NIST 2005
Speaker Recognition Evaluation

University of New Brunswick
System Submission

Kevin Englehart, Ryan Reynolds,
Kim Briggs, Arvind Kizhanatham, Eduardo Castillo-Guerra, Erik Scheme

UNB

Outline

- Background
- Detection systems
  - Core system description
  - Development data
  - Pitch / energy system
  - Fusion
- Performance
- Current research and future directions

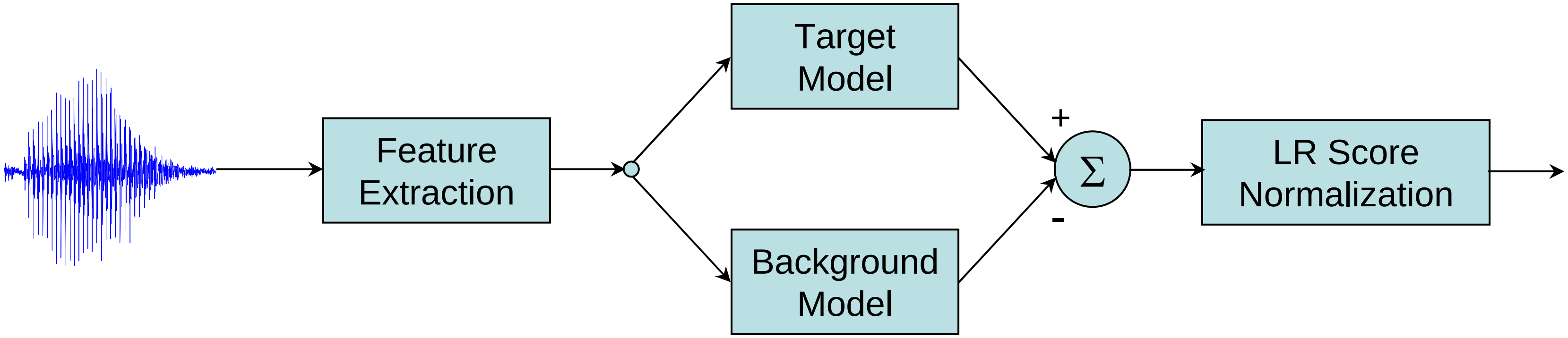
Background

- UNB R&D efforts began in October 2004
- Objectives: robust speaker identification and verification, with emphasis on short utterance length
- Team decided to participate in NIST SRE05
  - technical challenges are very close
  - motivated to become part of research community

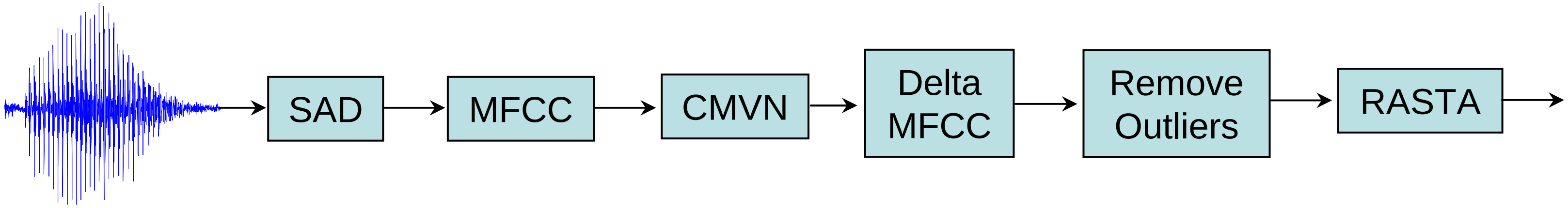
Timeline

- Core team assembled in January 2005
- MFCC-GMM system prototyped in February 2005
- Development data
  - Jan-March 2005: NIST SRE 1998 (landline only)
  - April 2005: NIST SRE 2004 (landline & cellular)

Scoring: Log-Likelihood Ratio



Front-End Processing



SAD → MFCC → CMVN → Delta MFCC → Remove Outliers → RASTA →

- Speech activity detection
  - Energy and frequency (zero crossings)

- Feature extraction
  - 20 ms window, 10 ms increment
  - 19 cepstral coefficients extracted from 300-3300Hz band
  - 19 delta cepstral coefficients
  - MFCC outliers removed (bottom, top 10%)

- Channel compensation
  - RASTA filtering
  - Cepstral mean and variance normalization (speech frames only)
  - Attempted Feature Mapping*

* Reynolds, D.A., “Channel robust speaker verification via feature mapping,” ICASSP 2003

