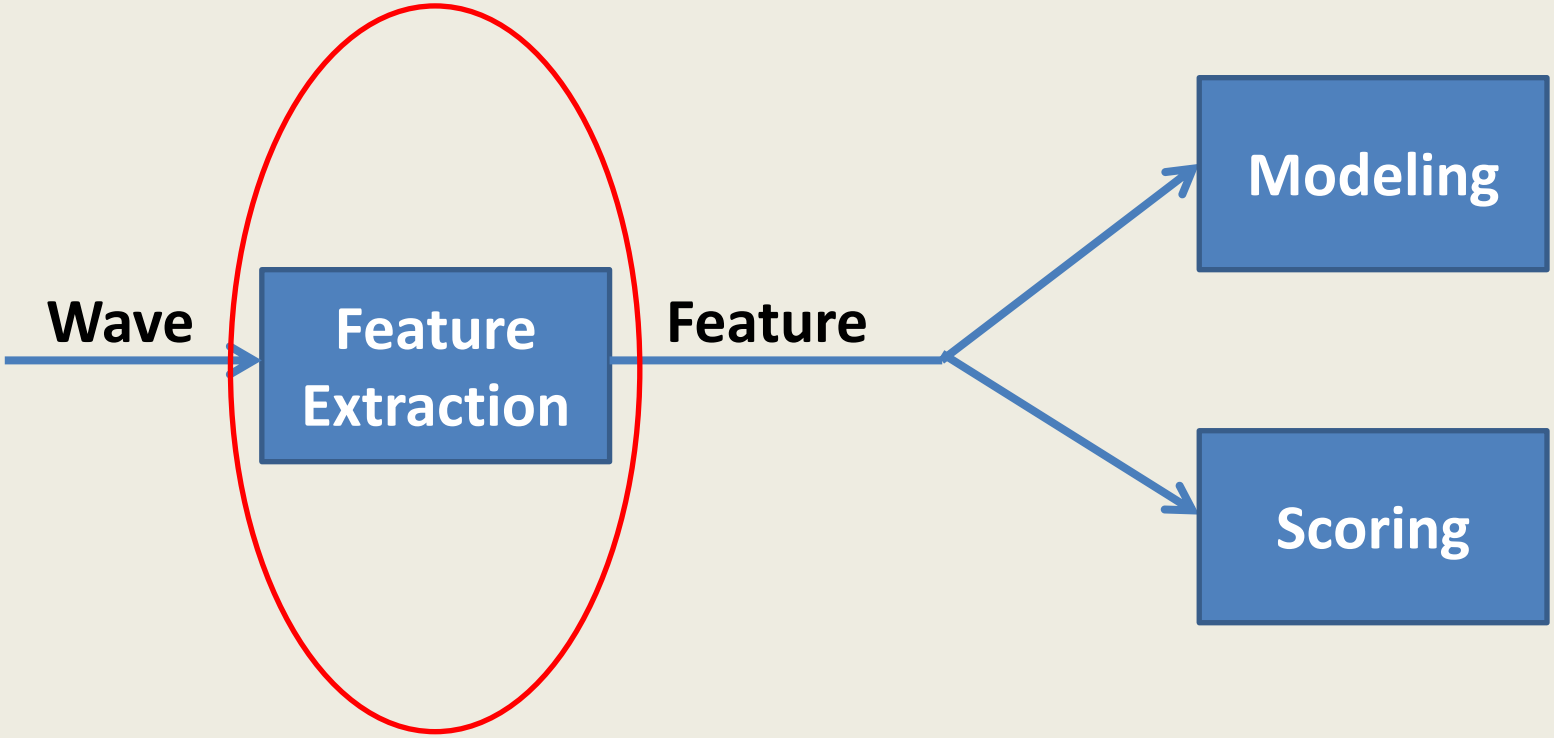
- Masters
  - WU, Wei in 2007
  - BAO, Huanjun in 2007 (SVM)
  - LUO, Canhu in 2011 (embedded, text dependent)
- PhDs
  - XIONG, Zhenyu in 2005
  - DENG, Jing in 2006
  - WANG, Gang in 2011

