AFRL/HECP 2005
Speaker Recognition Systems

Raymond E. Slyh
Human Effectiveness Directorate
Air Force Research Laboratory

Team Members

- **AFRL/HECP:**
  - **Eric Hansen**
  - **Raymond Slyh**
- **General Dynamics Advanced Information Systems**
  - **Brian Ore**
  - **Anthony Halley**

Components of Submitted Systems

Yellow = switch train/test		TESTING 10sec4w	TESTING 1conv4w	TESTING 1conv2w
TRAINING	10sec4w	FMBWF0 LPCC MFCC	FMBWF0 LPCC MFCC	MFCC
TRAINING	1conv4w	FMBWF0 LPCC MFCC	FMBWF0 LPCC MFCC PS-MFCC WLM	MFCC
TRAINING	3conv4w	FMBWF0 LPCC MFCC	FMBWF0 LPCC MFCC PS-MFCC WLM	MFCC
TRAINING	8conv4w	FMBWF0 LPCC MFCC	FMBWF0 LPCC MFCC PS-MFCC WLM	MFCC
TRAINING	3conv2w	MFCC	MFCC	MFCC

KEY

FMBWF0: F1–F3, BW1–BW3, F0

LPCC: 16 Coeffs + Deltas (from Closed-Phase Analysis)

MFCC: 19 Coeffs + Deltas

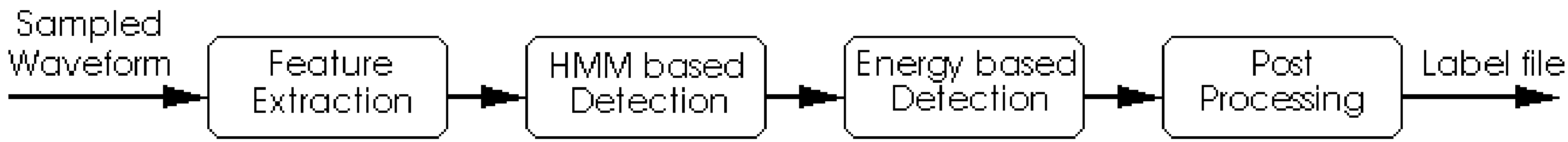
PS-MFCC: MFCCs Using Phoneme-Specific GMMs

WLM: Language Modeling on BBN Words

GMM-Based Systems

- Version 2.1 of MIT Lincoln Laboratory system:
  – Gaussian mixture models (GMMs)
  – Diagonal covariance matrices
- Background, target, & T-norm models: 2048 mixtures
- Model adaptation from background:
  – FMBWF0: Weights, means & variances adapted
  – LPCC, MFCC, & PS-MFCC: Only means adapted

MFCC/HMM++ SAD (1)



- **Features: 19 MFCCs (300–3138 Hz) & deltas (No RASTA or feat map)**
- **HMM-based speech activity detector (SAD):**
  - **Two-state HMM built with HTK (64 mixtures/state)**
  - **Trained on background model data using SONIC labels as truth**
- **Energy-based detector:**
  - **Refines the output from the HMM-based detector**
  - **Noise floor set using the average frame energy from the top ten non-speech segments from the HMM-based detector**
  - **Energy-based detection performed using MIT-LL *xtalkN***
- **Post-Processing: Removes speech segments < 20 msec in duration**
- **Only used for PS-MFCC system if SONIC SAD gave no speech frames**

GMM Systems: Background Model

- Approx. 16 hours of data
- Gender-balanced
- Channel-balanced
- Sources:
  - NIST 2001–2003 evaluations (for carbon button, electret, and digital cellular channels)
  - OGI National Cellular Corpus (for analog cellular)
- Gender/channel models used for feature mapping

GMM Systems: T-norm Models

- In general (other than 10sec4w training):
  - Gender-dependent
  - 120 models for each gender
  - Data for each model:
    - From NIST 2001–2003 evaluations
    - Single conversation side
- For 10sec4w training conditions:
  - Gender-independent
  - 240 models
  - 10sec4w and 1conv2w testing: Built from the first 30 sec of data from original set of T-norm models

FMBWF0 & LPCC Systems

- **FMBWF0:**
  - **F1–F3 in radians, BW1–BW3 in radians, and log(F0)**
  - **F0 & probability of voicing from ESPS *get_f0***
  - **Formant center frequencies & bandwidths from Snack 2.2.2 from KTH**
- **LPCC:**
  - **LP params from closed-phase analysis (Odyssey 2004)**
  - **16 cepstral coefficients (no 0^{th}) with RASTA & deltas**
  - **Feature mapping (using channel from MFCCs) and mean and variance normalization**