Target and Background Models

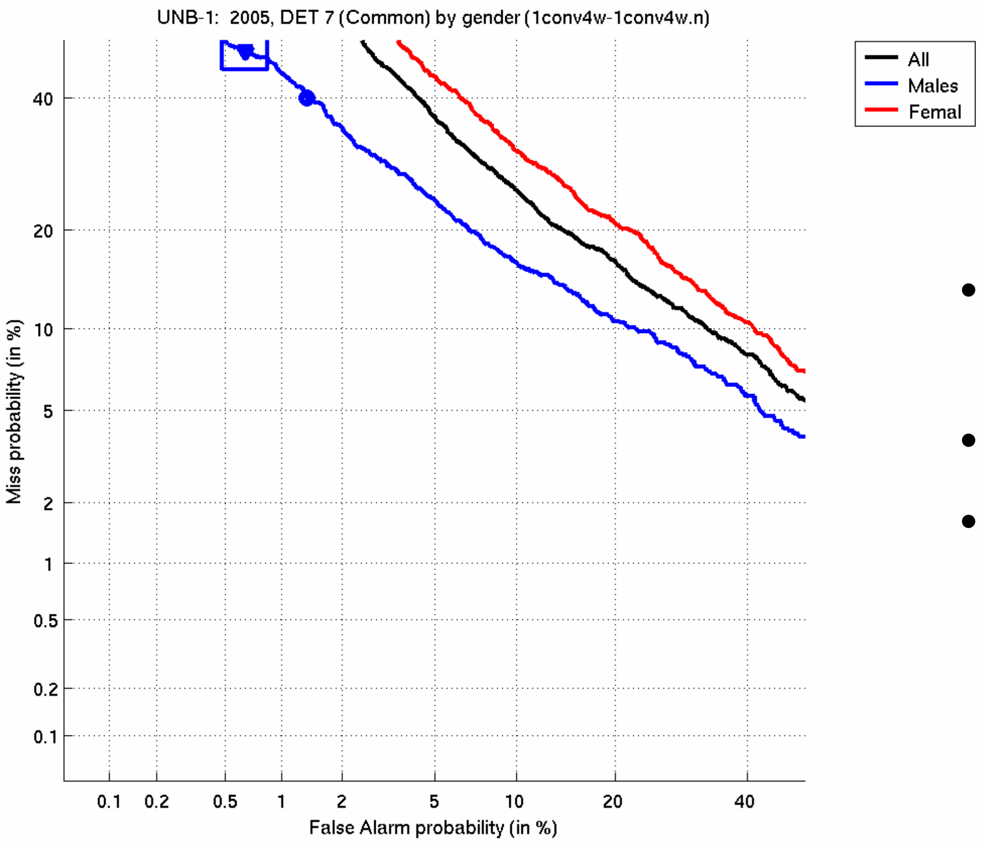
- Background models are 64 mixture GMMs
  - Trained using NIST 2004 data
  - Approximately 4 hours of data
    - 248 speakers, one minute per speaker (to avoid speaker bias)
  - Unbalanced (8:1 landline vs. cellular)
    - Performance degraded with more UBM mixtures
- Target models derived from UBM using MAP adaptation
  - One iteration
  - Variance limiting
  - Only mean vectors adapted

Prosodic Features

- Pitch + energy (1 minute of data)
- GMM classifier (64 mixtures)
- Fusion using a 2-layer multilayer perceptron network with 12 hidden layer nodes
- The training performance measure was adapted to minimize the DCF

