MIT Lincoln Laboratory
Site Presentation

**Doug Reynolds, Bill Campbell, Wade Shen,
Doug Sturim, Pedro Torres-Carrasquillo,
Andre Adami***

NIST Speaker Recognition Workshop

07 June 2005

*OGI

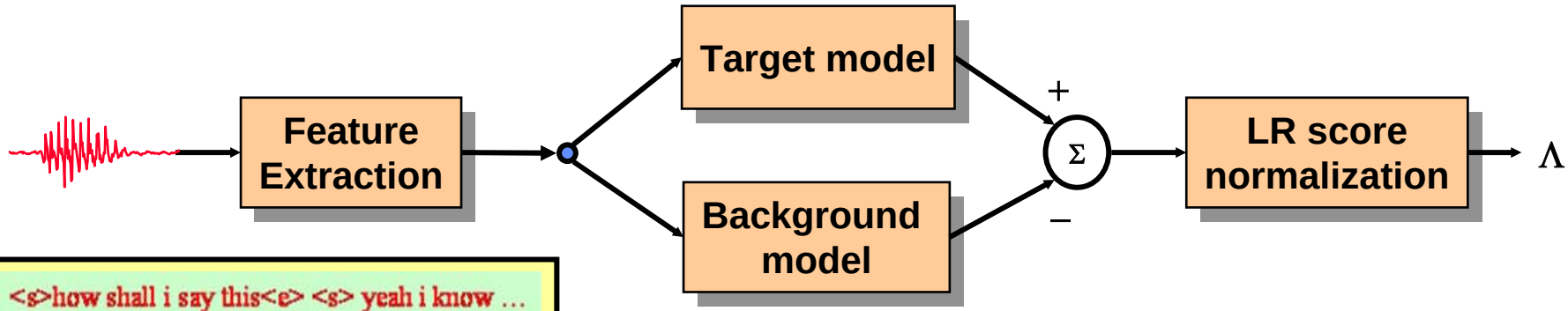
This work was sponsored by the Department of Defense under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

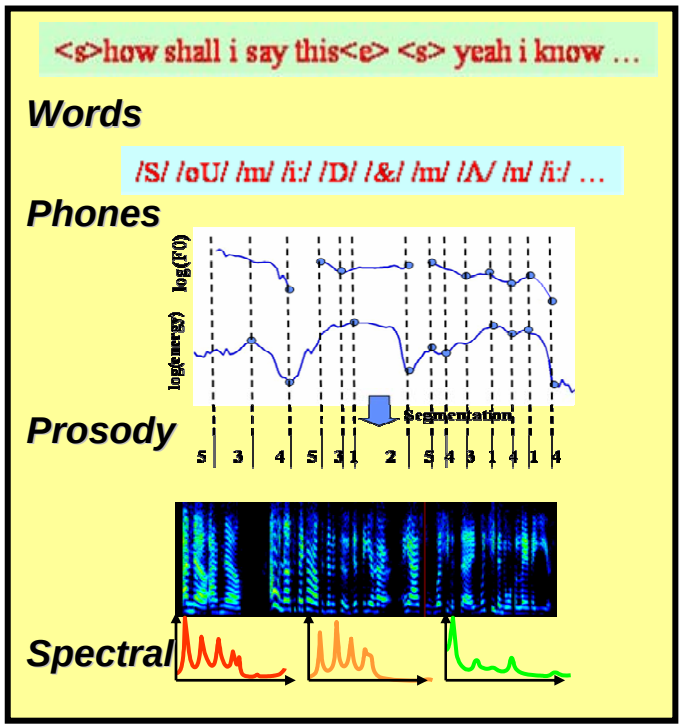
Outline

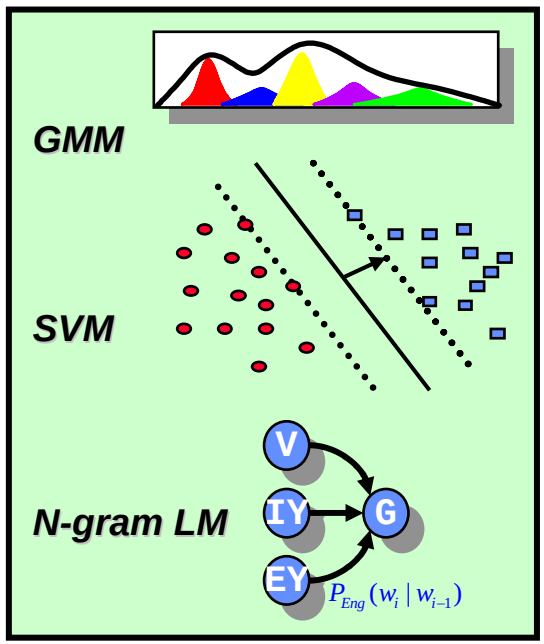
- System Overview
- New for 2005
  - Data Sets
  - SVM NAP
  - SVM Text-constrained
  - Phone and word lattice
  - Metadata for fusion
- Analysis
  - 2 wire vs. 4 wire
  - Conv. mic vs. main conditions
  - AT-Norm cohort selection
  - Metadata fusion
  - Word LLR Smoothing
  - Phonetic Refraction
- Conclusion

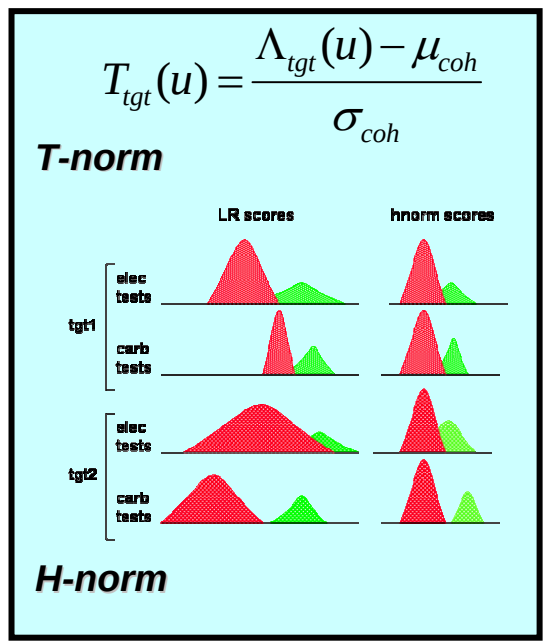
System Overview
Likelihood Ratio Detector