Speaker Recognition Steps



Motivations and Approaches

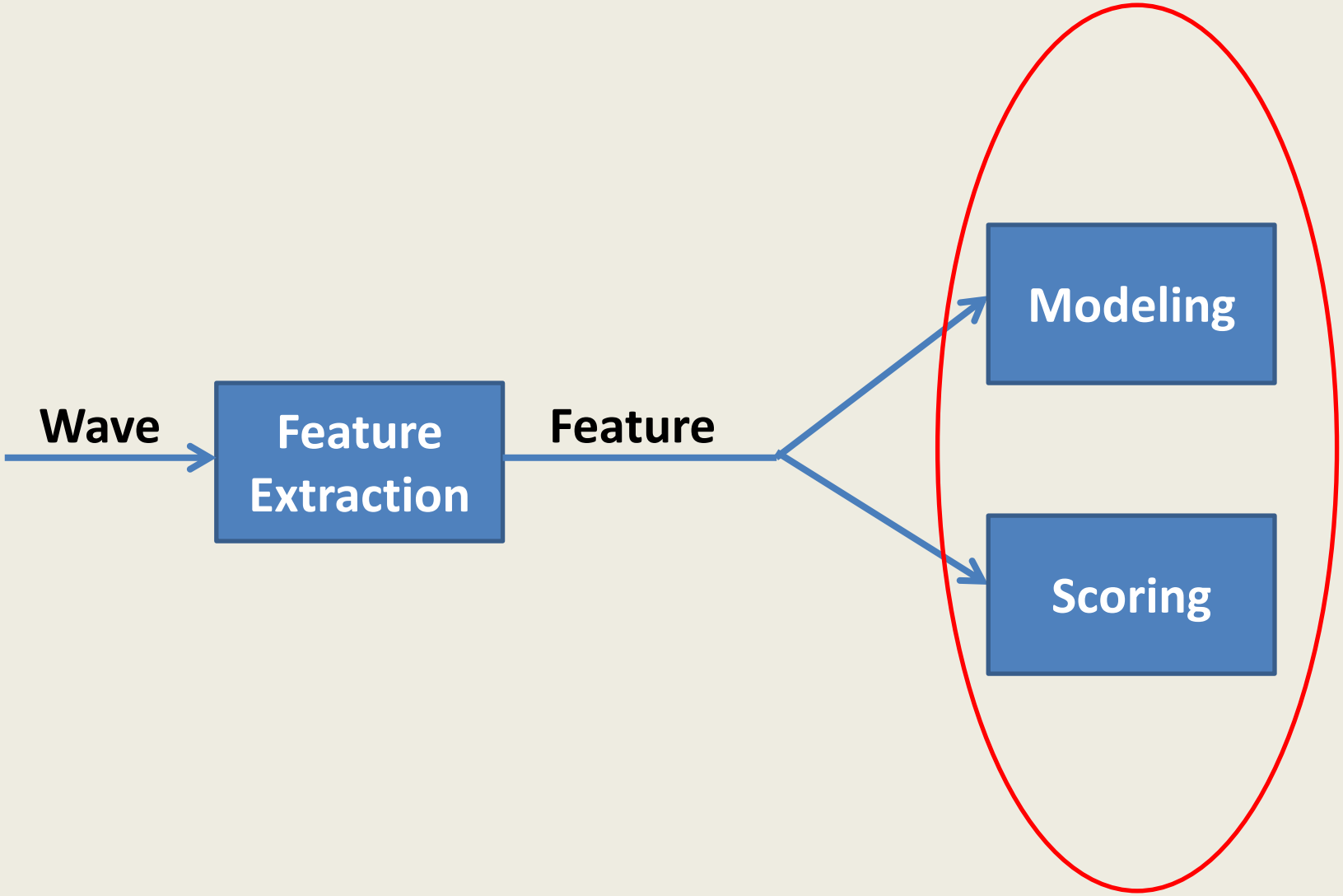
- Motivations
  - Better recognition results
  - Robustness of noise and cross-channel
  - Speeding-up
- Approaches
  - New processing methods
  - Additional processes
  - Data structures



