The 2006 AFRL/HEC Speaker Recognition Systems

Raymond E. Slyh
Human Effectiveness Directorate
Air Force Research Laboratory

Team Members

- Eric Hansen, Human Effectiveness Directorate, Air Force Research Laboratory (AFRL/HEC)
- Brian Ore, General Dynamics Advanced Information Systems
- Raymond Slyh, Human Effectiveness Directorate, Air Force Research Laboratory (AFRL/HEC)

Components of Submitted Systems

Conditions		TESTING 10sec4w	TESTING 1conv4w	TESTING 1conv2w
TRAINING	10sec4w	FMBWF0/GMM MFCC/GMM MFCC/SVM		
TRAINING	1conv4w	FMBWF0/GMM MFCC/GMM MFCC/SVM	FMBWF0/GMM MFCC/GMM MFCC/SVM MFCC/PS-GMM WLM	FMBWF0/GMM MFCC/GMM MFCC/SVM
TRAINING	3conv4w	FMBWF0/GMM MFCC/GMM MFCC/SVM	FMBWF0/GMM MFCC/GMM MFCC/SVM MFCC/PS-GMM WLM	FMBWF0/GMM MFCC/GMM MFCC/SVM
TRAINING	8conv4w	FMBWF0/GMM MFCC/GMM MFCC/SVM	FMBWF0/GMM MFCC/GMM MFCC/SVM MFCC/PS-GMM WLM	FMBWF0/GMM MFCC/GMM MFCC/SVM
TRAINING	3conv2w		FMBWF0/GMM MFCC/GMM MFCC/SVM	FMBWF0/GMM MFCC/GMM MFCC/SVM

KEY

FMBWF0: F1–F3, BW1–BW3, log(F0)

MFCC: Mel-Frequency Cepstral Coeffs & Δs

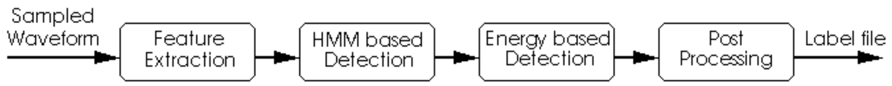
GMM: Gaussian Mixture Models

SVM: Support Vector Machines

PS-GMM: Phoneme-Specific GMMs

WLM: Language Modeling on Transcripts Output by SONIC

MFCC/HMM SAD



- **Features: 19 MFCCs (300–3138 Hz) & Δs (No RASTA)**
- **HMM-based speech activity detector (SAD):**
  - **Two-state HMM built with HTK (64 mixtures/state)**
  - **Trained on background model (BKG) data using SONIC labels as truth**
- **Energy-based detector:**
  - **From MIT-LL *xtalkN***
  - **Refines the output from the HMM-based detector**
- **Post-Processing: Removes speech segments < 20 msec**

GMM-Based Systems

- Gaussian mixture models from MIT Lincoln Laboratory (MIT-LL) system with:
  - 2048 mixtures per model
  - Diagonal covariance matrices
- T-norm applied to output scores
- (Initial) speaker & T-norm models built using MAP adaptation from BKG with:
  - Relevance factor of 16
  - Only mixture means adapted

GMM-Based Systems: Models

- BKG:
  - 16 hours of data balanced for gender & channel from:
    - NIST 2001–2003 Evals (digital cell, electret, & carb.)
    - OGI National Cellular Corpus (for analog cellular)
  - Gender/channel models used for feature mapping
- T-norm:
  - Other than 10sec4w training:
    - Gender-dependent: 120 models per gender
    - Single conversation sides from NIST 2001–2003 Evals
  - 10sec4w training:
    - 240 gender-independent models
    - First 30 sec of data from original set of models

MFCC/GMM System: Features

- 19 MFCCs every 10 msec with:
  - Bandwidth of 300–3138 Hz
  - No 0^{th} coefficient
- Applied RASTA filtering & calculated Δs of features
- Kept a frame if labeled as speech by MFCC/HMM SAD
- Applied feature mapping and mean & variance norm.