- Basic decision statistic in core detectors is the likelihood-ratio





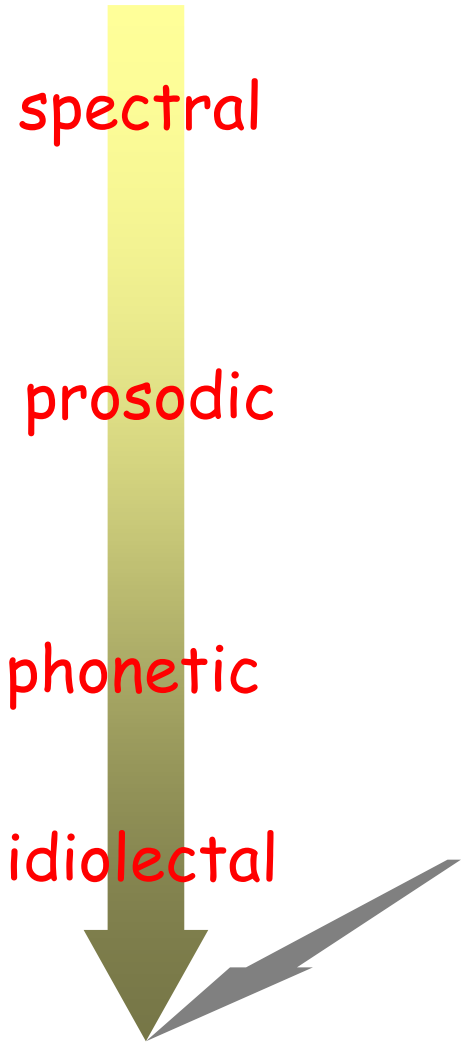




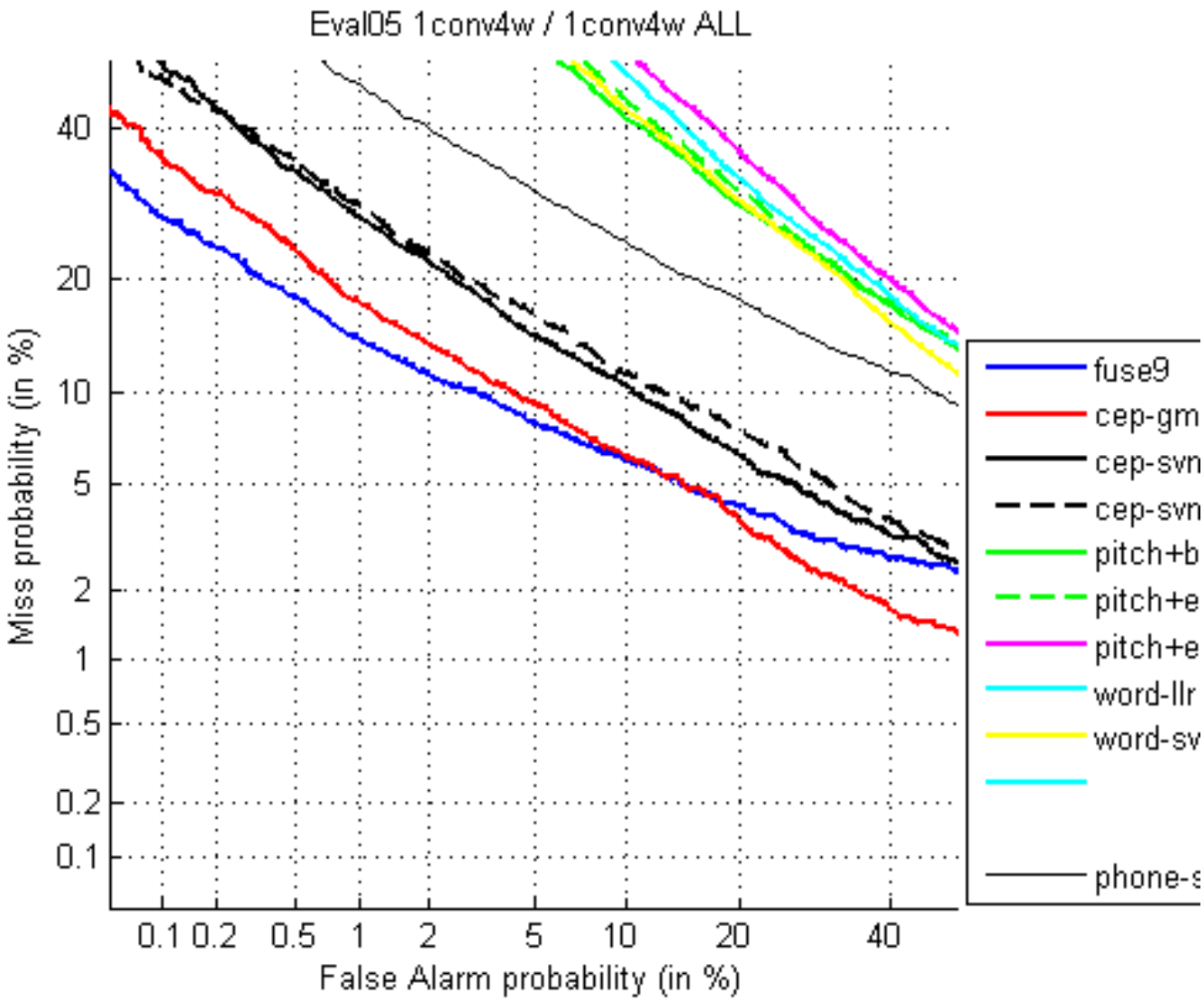
System Overview

Core Detectors

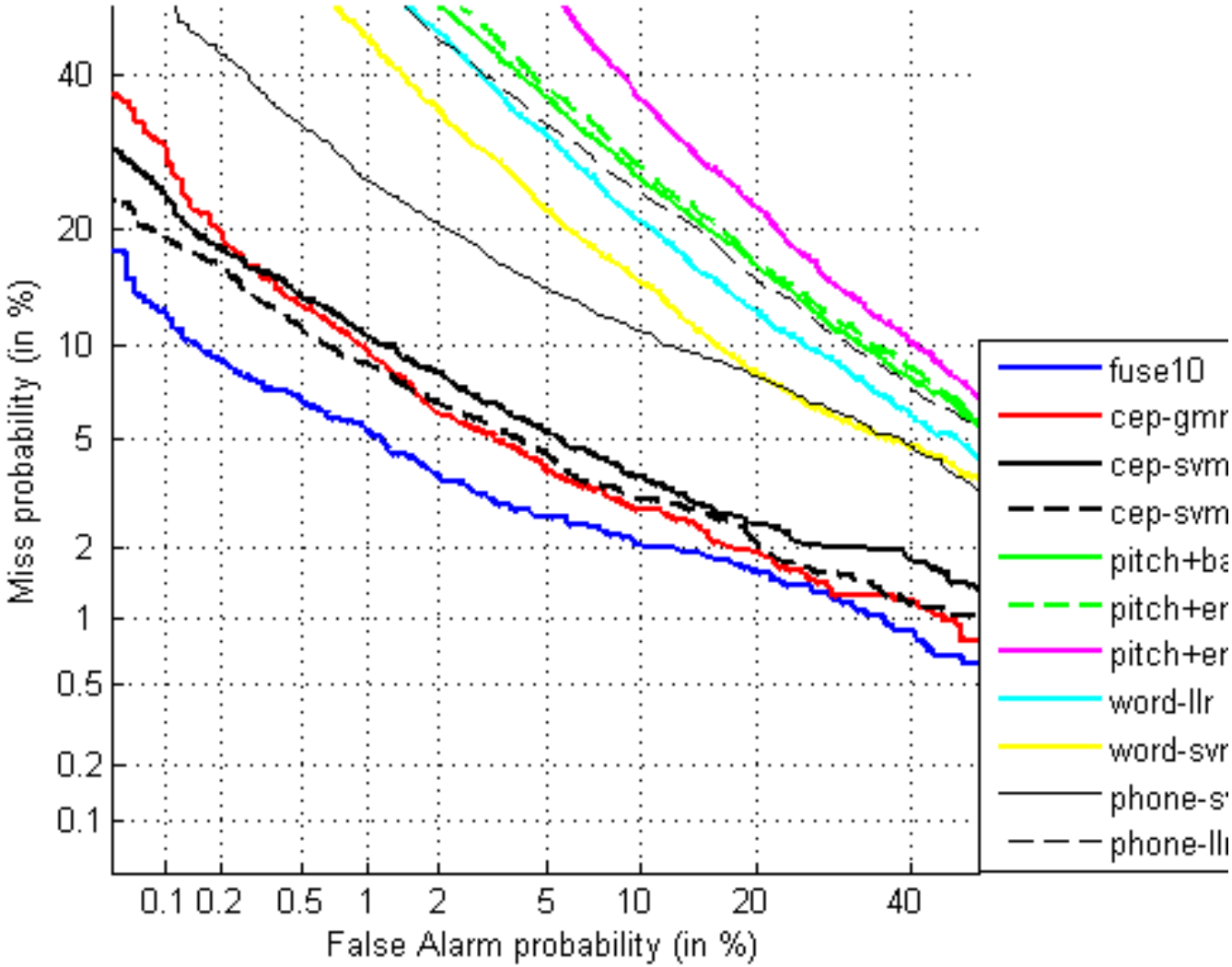
- MFCC GMM-UBM system
- MFCC NAP SVM GLDS kernel system
- MFCC SVM GLDS kernel system
- MFCC SVM Text-Constrained
- Pitch/Energy GMM-UBM system
- Pitch/Energy Slope N-gram system
- Sub-band prosodic modeling
- Phone N-gram system
- Phone SVM system
- Word SVM system
- Word N-gram system
- Fusion



1conv4w / 1conv4w



8conv4w / 1conv4w



Outline

- System Overview
- **New for 2005**
  - **Data Sets**
  - **SVM NAP**
  - **SVM Text-constrained**
  - **Phone and word lattice**
  - **Metadata for fusion**
- Analysis
  - 2 wire vs. 4 wire
  - Conv. mic vs. main conditions
  - AT-Norm cohort selection
  - Metadata fusion
  - Word LLR Smoothing
  - Phonetic Refraction
- Conclusion

Data Sets

Dev Set Design

- Evaluation ’04 indices were redesigned to focus on harder impostor trials
  - Target trials were same as Eval04
  - Imposter trials were chosen only from the same dialect as the target speaker (L1)