Noise Robust Feature Extraction

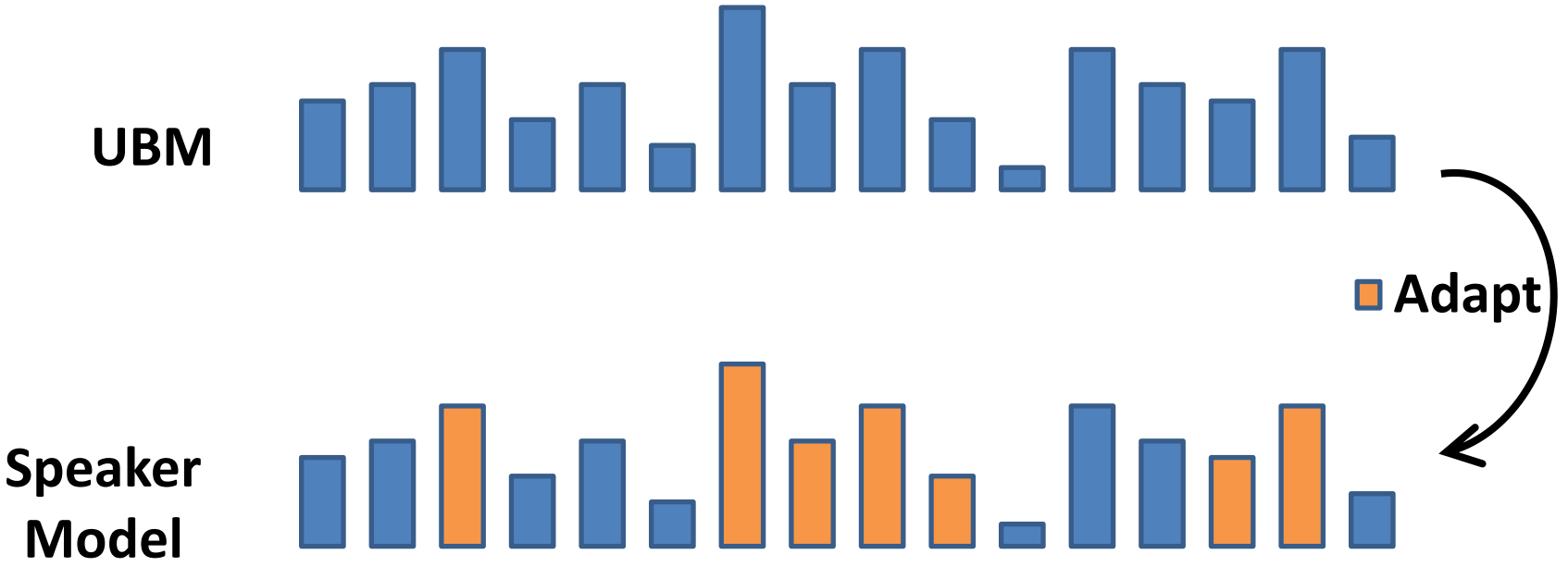
DENG's Idea

- Wave peaks and troughs contain speaker information more than other parts
- Noise changes troughs greatly
- Use a sine filter to help detect peaks and troughs with differential power estimation
- Estimate the clean speech power spectrum by accumulating the differential power values