MFCC & PS-MFCC Systems

- **MFCC:**
  - From Version 2.1 of MIT-LL MFCC/GMM system
  - 19 mel-frequency cepstral coefficients (BW: 300–3138 Hz, no 0th coeff.) with RASTA & deltas
  - Feature mapping and mean and variance normalization
- **PS-MFCC:**
  - Features as in MFCC system
  - Used SONIC SAD generally
  - "Top 15" phonemes from SONIC (Ver. 2.0-beta2) run as an English-language speech recognizer:

{AE, AH, AX, AY, DH, EH, EY, IH, IY, L, M, N, OW, S, Y}

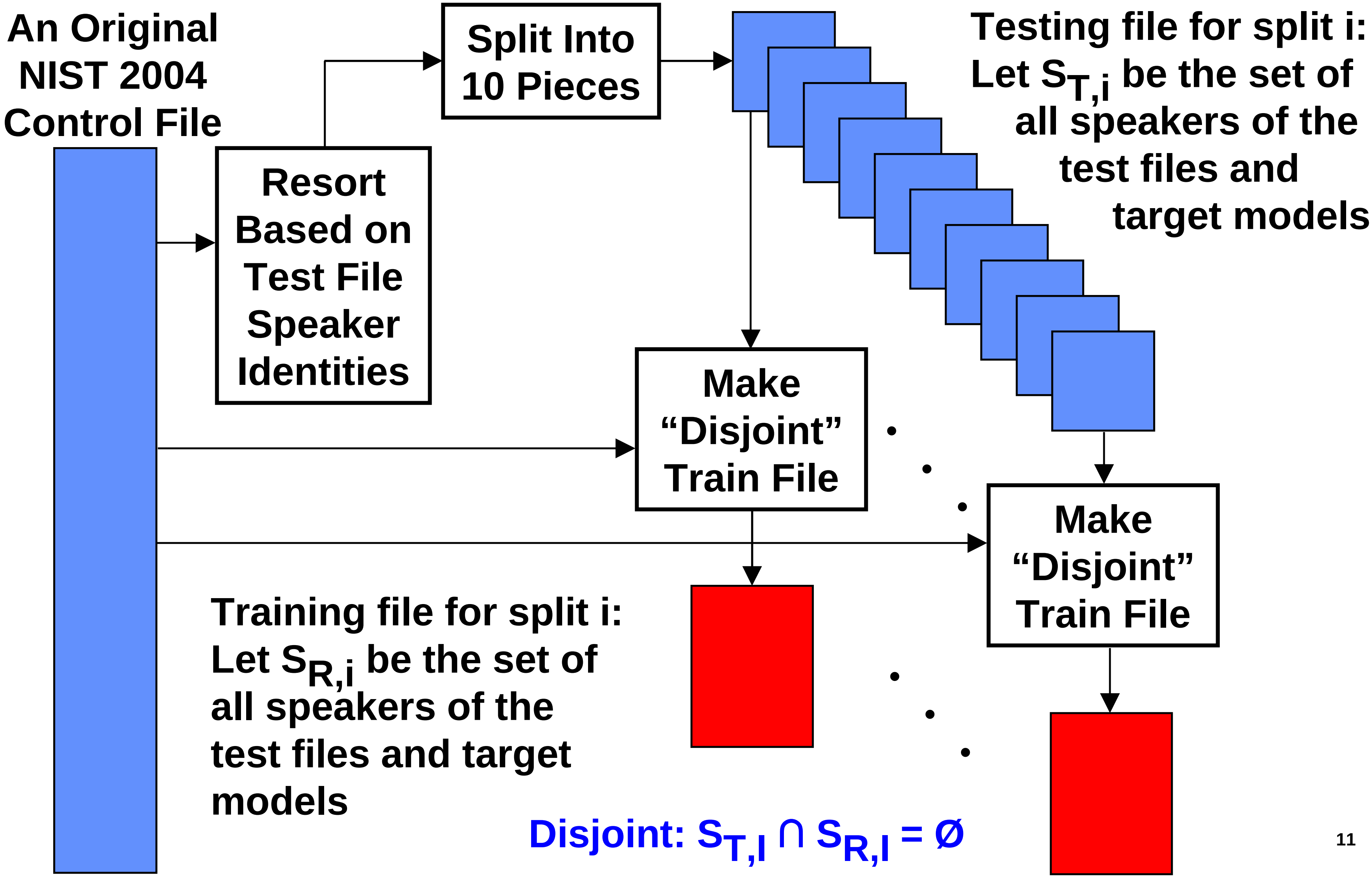
WLM System

- Used BBN transcripts provided by NIST
- Pseudo sentence breaks were added
- Bigram language models with back-off
- CMU-Cambridge Language Modeling Toolkit (Ver. 2.05) with top 20,000 words, Witten-Bell discounting, & zero cut-offs
- Score a test file vs. claimant model as:

$$\frac{1}{K}\sum_{k=1}^{K}\log(\Pr_{\text{Claimant}}(k)) - \log(\Pr_{\text{Background}}(k))$$

- K is the number of matching bigrams
- Background & 100 gender-independent T-norm models from SWB II

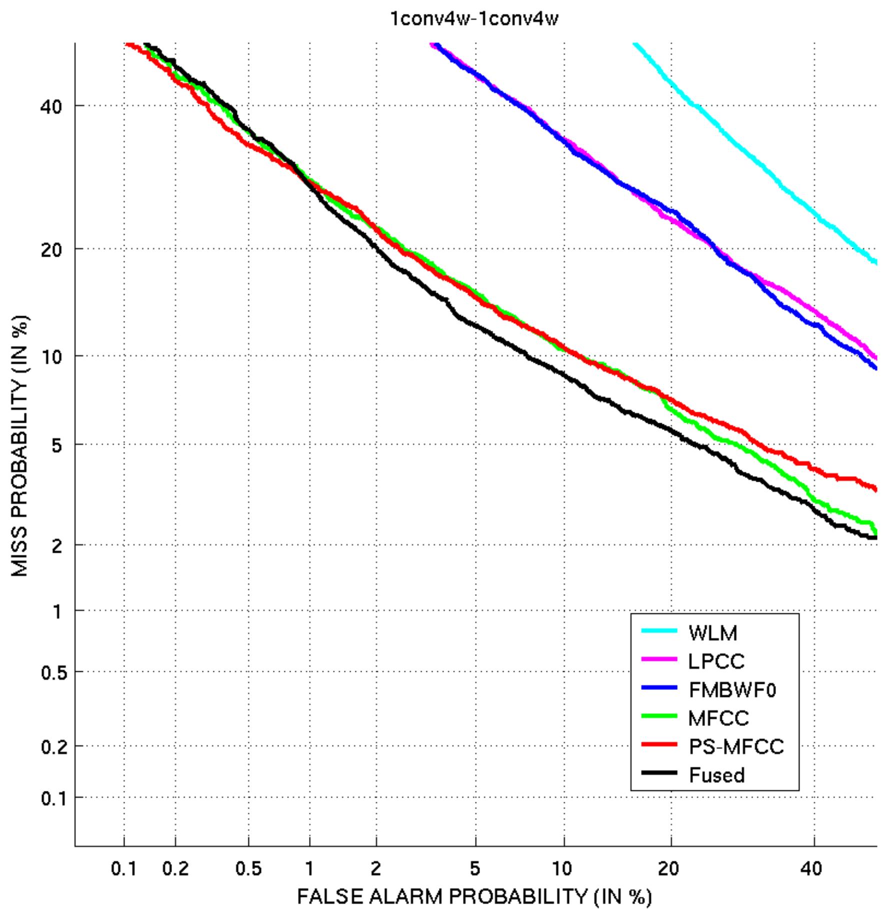
Splitting NIST 2004 Control Files

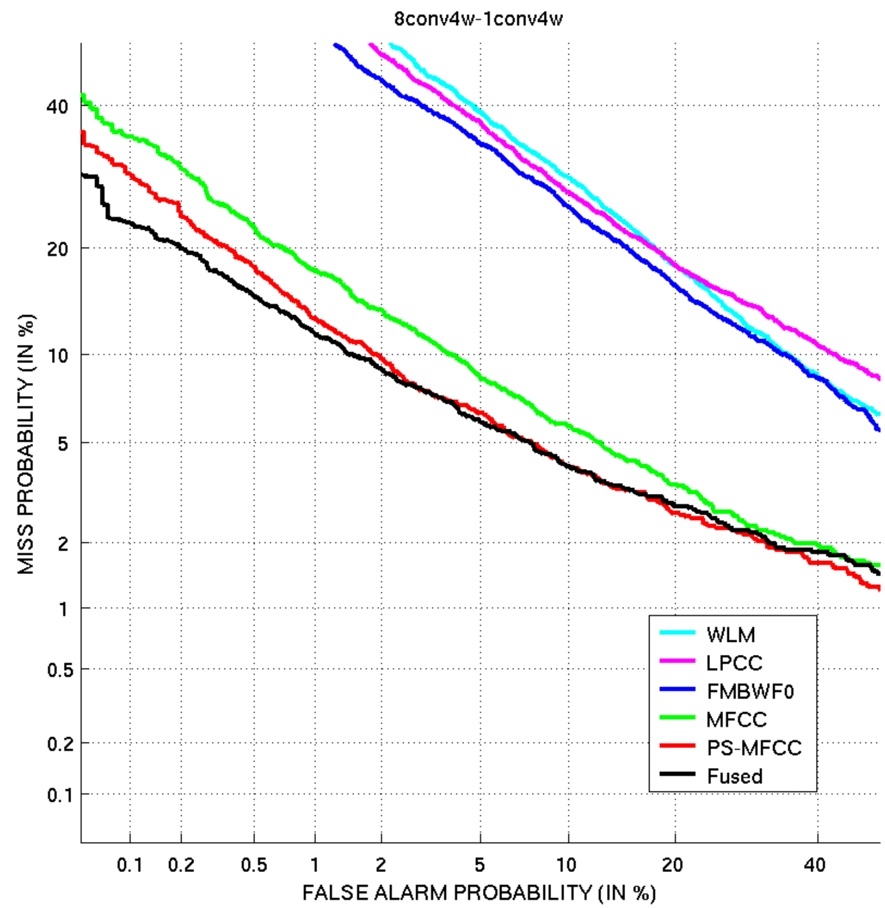


System Fusion and Thresholds

- For each split:
  - Build a single-layer perceptron (SLP) on the training file
  - Apply SLP to system scores for the test file
- Concatenate score files for the ten splits
- Determine threshold for minDCF (this is the threshold used for the 2005 Eval)
- Build new SLP over the entire control file for the condition (this is the SLP used for the 2005 Eval)
- SLPs built using LNKnet