- Focus on “all” pooling and cross-language conditions
  - Turns out this was not really represented in eval05 data
- Development using two equal splits with non-overlapping speakers

GMM-UBM	8conv4w/1conv4w	1conv4w/1conv4w
eval04	eer=7.71 dcf=0.0327	eer=11.97 dcf=0.0478
dev05	eer=8.71 dcf=0.0380	eer=12.82 dcf=0.0516
eval05	eer=6.82 dcf=0.0270	eer=10.56 dcf=0.0392

Data Sets
Model Training

- GMM UBM background model and cohort training
  - SwbII, Fisher and Eval04
- SVM background
  - Fisher, CallFriend and CallHome
- Word and Phone SVM/LLR backgrounds
  - Fisher, CallFriend and CallHome
- Fusion
  - Eval04

Dev05 results	8conv4w/1conv4w	1conv4w/1conv4w
GMM-UBM SWBII	eer=8.71 dcf=0.0380	eer=12.82 dcf=0.0516
GMM-UBM SWBII, new SAD	eer=7.89 dcf=0.0373	eer=11.24 dcf=0.0493
GMM-UBM SWBII+eval04, new SAD	eer=7.26 dcf=0.0371	eer=10.94 dcf=0.0477

Nuisance Attribute Projection (NAP)*

- Remove directions which have significant channel variability using a projection:

$$\mathbf{P} = \mathbf{I} - ww^t$$

- The new kernel is:

$$K(x, y) \Rightarrow \mathbf{Pb}(x) \cdot \mathbf{Pb}(y)$$

- The projection is designed with the following criterion:

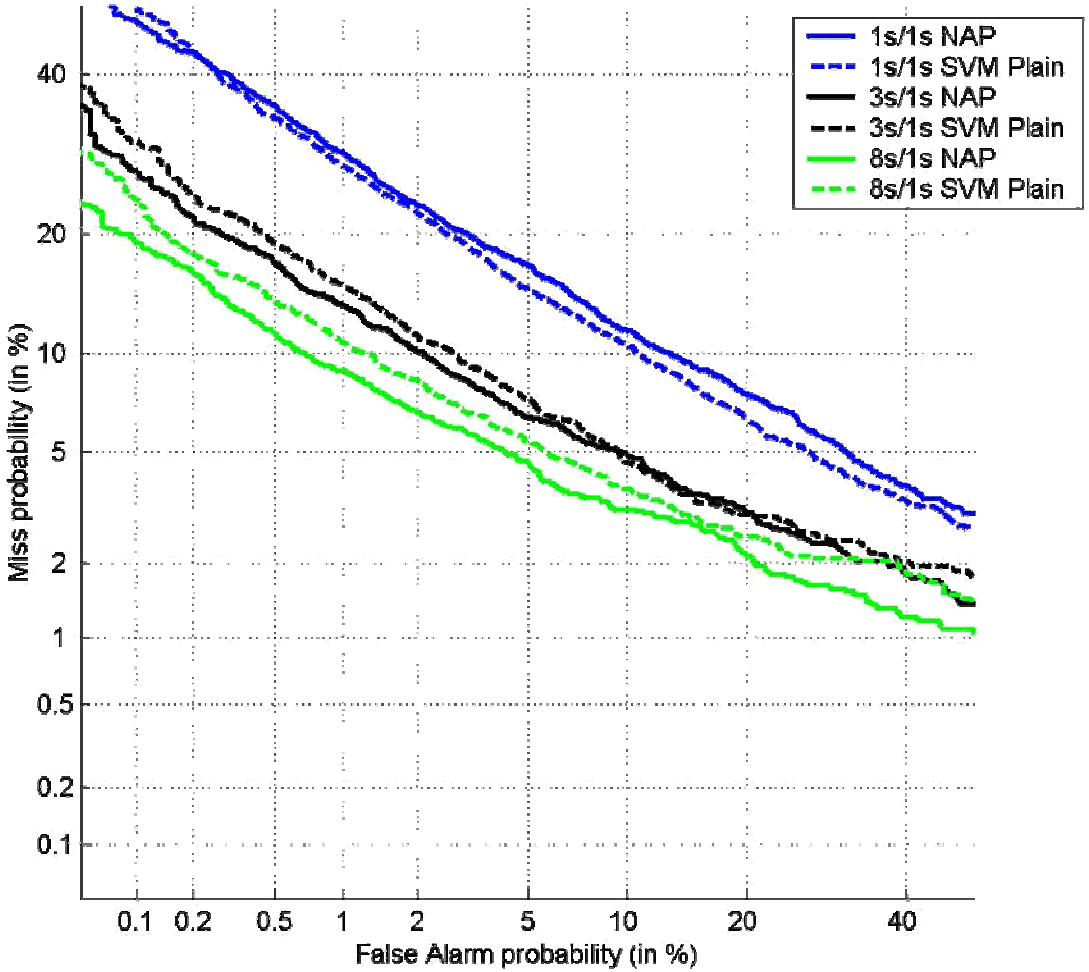
$$\delta = \sum_{ij} W_{ij} \left| \mathbf{P}(b(x_i) - b(x_j)) \right|^2 \quad W_{ij} = \begin{cases} 1 & \text{for } i, j \text{ same channel} \\ 0 & \text{for } i, j \text{ different channel} \end{cases}$$

- We trained the projection using automatically labeled data from the Fisher and CallFriend corpus. Labels were cell, electret, and carbon button.

*Alex Solomonoff, W. Campbell, I. Boardman, "Advances In Channel Compensation For SVM Speaker Recognition," ICASSP 2005, pp. 629-632.