Wave
Feature Extraction
Feature
Modeling
Scoring

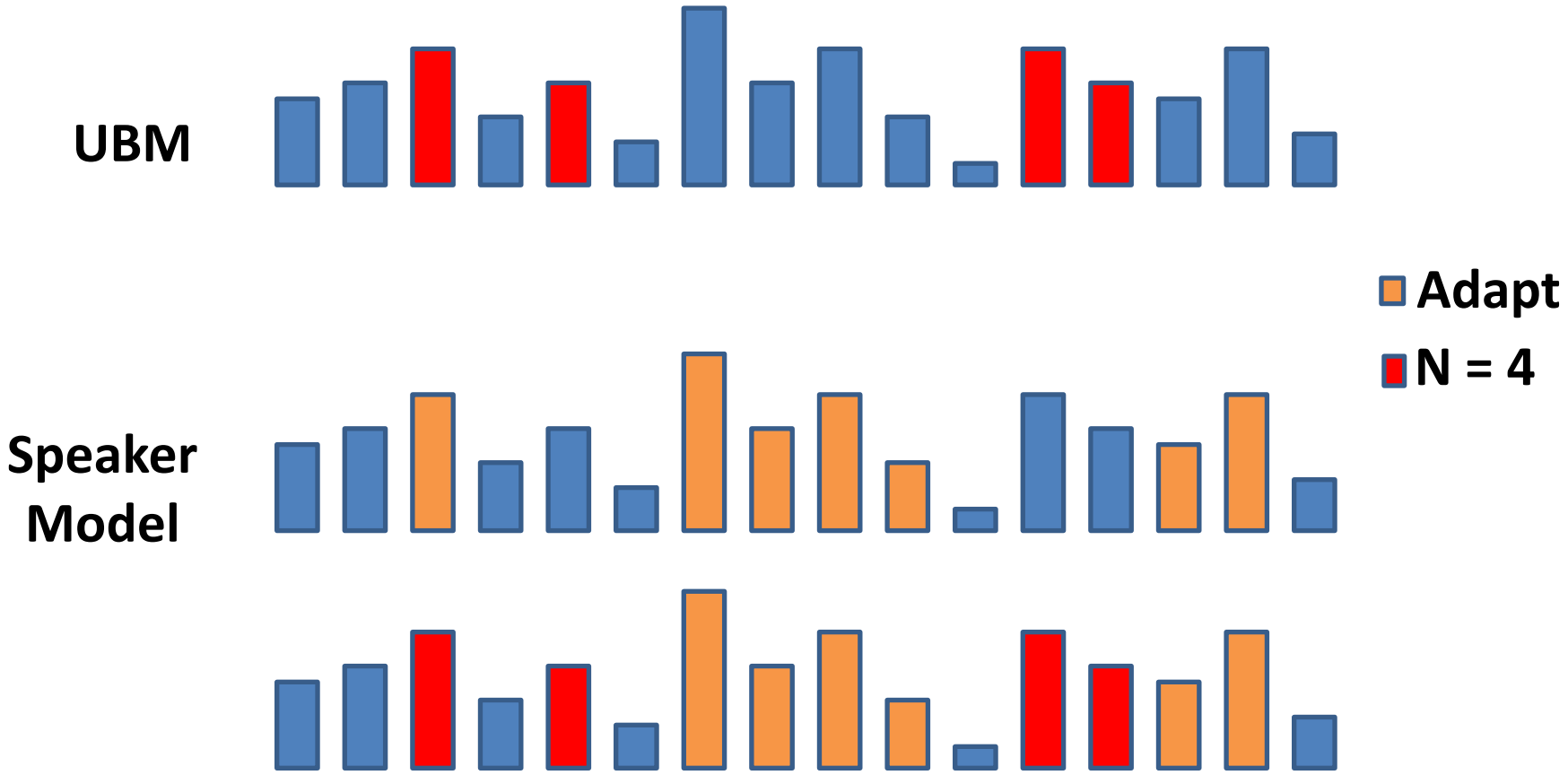
GMM-UBM Modeling

- Gaussian Mixture Model
  - A weighted group of multi-dimensional Gaussian distributions to represent any distribution of vectors
- Universal Background Model
  - Trained from feature vectors of a large amount of speakers, the UBM represents an average distribution of features in the GMM format
- Speaker models
  - Adapted from the UBM with the speaker's own feature vectors, also in the GMM format


UBM
Adapt
Speaker
Model

GMM-UBM Scoring

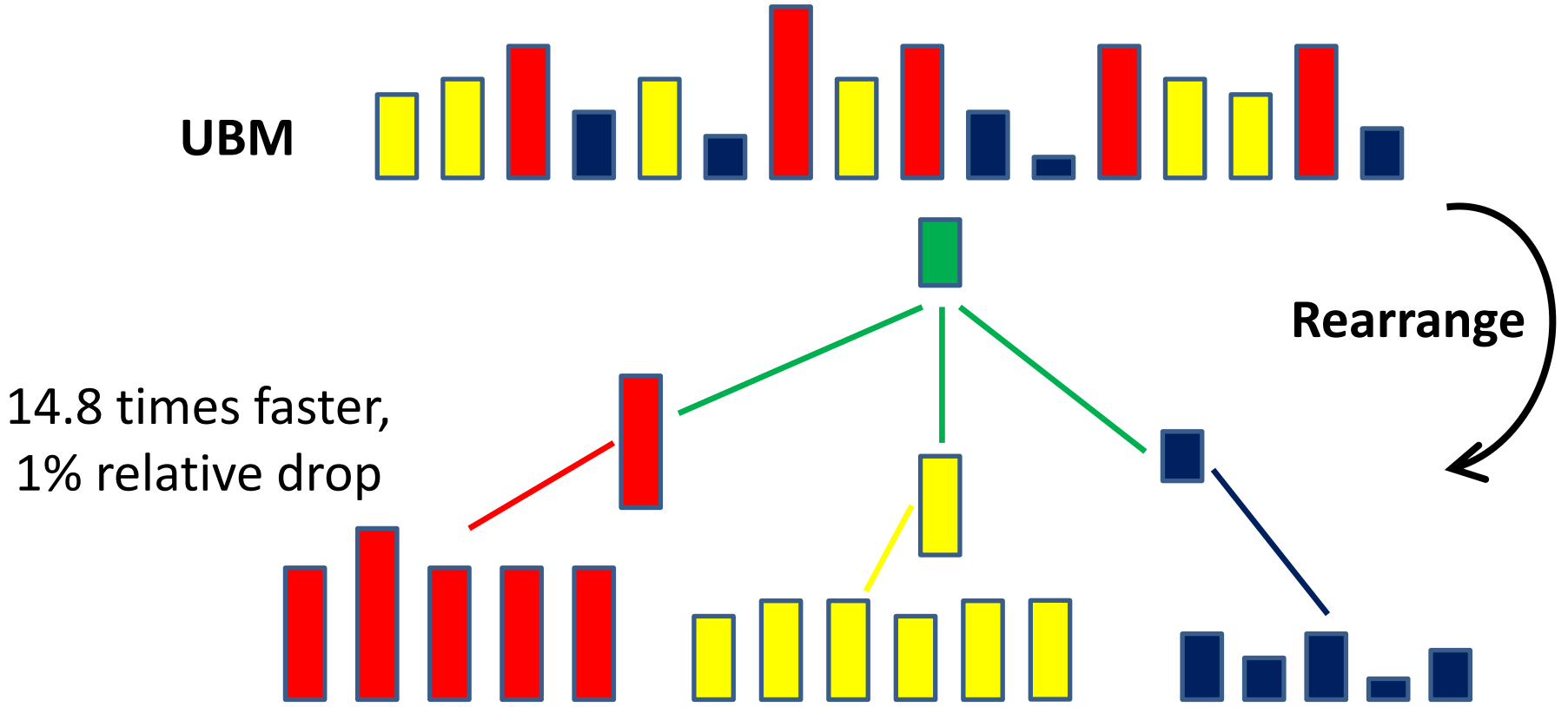
- Recognition score = average of all frame scores
- Frame scoring (to score the feature of the frame)
  - Find top N Gaussian components in the UBM with the highest probabilities (N equals 4 or 5 commonly)
  - score against the UBM = the product of the N probabilities
  - score against the speaker model = the product of the N probabilities of the corresponding N Gaussian components in the speaker model
  - Frame score = log (score against the speaker model) - log (score against the UBM)