Component Systems & Fusion '05

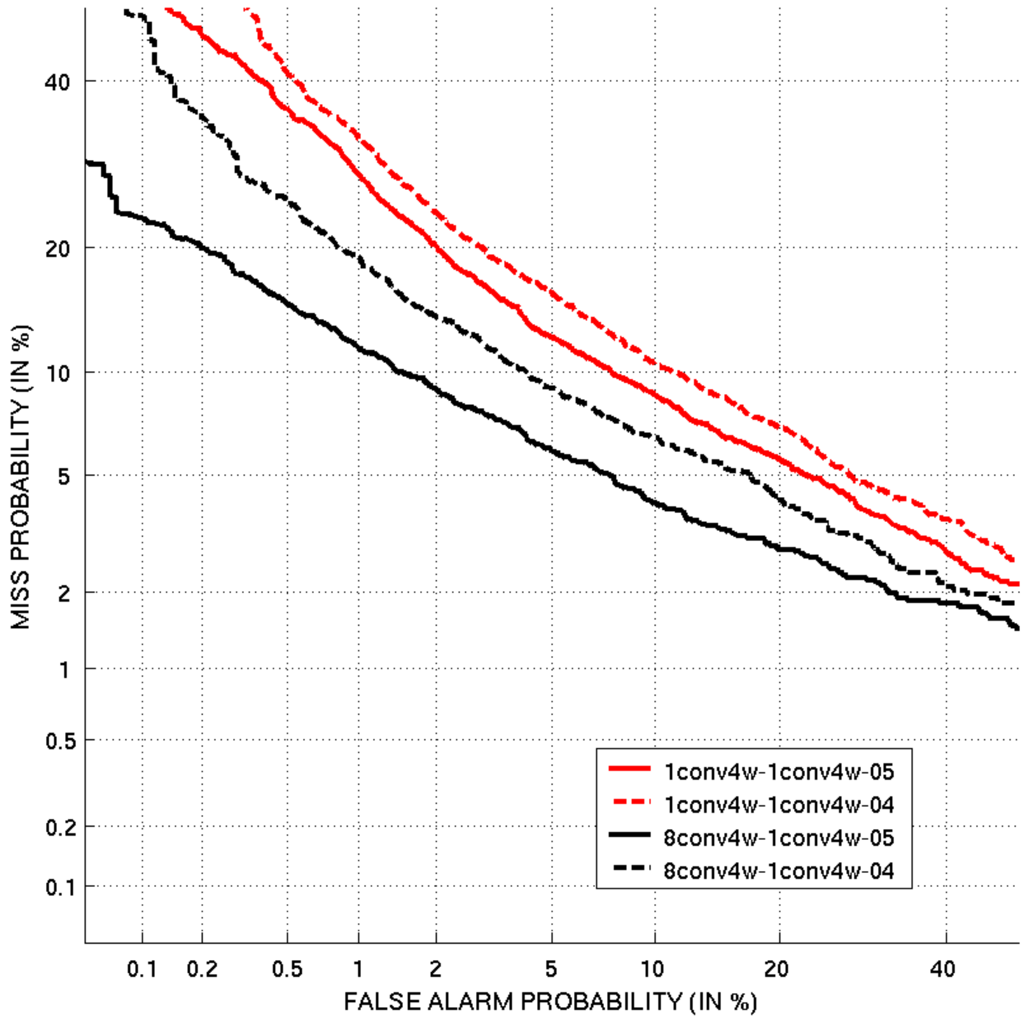




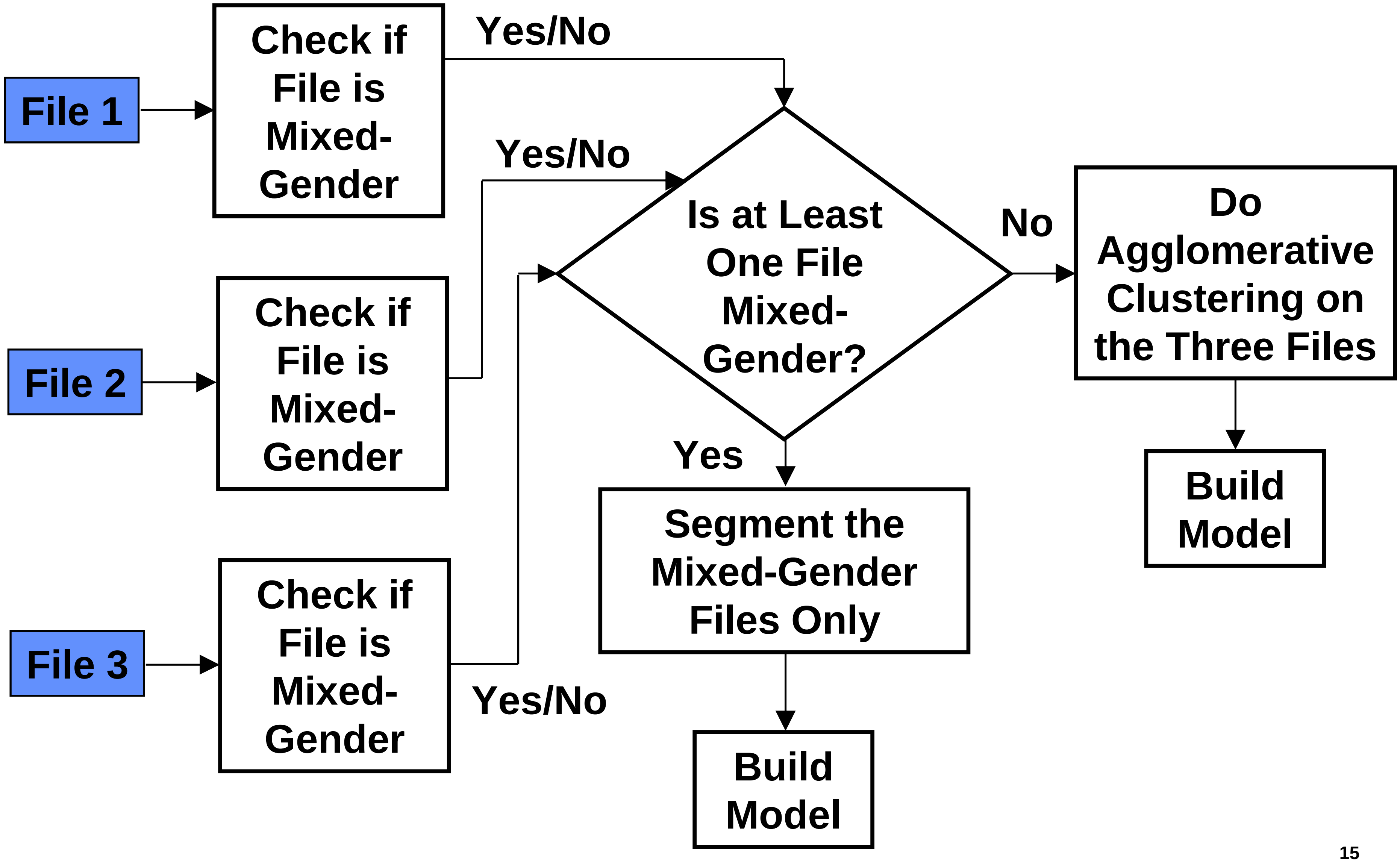
- **PS-MFCC system outperforms MFCC system for 8conv4w training**
- **PS-MFCC provides some benefit in fusion, even for 1conv4w training**