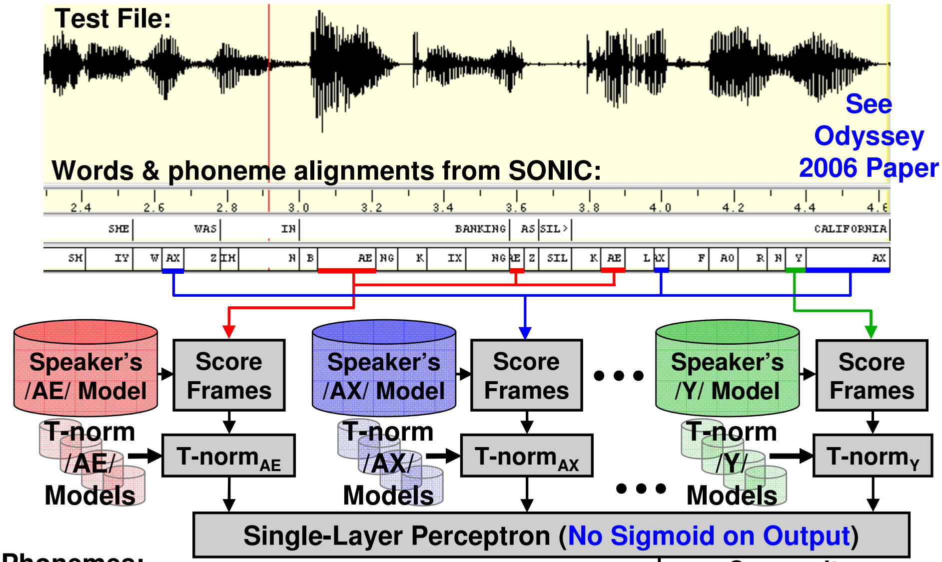
FMBWF0/GMM System: Features

- Every 10 msec:
  - Formant center frequencies (F1–F3) & bandwidths (BW1–BW3) using Snack toolkit from KTH
  - F0 & probability of voicing using *get_f0* from ESPS
- Kept a frame if:

  (speech) AND (voiced) AND (F0 < 250 Hz) AND
  {(F1, F2, F3) != (500 Hz, 1500 Hz, 2500 Hz)}

- Converted F1–F3 & BW1–BW3 to radians & took log(F0)
- Applied feature mapping (with channel picked by MFCCs)

MFCC/PS-GMM System



Phonemes:
{AE, AH, AX, AY, DH, IH, IY, L, M, N, OW, S, TD, W, Y}

MFCC/SVM System

- Features as in MFCC/GMM system
- Support vector machine classifier:
  - Generalized linear discriminant sequence kernel
  - From MIT-LL speech tools
- T-norm applied to scores (with T-norm models built using same data as for GMM systems)

WLM System

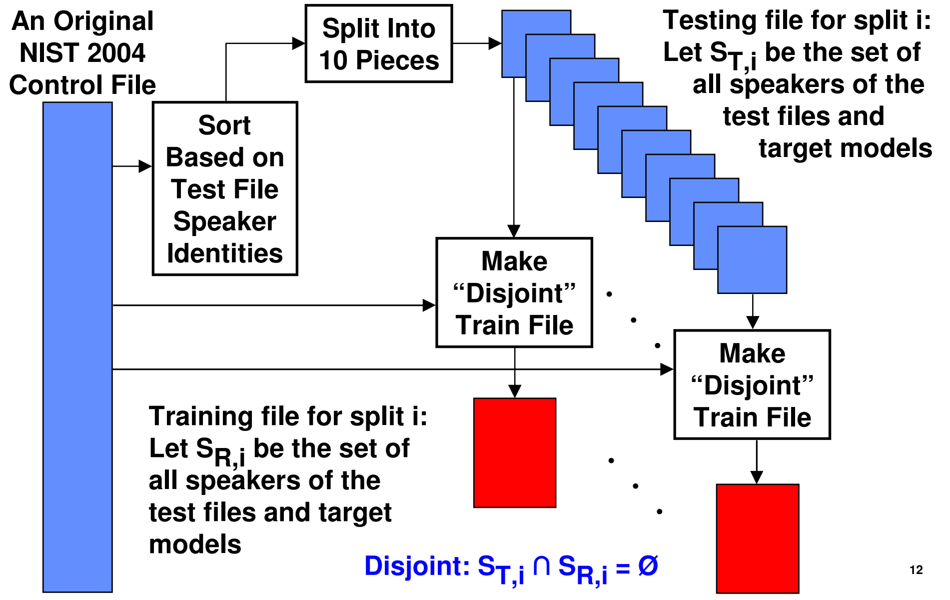
- Used (English) transcripts generated by SONIC
- Pseudo sentence breaks were added
- Bigram language models with back-off
- CMU-Cambridge Language Modeling Toolkit with top 20,000 words, Witten-Bell discounting, & zero cut-offs
- Score a test file vs. claimant model as:

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

where K is the number of matching bigrams

- 100 gender-independent two-conversation T-norm models from SWB II

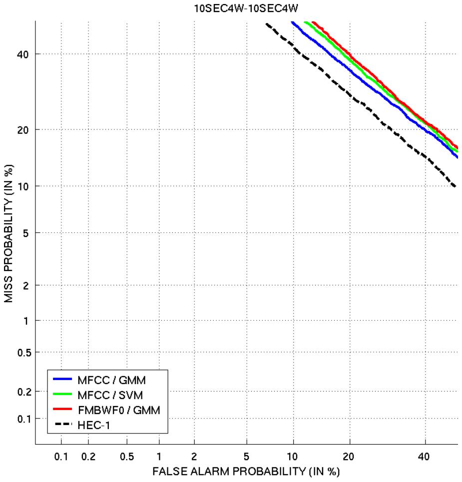
Splitting NIST 2004 Control Files

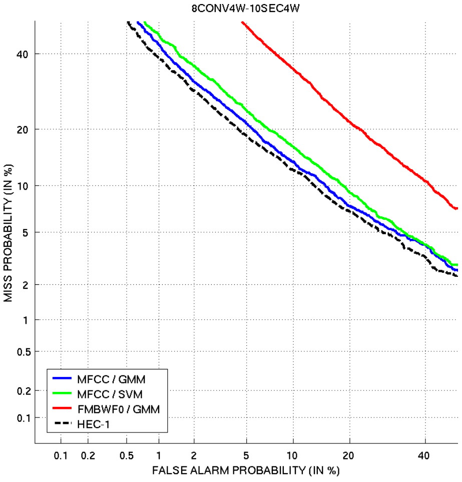


Four-Wire Fusion & Thresholds

- For each split:
  - Built a single-layer perceptron (SLP) on training file
  - Applied SLP to system scores for the test file
- Concatenated score files for the ten splits
- Determined threshold for minDCF (this was the threshold used for the 2006 Eval)
- Built new SLP over the entire control file for the condition (this was the SLP used for the 2006 Eval)
- SLPs built using LNKnet from MIT-LL

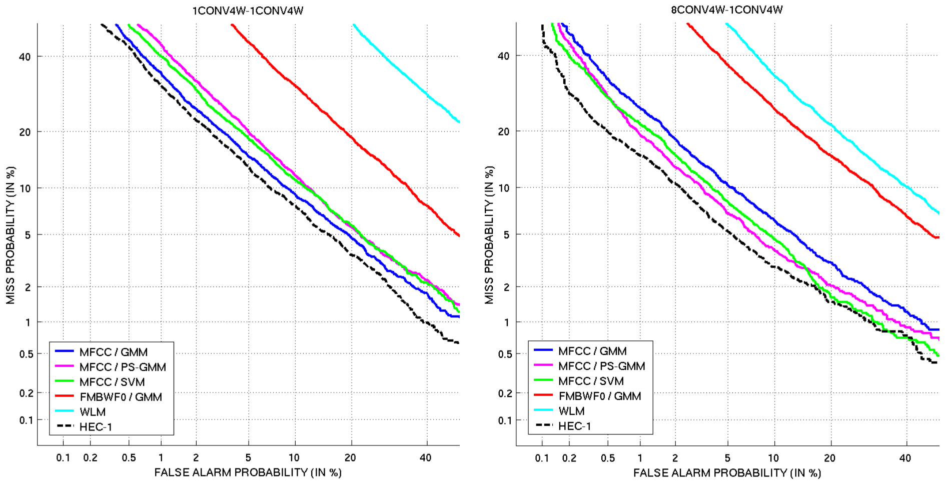
10sec4w Testing





Score combination provided considerable benefit for 10sec4w training but less benefit for larger amounts of training data