UBM
Speaker Model
Adapt
N = 4

New Data Structure to Speed Up

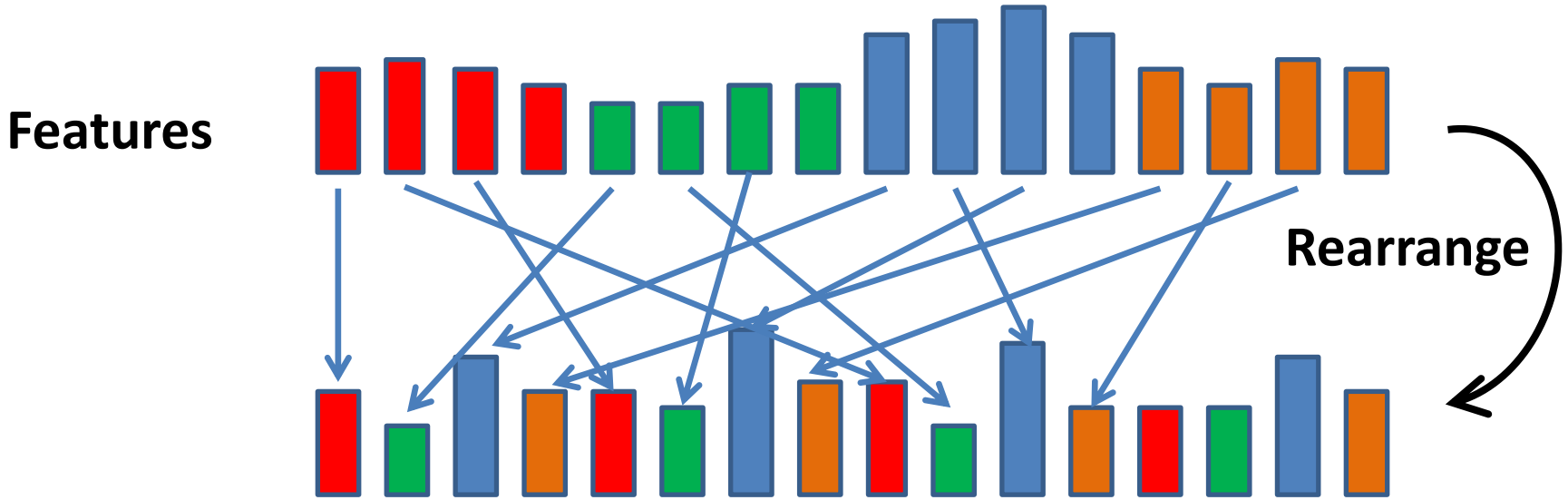
XIONG's Idea (1)