Comparison of Fusion '04 & '05

- **'05 Fusion Results: Scores from single SLP built on 2004 data for a given condition**
- **'04 Fusion Results: Concatenation of scores from fusion on the splits for a given condition**
- **Threshold differences between '04 and '05:**
  - **Differences in data**
  - **Differences in '04 and '05 fusion methods**



3conv2w Training



Gender Determination/Segmentation

- **For a file:**
  - **MFCC/HMM++ SAD (1) to find speech/non-speech segments**
  - **Score each speech segment against male and female GMMs**
  - **Suppose target speaker is male: Label a segment female if**

    $Score_{Female}$(segment) > $Score_{Male}$(segment) + Threshold(lang)
  - **Similar procedure if target is female**
  - **If less than approx. 90% of the frames are classified as the same gender, declare the file to be mixed-gender**
- **If one or more files are mixed-gender: Top 90% of segments of proper gender from mixed-gender files used for target model**
- **MFCCs, 300–3138 Hz, RASTA, deltas, feat map, & mean & var norm**

Agglomerative Clustering

For each file:

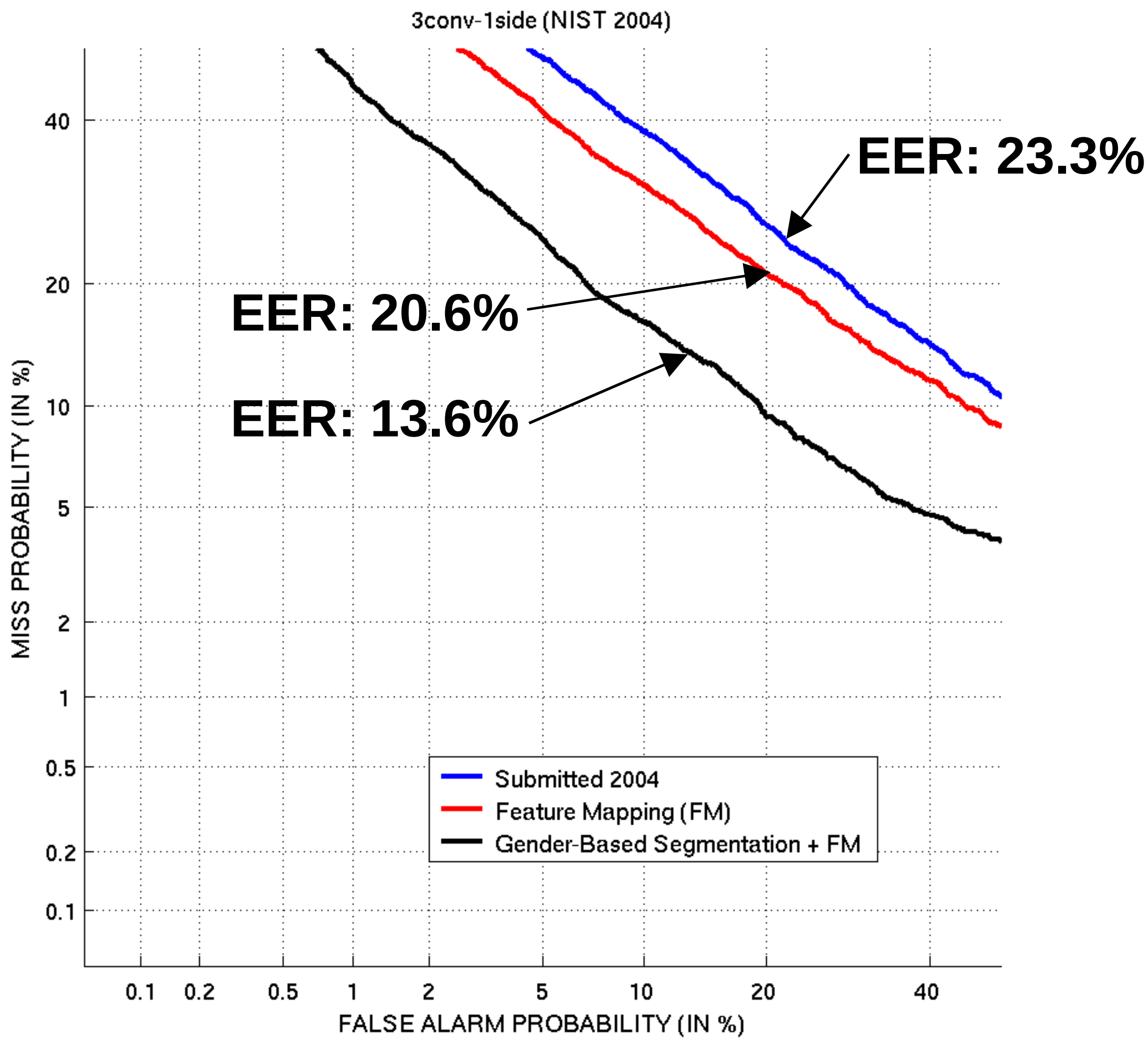
- **Determine speech/non-speech segments: MFCC/HMM++ SAD (2)**
  - **MFCCs, 200–2860 Hz, deltas, no RASTA, no feature mapping**
  - **80 mixtures/state trained from SWB II data & SRI transcripts**
- **64-mixture GMM trained using all speech vectors**
- **Weights then adapted for each speech segment**
- **In each clustering stage, vectors for each segment scored against all models & highest scoring feature vector/model pair merged**
- **Repeat the process until three sets of segments left (presumably, one for each speaker and a "garbage" set)**