1conv4w Testing



MFCC/PS-GMM system outperformed MFCC/GMM system for 8conv4w training even though it used only 15 out of 50 *English* phonemes

Unsupervised Adaptation

- **HEC-2 System: MFCC/GMM system with & without unsupervised adaptation (UA) of mixture means, $\overline{\mu}_m$:**

$$\overline{\mu}_m^{\mathrm{NEW}} = \alpha_m E_m(X) + (1-\alpha_m)\overline{\mu}_m$$

$E_m(X)$: **Mean of vectors prob. assigned to mixture m**

- **Initial speaker models built using MAP adaptation from BKG with:**

$$\alpha_m = \frac{n_m(X)}{n_m(X)+r}$$

$n_m(X)$: **Probabilistic "count" of vectors in mixture m**

r: **Relevance factor = 16**

- **Updated speaker model built using MAP adaptation from current speaker model with:**

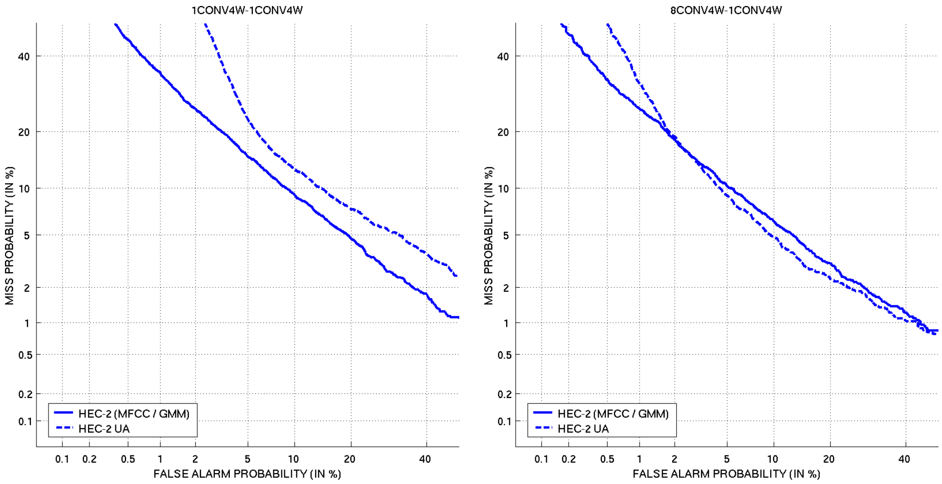
$$\alpha_m = \begin{cases} 0.1, & \beta < 0.1 \\ 0.5, & 0.5 < \beta \\ \beta, & \text{otherwise} \end{cases} \qquad \beta = \frac{T_T}{T_T + T_M}$$

T_T : **# speech frames in test file**

T_M : **# speech frames used for current model**

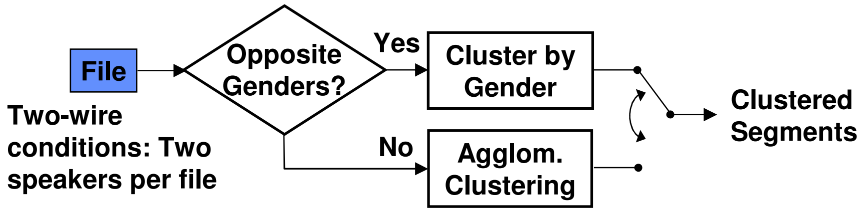
- **See Odyssey 2006 paper for more details**

Unsupervised Adaptation



- Model updating threshold: minDCF threshold from NIST 2004 Eval data
- UA degraded performance: Need a different updating threshold?