- Rearrange the Gaussian components in the UBM into a tree



XIONG's Idea (2)

- Rearrange the sequence of feature vectors so as to prune as early as possible

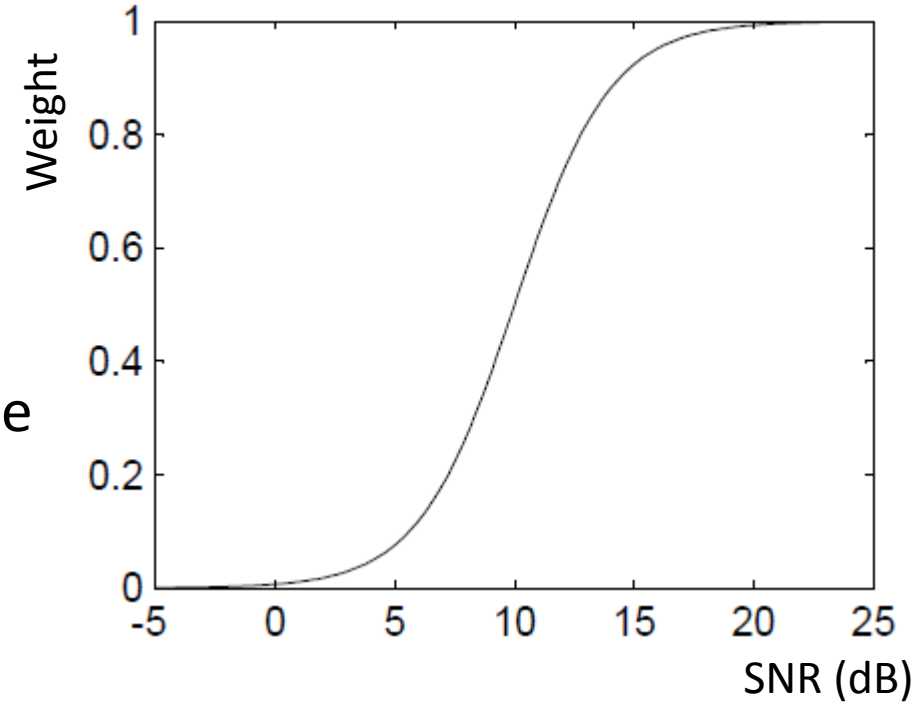


25 times faster,
no accuracy drop

Noise Robust Scoring

XIONG's Idea

- Clean frames are more reliable than noisy frames



Error rate drops 8.7% in average and 13.4% in [0dB, 10dB] than spectral subtraction.

New Ideas in Scoring

XIONG's Idea (for Open-set SI)

- Every part of feature vectors should be with high confidence measure for the true speaker
- Use segmental confidence measures to train an ANN as a classifier
- EER is 30% relative lower than log-likelihood