Agglomerative Clustering: Use

- **Use:**
  - **If no mixed-gender files in 3conv2w training**
  - **In 1conv2w testing**
- **3conv2w training: After each file segmented & clustered, cluster across the three speech files using final features**
- **1conv2w testing: Test each of the three segments against the claimant model and take the maximum score**
- **Final features: MFCCs, 300–3138 Hz, RASTA, deltas, feature mapping, & mean & variance normalization**

Segmentation on NIST 2004 Data

- 2004 version of 3conv2w-1conv4w
- Blue line: Submitted 2004 system:
  - Agglomerative clustering only
  - No feature mapping
- Red line: 2004 system with feature mapping
- Black line: 2005 system on 2004 data (*i.e.,* using gender-based segmentation)
- Gender-based segmentation helped significantly



Gender-Based Segmentation: Stats

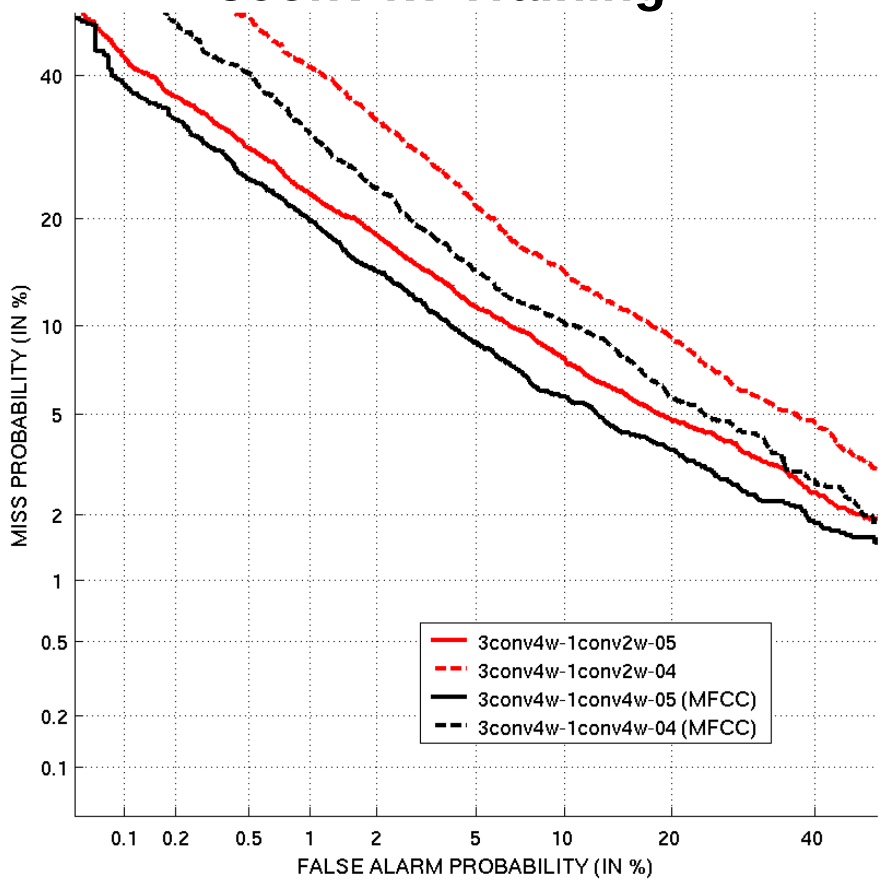
NIST 2004 3conv Training		
Number of Mixed-Gender Files/Model	True Percentage	Estimated Percentage
0	39.4	31.4
1	22.1	21.6
2	17.7	24.5
3	20.8	22.5