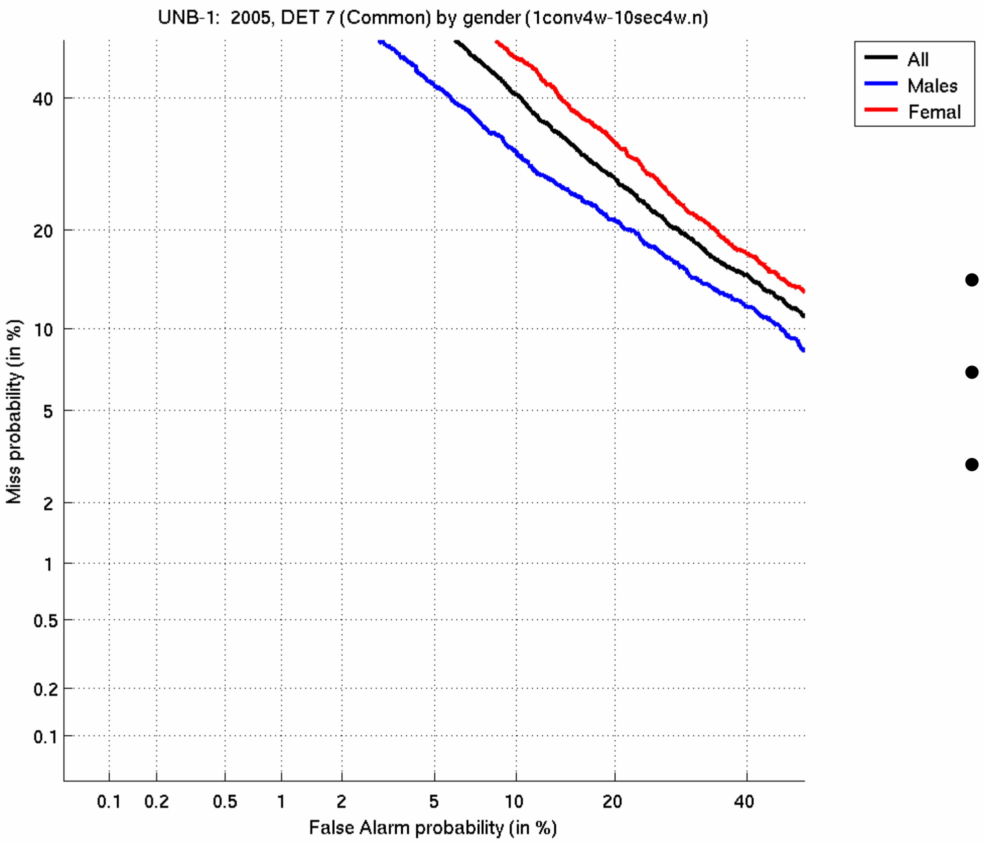
NIST SRE'05 Results

1conv4w-1conv4w



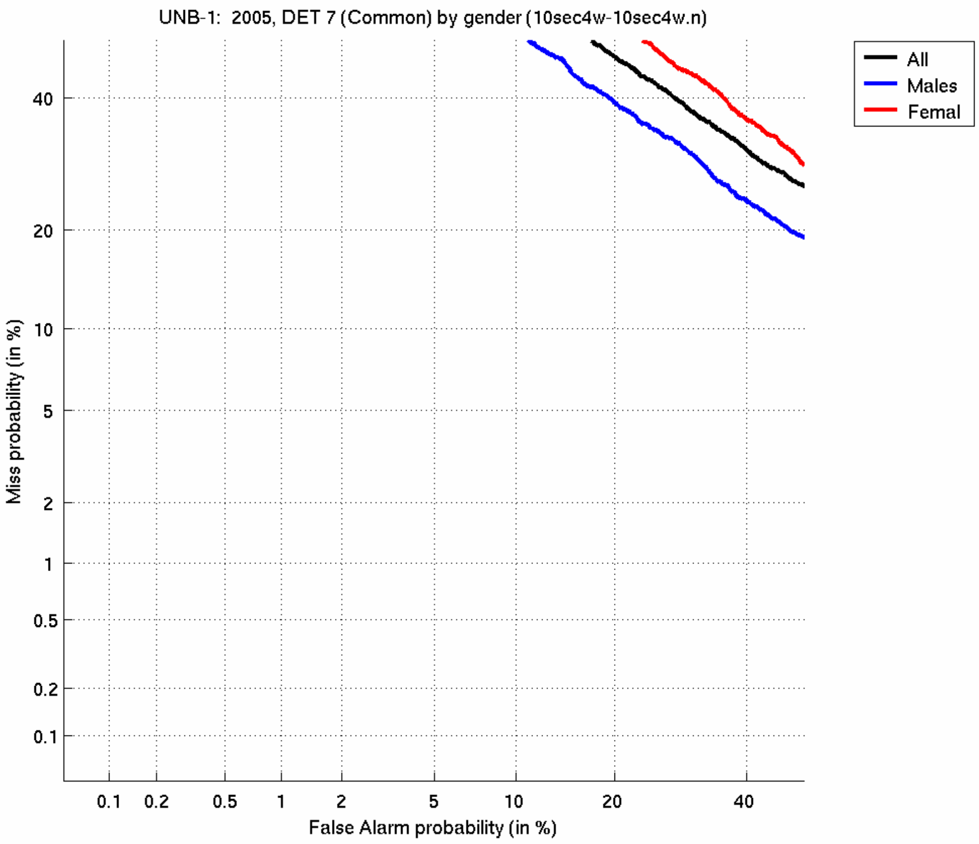
- 2 minutes for training and 1 minute for verification
- No score normalization
- No handset compensation