Cross-channel Robust Modeling

DENG's Idea

- For any model *M* in channel *c*

  $M(c) = m + Uz(c)$

- Where, *m*: the channel free model

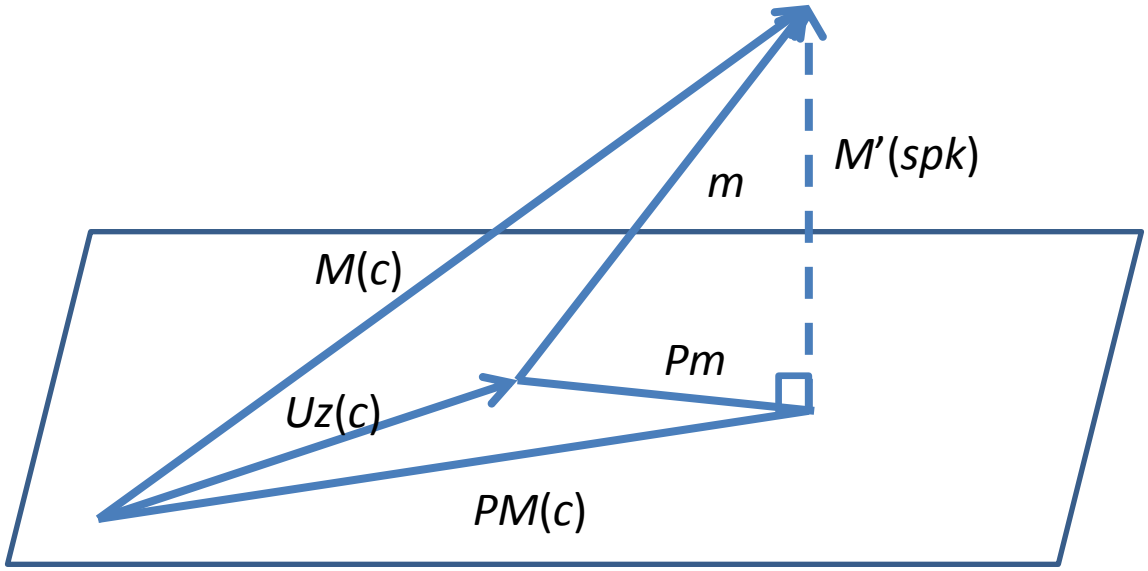
  U : basis matrix for channel space

  Uz(*c*): channel projection

- By training to get *U*, we find the projection matrix

  $P = UU^t$, and $PU = UU^tU = U$

DENG's Idea – cont.

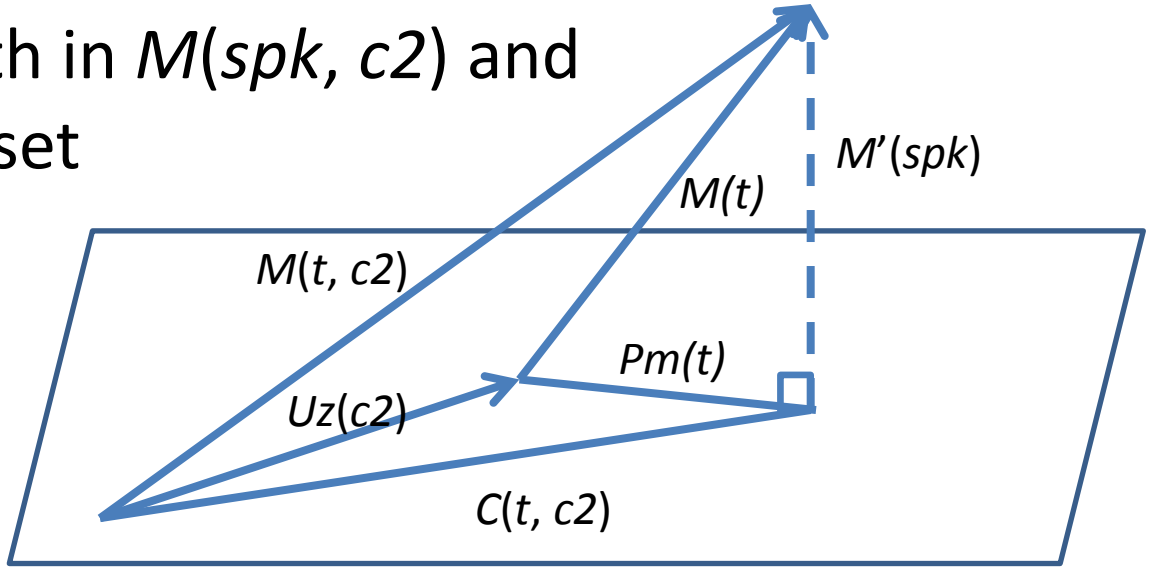


- Define: $C(spk, c) = PM(spk, c) = Pm(spk) + Uz(c)$
- Store: $M'(spk) = (I\text{-}P)M(spk, c1) = (I\text{-}P)m(spk)$