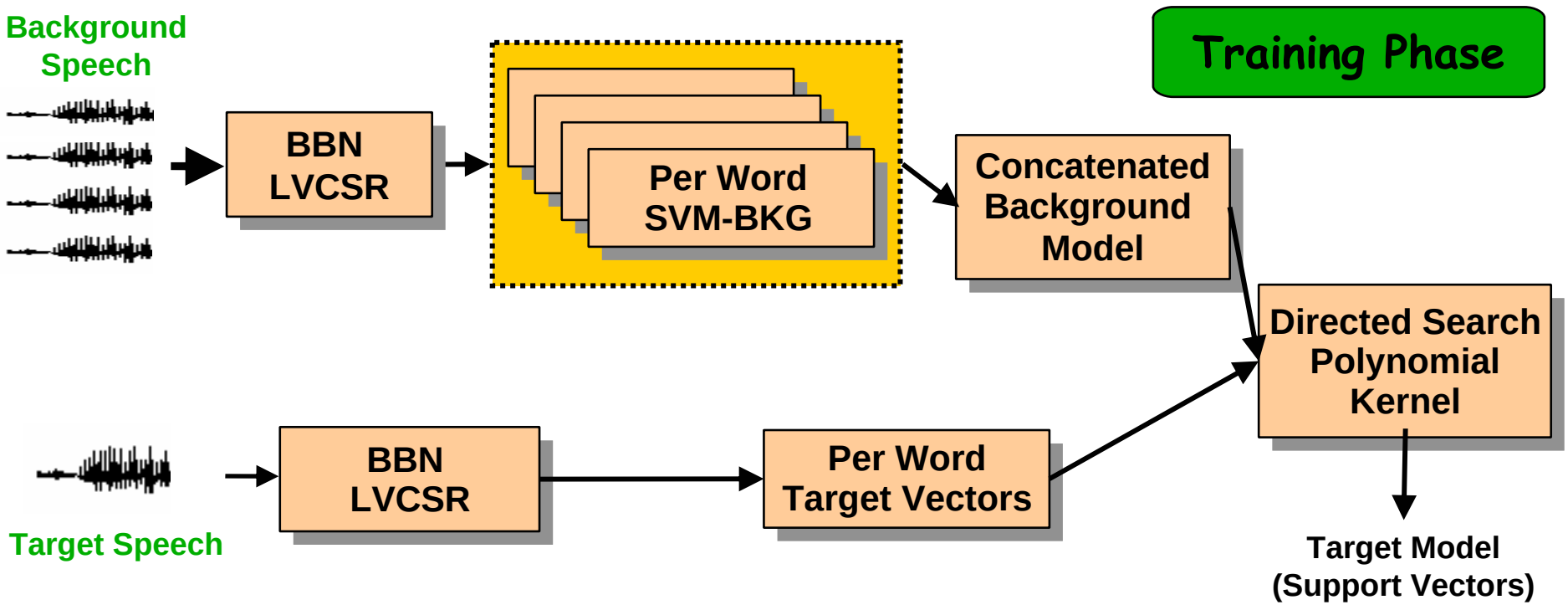
SVM NAP

Performance – All Trials



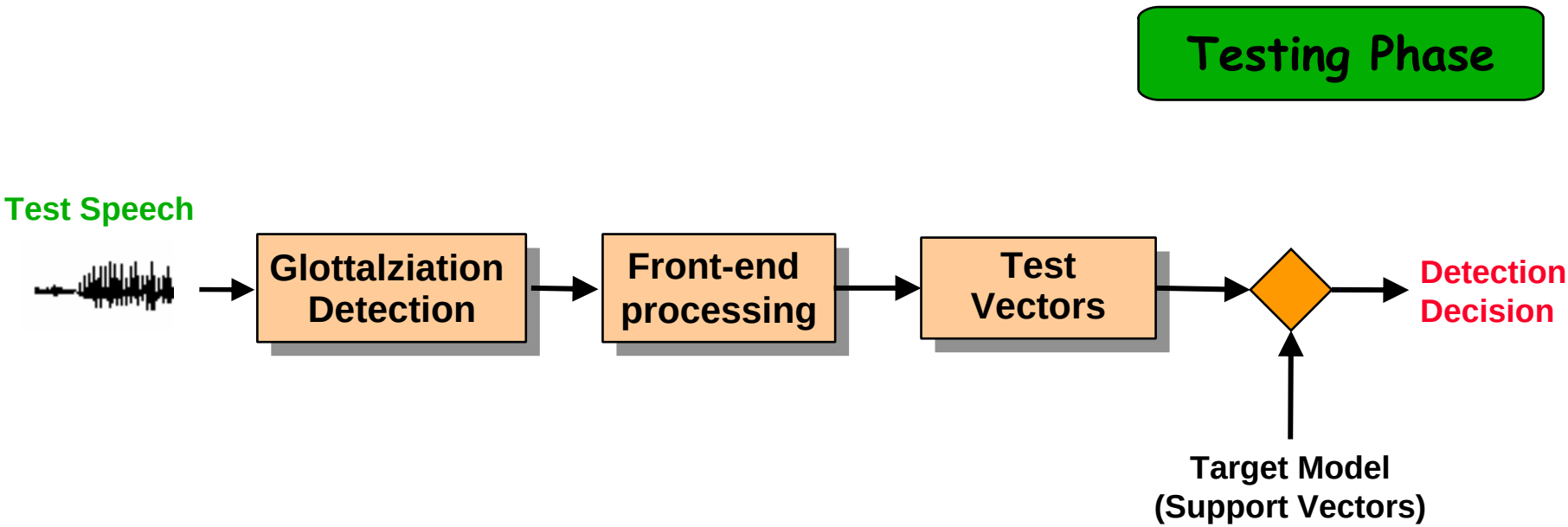
Text Constrained SVM

- Uses Base-Generalized Linear Discriminant Sequence (GLDS) Kernel with a directed search during training
- Expansion vectors are average per word per utterance
- Background trained using Fisher 2004 collect



Text Constrained SVM

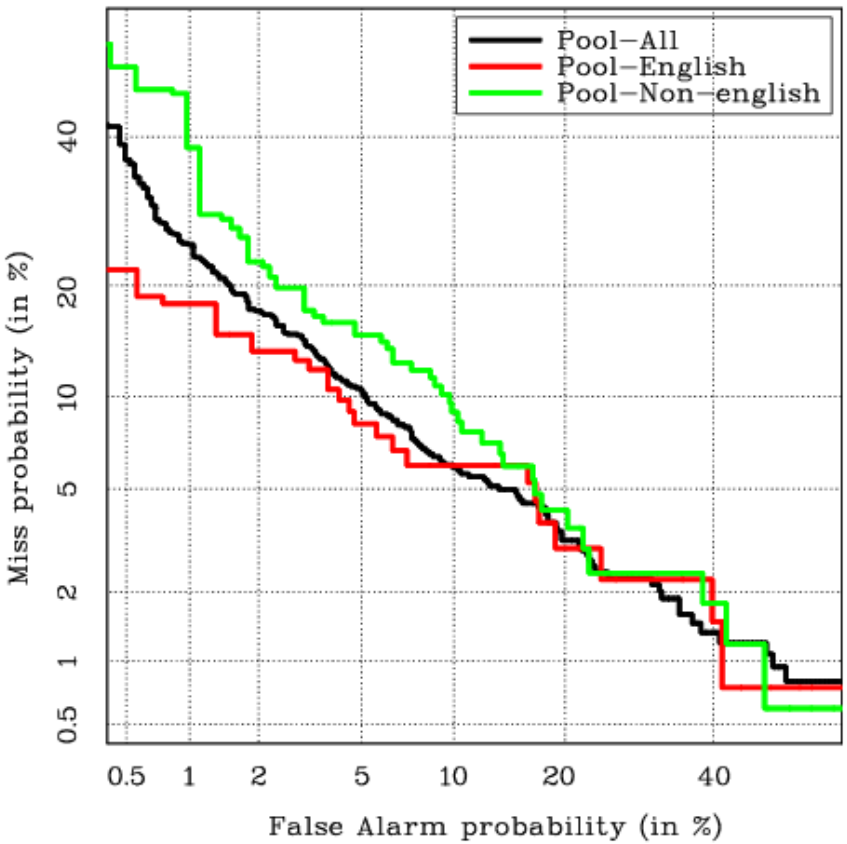
- Testing utilizes the same two step front-end
  - Detection
  - Feature Generation