NIST 2005 3conv2w Training	
Number of Mixed-Gender Files/Model	Estimated Percentage
0	19.1
1	32.2
2	33.6
3	15.1

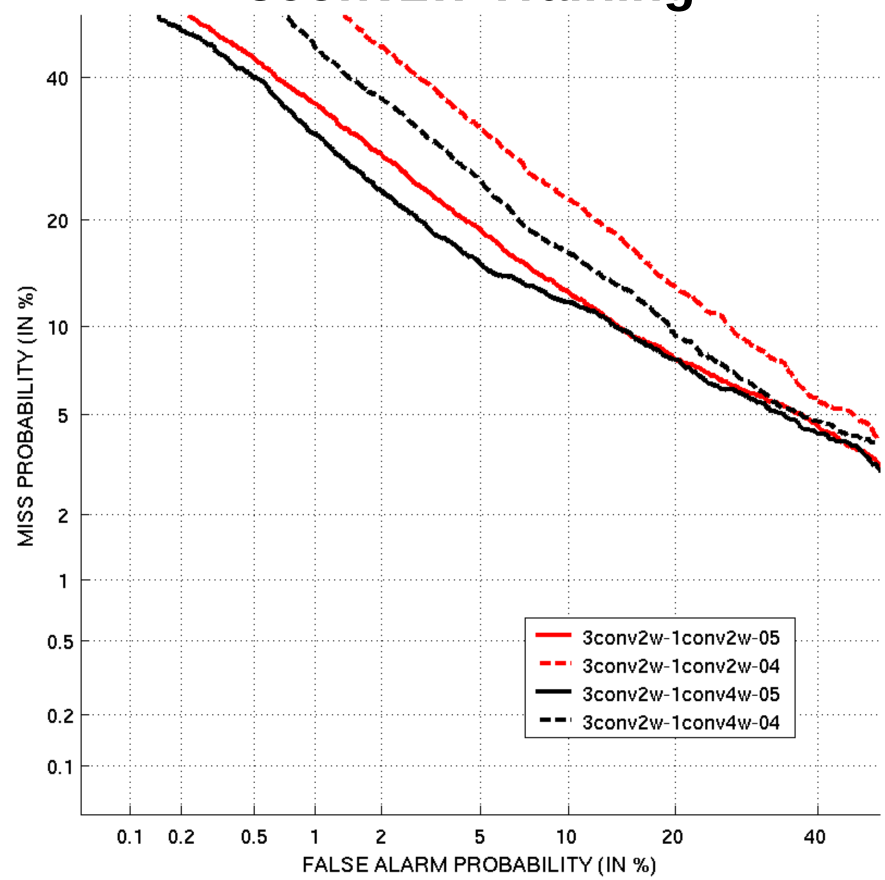
Required Agglomerative Clustering

Segmentation Results

3conv4w Training



3conv2w Training



- **MFCC systems only (no fusion here)**
- **2005 conditions considerably easier than corresponding 2004 conditions**

Acknowledgements

- MIT Lincoln Laboratory:
  - MFCC/GMM and feature mapping code
  - LNKnet
- Bryan Pellom, Univ. of Colorado at Boulder: SONIC speech recognizer, acoustic & language models
- Cambridge Univ.:
  - Statistical Language Modeling Toolkit (with CMU)
  - HTK
- KTH: Snack toolkit