1conv4w-10sec4w



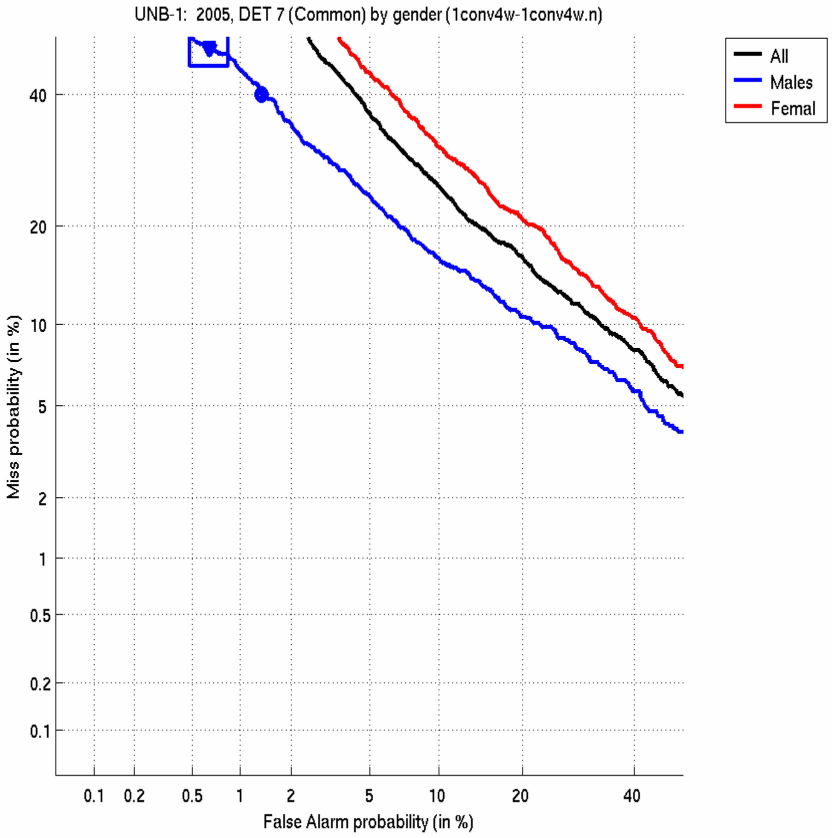
- 2 minutes for training
- No score normalization
- No handset compensation

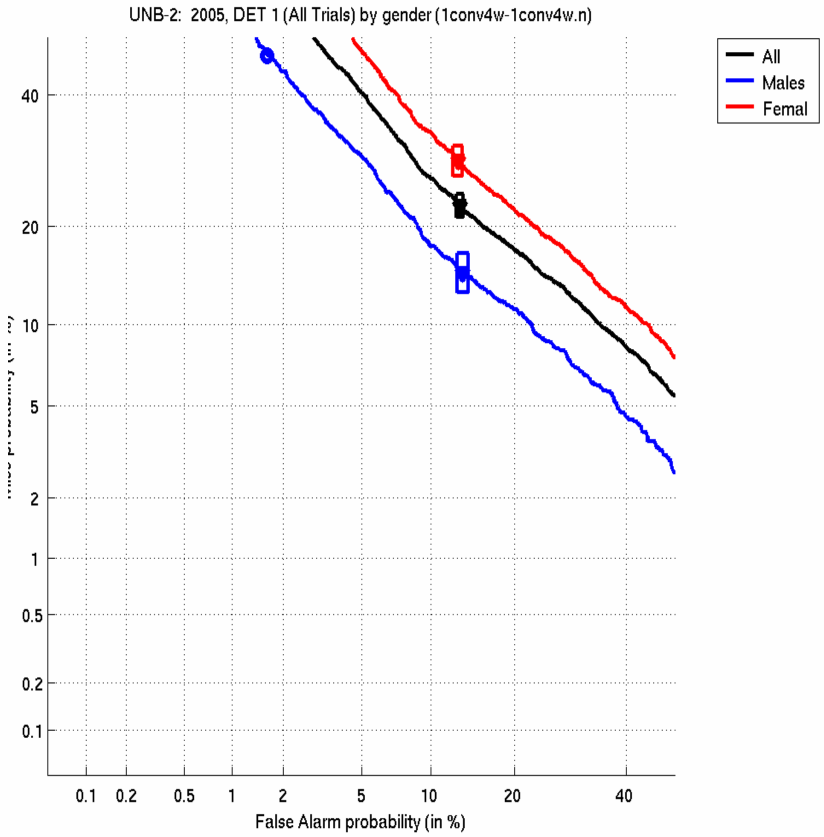
10sec4w-10sec4w



- No score normalization
- No handset compensation

Fusion


UNB-1: 2005, DET 7 (Common) by gender (1conv4w-1conv4w.n)
All
Males
Femal
Miss probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1
0.2
0.5
1
2
5
10
20
40
False Alarm probability (in %)


UNB-2: 2005, DET 1 (All Trials) by gender (1conv4w-1conv4w.n)
All
Males
Femal
40
20
10
5
2
1
0.5
0.2
0.1
0.1
0.2
0.5
1
2
5
10
20
40
False Alarm probability (in %)

Current Areas of Research

- Score normalization
  - T-norm with cohort selection
- Feature mapping
- Handset and phone type normalization
  - Cell (analog & digital)
  - Landline (regular & cordless)

Future Directions

- Exploit more development data
  - Balanced UBM creation
  - Handset dependent UBMs
- Speaker dependent threshold selection