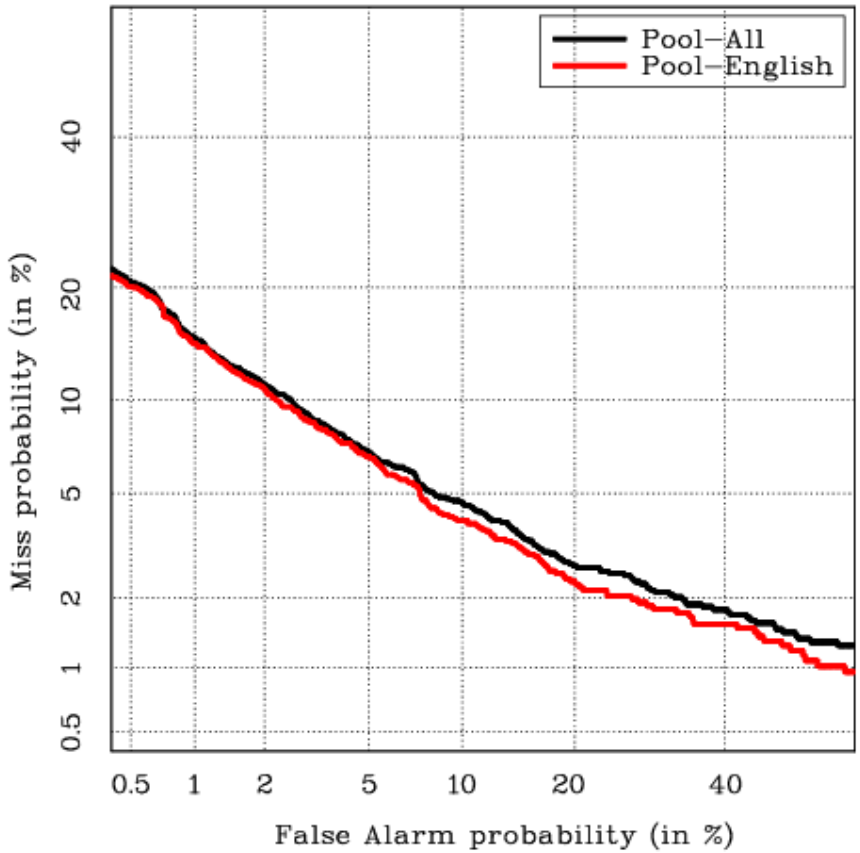
Text-Constrained SVM

Results

Dev05 Train-8c / Test-1c

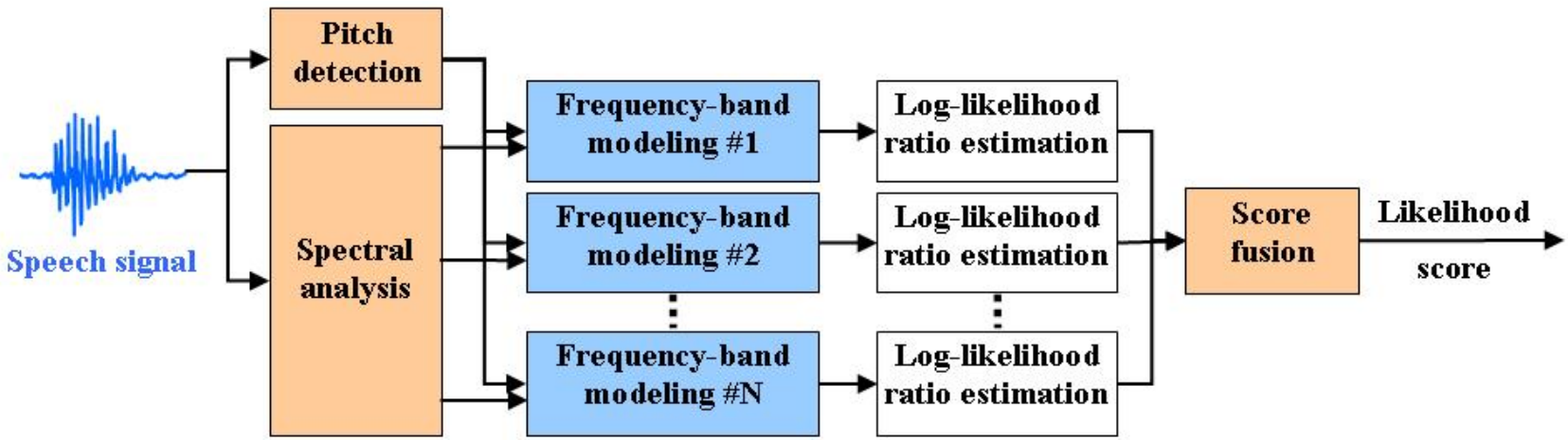


Eval05 Train-8c / Test-1c



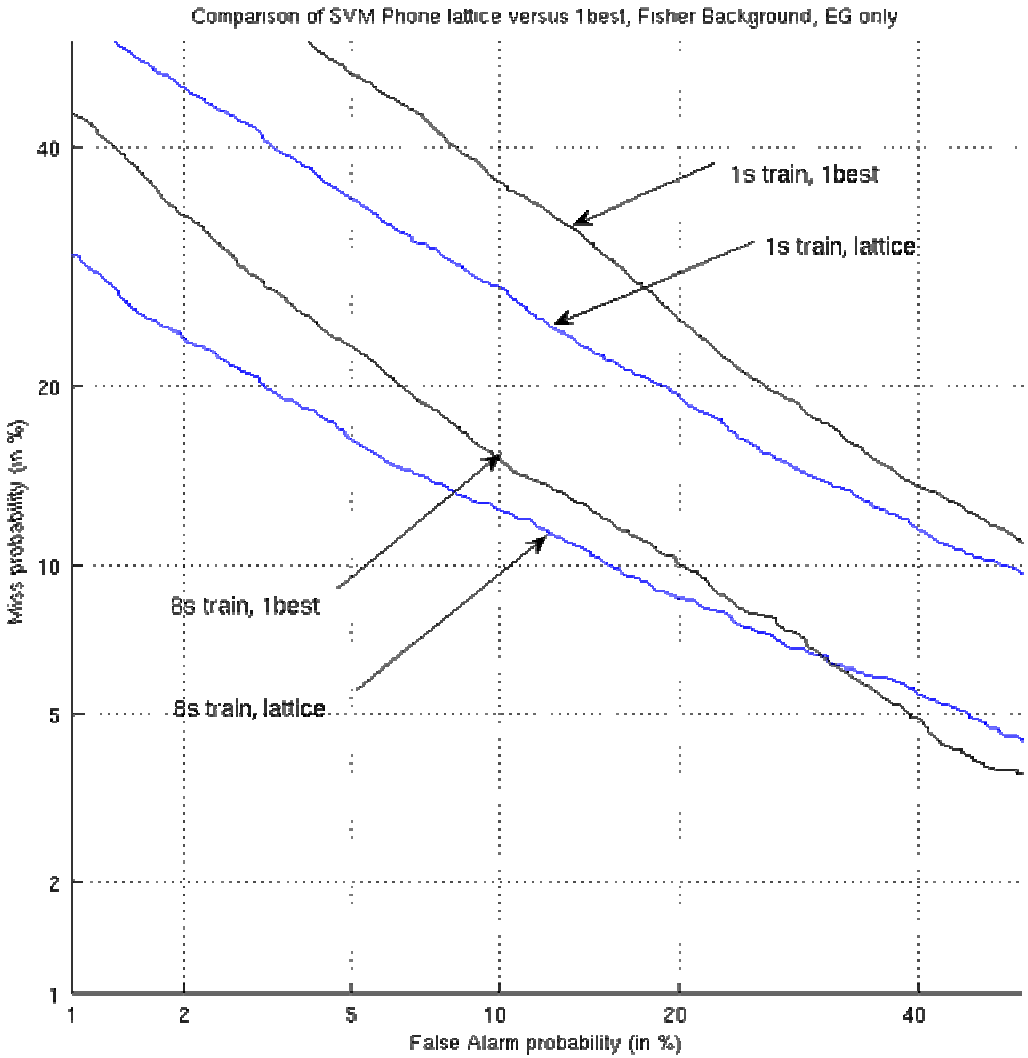
Sub-band Prosodic Modeling

- Aim: capture frequency-localized energy and pitch correlations characteristics
- Approach: apply n-gram modeling to the sequence of symbols estimated from the sub-band frequency energy and F0
- Description
  - Critical-band energy instead of short-term energy
    - 15 Bark-scale critical bands (1-Bark spacing between filters)
  - Log likelihood ratio between the target-speaker and the background bigram models per band
  - Frequency-bands fusion using linear combination
  - Background model trained using (Switchboard 2 (phases 1, 2, 4, and 5) corpora and the NIST SRE'04)
  - T-norm (gender and training condition dependent) applied using speakers from Background model