Two-Wire Segmentation/Clustering



- 1conv2w testing:
  - If gender-based clustering used for a file: Test correct-gender cluster against target model
  - If agglomerative clustering used for a file: Cluster into three sets, test each set against the target model, & pick the highest score
- 3conv2w training:
  1) Segment & cluster each of the three files individually
  2) Cluster across the three files
  3) Build model

Opposite-Gender Files & Clustering

- **Opposite-Gender File Determination:**
  - MFCC/HMM SAD determines speech/non-speech segments
  - Score files against male, female, & BKG GMMs
  - If target speaker is male, label a file opposite-gender if:

    $\text{Score}_{\text{BKG}} - \text{Score}_{\text{Male}} >$ Gender-dependent threshold
  - Similar procedure if target is female
- **Gender-Based Clustering:**
  - MFCCs, 300–3138 Hz, RASTA, Δs, but no feature mapping
  - Score each segment individually against male & female GMMs
  - Take top 90% of the segments of proper gender for target model
- See Odyssey 2006 paper for more details

Agglomerative Clustering Within File

- MFCC/HMM SAD determined speech/non-speech segments
- 64-mixture GMM trained with all speech vectors from the file using MFCCs band limited to 200–2860 Hz and Δs, but without RASTA filtering, feature mapping, or mean & variance normalization
- Weights then adapted for each speech segment
- In each clustering stage:
  - Let X and Y be two segments, and let $Z = X \cup Y$
  - $\forall X, Y$ calculate:

    $$\Lambda(X,Y) = \frac{L(Z|\theta_Z)}{L(X|\theta_X)L(Y|\theta_Y)}$$

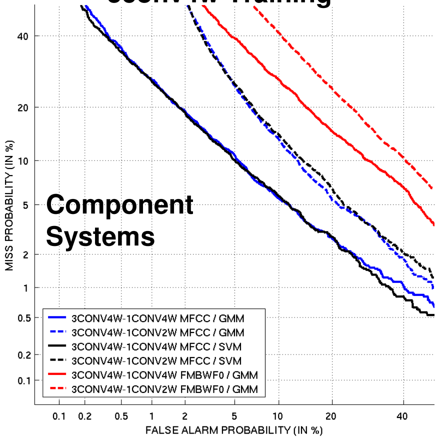
    $L(X|\theta_X)$: Likelihood of data for segment X given model for X
  - Merge the X and Y segments with the highest $\Lambda(X,Y)$
- Repeat the process until three sets of segments are left (presumably, one for each speaker and a "garbage" set)

3conv2w: Clustering Across Files

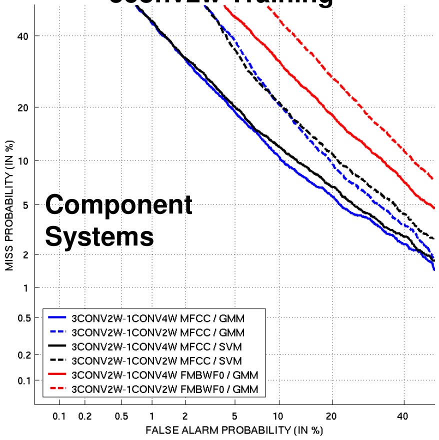
- Features: 19 MFCCs with a bandwidth of 300–3138 Hz, RASTA, Δs, feature mapping, and mean & variance normalization
- If any files were segmented by gender:
  - Correct-gender segments used to build an initial speaker model
  - Segments from other files tested against the initial speaker model
- If no files were segmented by gender:
  - Models were built for each of the three segment sets in each file by using MAP adaptation of mixture means from BKG
  - Segments were scored against the models (from other files) & highest scoring segment/model pair was clustered
  - Segments from third file tested against the clustered model

Segmentation Results

3conv4w Training



3conv2w Training



- Comparisons within a plot show effect of two-wire testing, while comparisons across the plots show the effect of two-wire training
- Substantial performance difference between 3conv4w & 3conv2w training and between 1conv4w & 1conv2w testing

Acknowledgements

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