  $UBM' = (I\text{-}P)UBM(c1)$

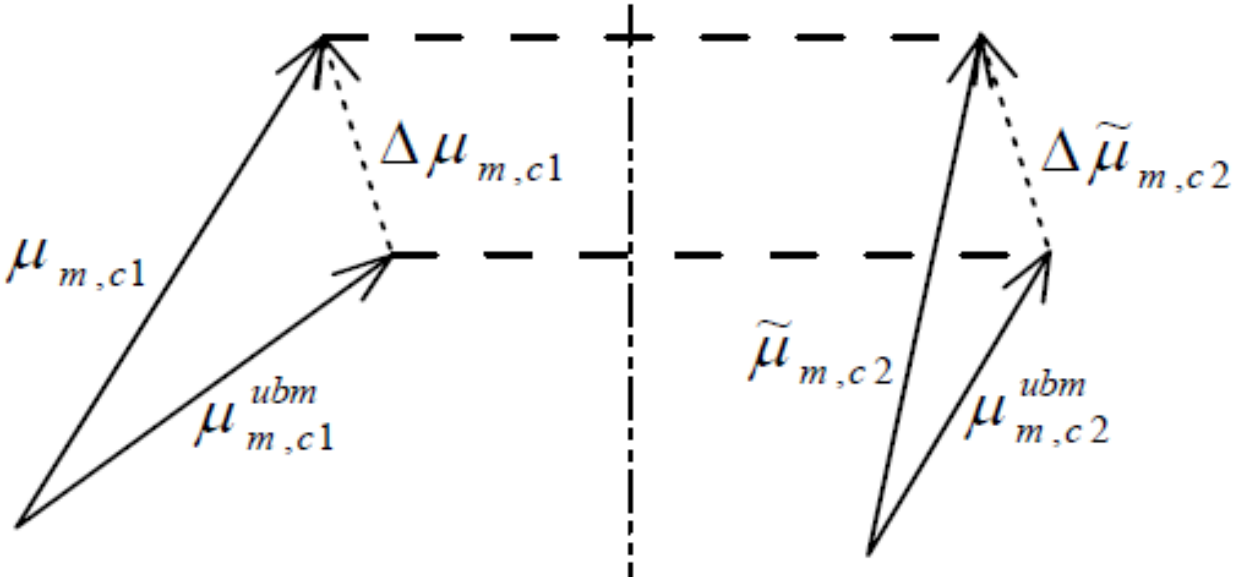
DENG's Idea – cont.

- When scoring utterance *t* in channel *c2*
  - Train *M*(*t*, *c2*) based on *UBM*(*c1*) to get
    $$C(t, c2) = PM(t, c2) = Pm(t) + Uz(c2)$$
  - Use *C*(*t*, *c2*) + *M'(spk)* to estimate *M*(*spk*, *c2*), and *C*(*t*, *c2*) + *UBM'* to estimate *UBM*(*c2*)
  - Same *Pm(t)* both in *M*(*spk*, *c2*) and *UBM*(*c2*) → offset



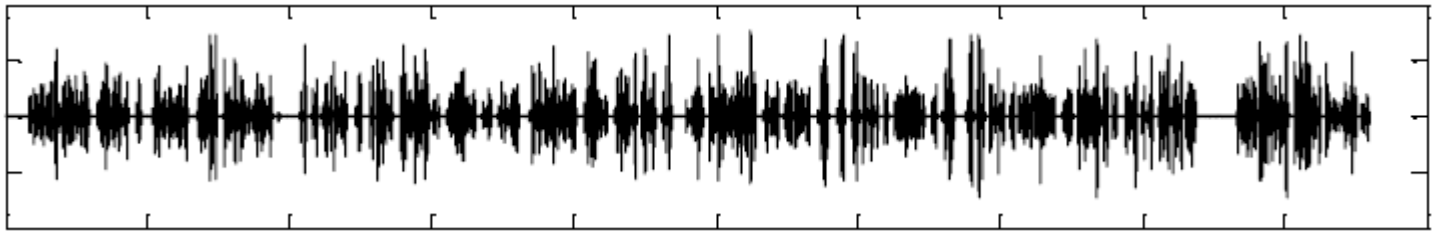
WU's Idea

- UBM-based speaker model synthesis assumes ⊿UBM = ⊿speaker across channels
- Really?
- ⊿Cohort ≈ ⊿speaker instead



Application: Speaker Detection

Common Steps



Wave

Speaker segmentation

Speaker clustering

Speaker Identification

Speaker X **Speaker Y** **Speaker Z**

Speaker Segmentation (1)

- Distance measure of a speech segment to a model is the key technology
  - To find long distance between adjacent segments
  - To find big difference between distances of adjacent speech segments to same models

- WANG used phoneme based text dependent recognition scores of the adjacent segments

Speaker Segmentation (2)

- DENG used the log likelihood ratio score of the adjacent segments to the UBM
- WANG proposed Reference Speaker Models
  - Speech segments from the same speaker have similar distances to other models
  - Cluster known speakers into a set of Reference Speaker Models
  - RSMs can also used to measure the distance between two speech segments

Speaker Clustering

- DENG used a small UBM for rough clustering and then a large UBM for fine clustering
- WANG used
  - RSMs to measure distances between segments
  - Within-class dispersion to keep high class purity

Speaker Identification

- WANG proposed
  - Pre-calculate the distances of every speaker's speech to RSMs
  - Calculate the distances of the test speech to RSMs
  - Find speakers with similar distances for further identification
  - Speed up: cluster speakers to save calculation of distance similarity

Reference

Can be found on Prof. Zheng’s homepage, including papers and thesises.