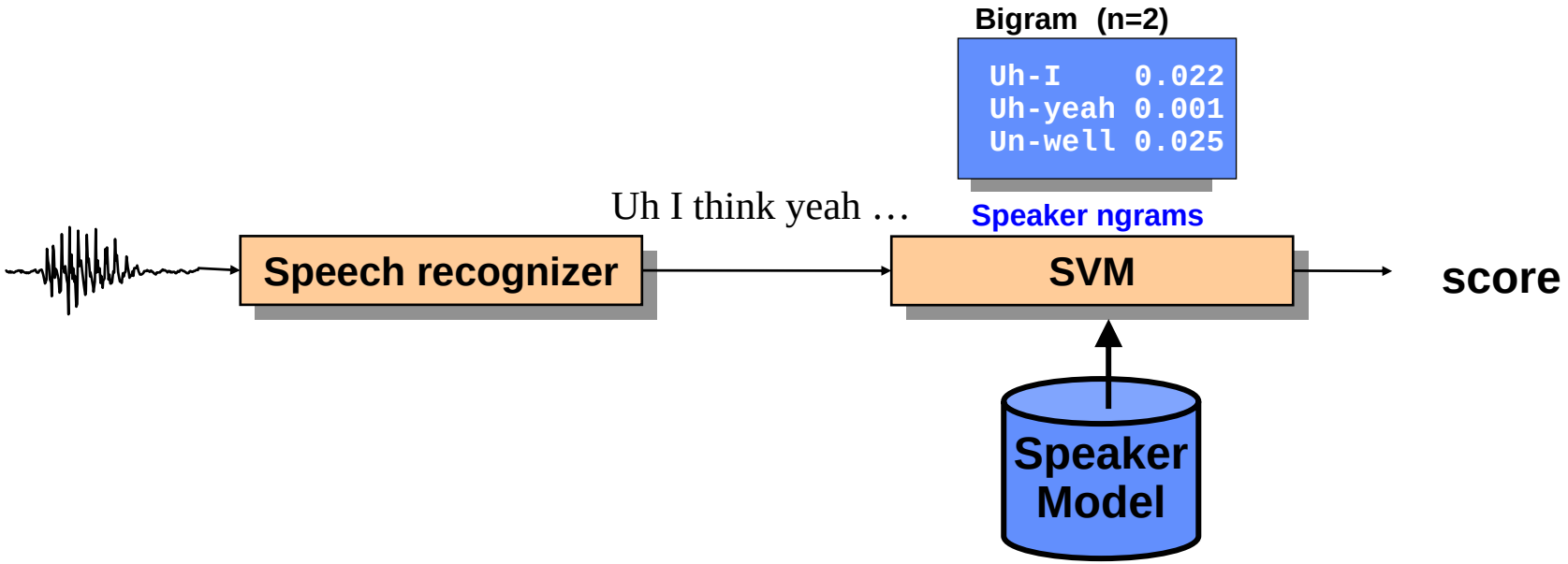


SVM Phone

Comparison: Lattice vs. 1best, English Only



SVM Word System

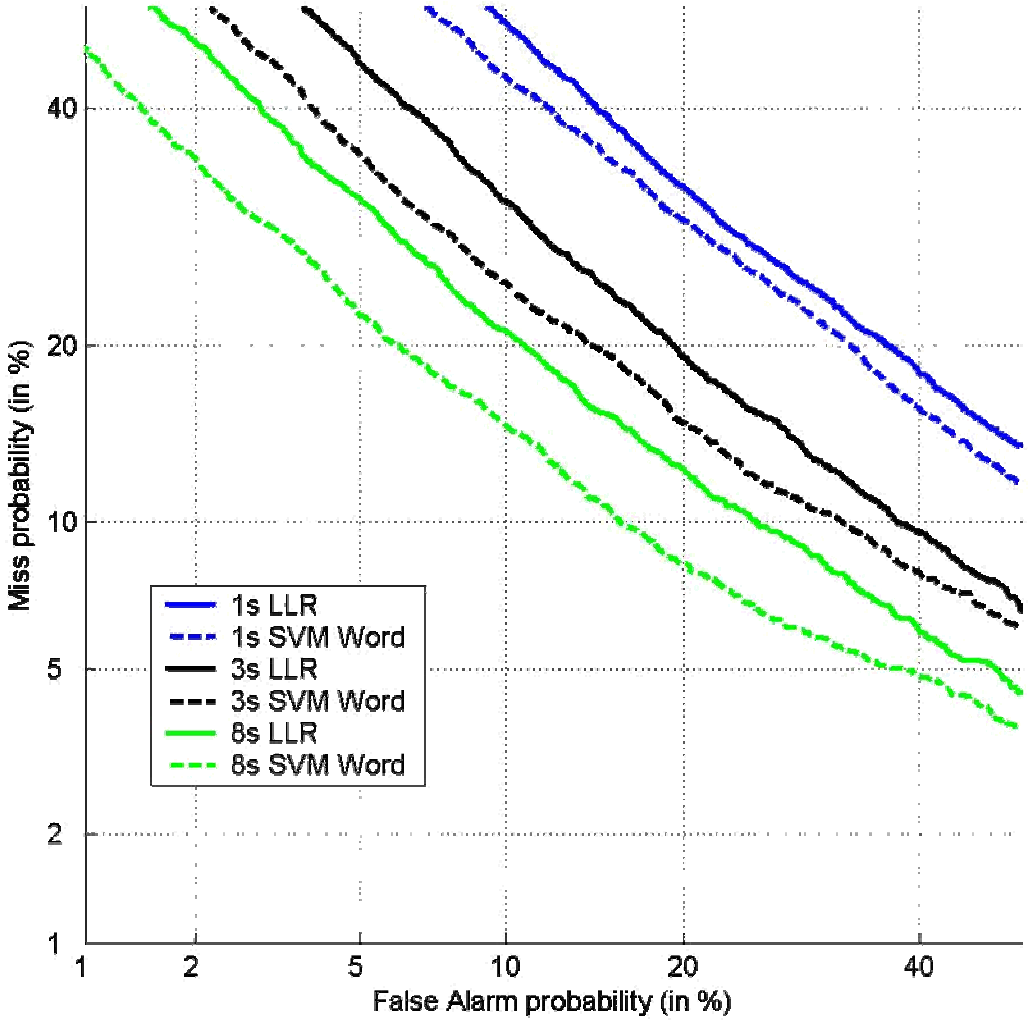


- Uses SVM Word system described in Campbell, et. al.
- Uses lattices and expected counts from Byblos ASR (same as LLR system)
- Unigram and Bigram Probabilities used
- Sparse Vector inner products used
- Weighting was similar to TFLLR; used a log squashing function instead of a square root
- Background trained from Fisher and CallFriend

W.Campbell, J.Campbell, D.Reynolds, D.Jones, T.Leek, "High Level Speaker Recognition with Support Vector Machines," ICASSP 2004.

SVM Word & LLR Word

Comparison: All Trials

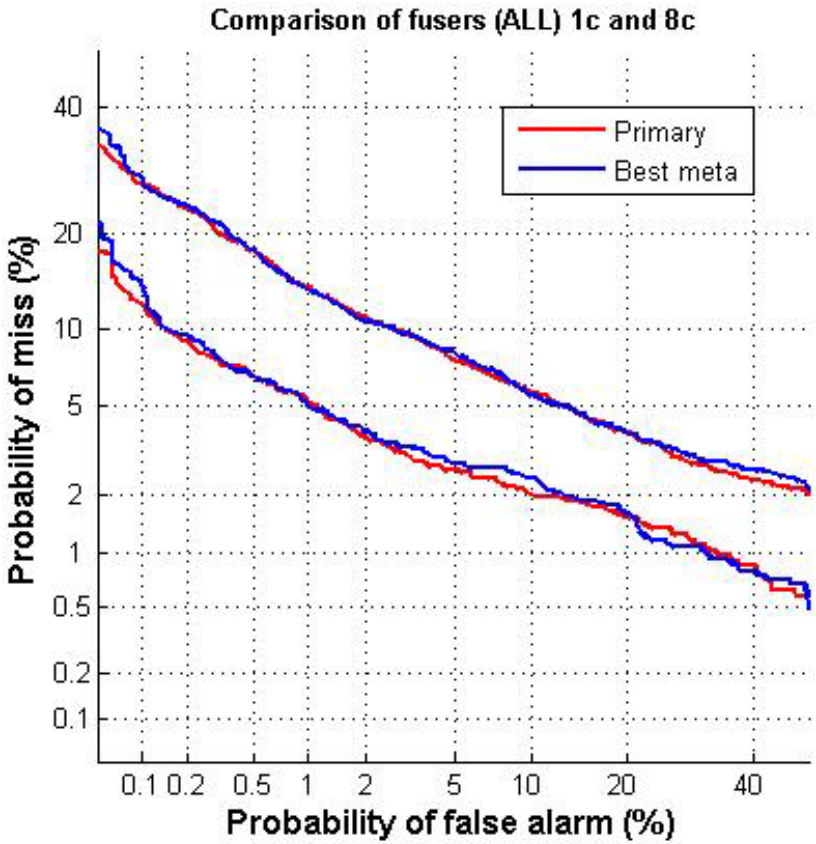


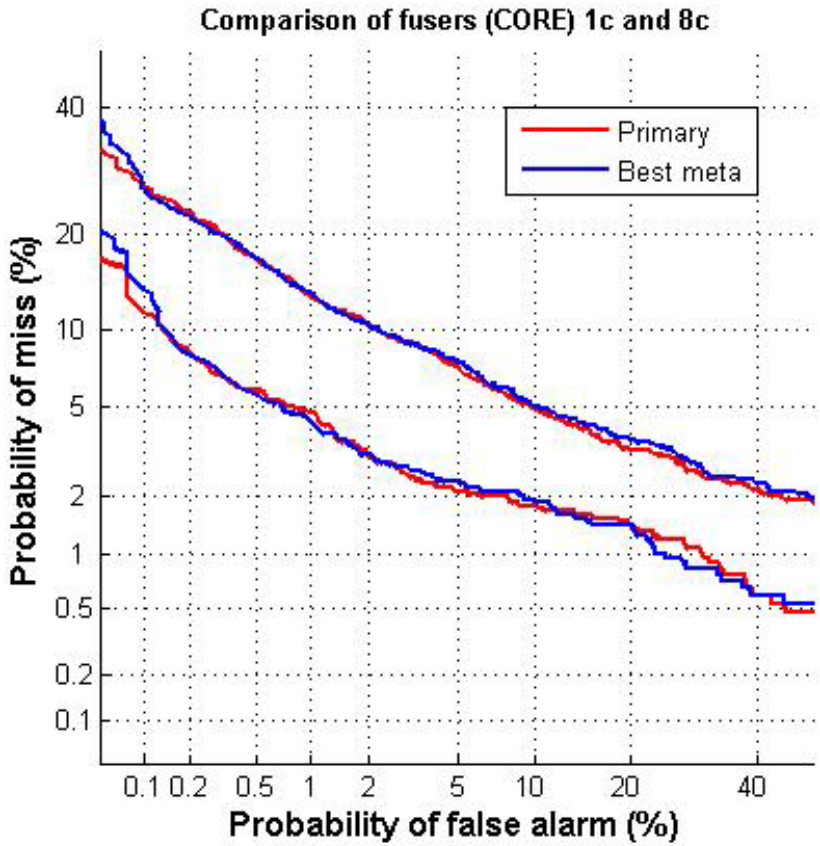
Fusion System Design

- **Metadata**
  - **Channel type (Landline/Cell)**
    - **Test: Encoded as single field (Land=0, Cell=1)**
    - **Model: Fraction of number of training files labeled cell**
  - **Duration (number of frames)**
    - **Test: Number of frames**
    - **Model: Total number of frames in all training files**
  - **Gender (Male/female)**
    - **Test: Male=0, Female=1**
    - **Model: Male/Female per index lists**
  - **Language (Arabic/English/Mandarin/Russian/Spanish)**
    - **Test: Encoded as five 1/0 fields**
    - **Model: Five fields represent fraction of number training files**

- **Development experiments show improvements on actual DCF for the language metadata**

Post Eval results





- **Metadata does not improve DET curve but improves actual DCF**
  - **Channel, gender, duration and language**

ALL	**1c 0.024120 / .023545**	**8c**	**0.022971 / 0.022378**
CORE	**1c 0.015458 / .012740**	**8c**	**0.014102 / 0.011362**