Natural Language Systems Group
IBM Research

NIST 2010 Speaker Recognition Evaluation

Jason Pelecanos **Mohamed K. Omar**

{jwpeleca, mkomar}@us.ibm.com

Natural Language Systems Group
IBM T. J. Watson Research Center
Yorktown Heights, New York

Roadmap

- Introduction
- Discriminatively Trained System
- Phonetically Inspired System
- Factor Analysis System
- Fusion Analysis
- Conclusions

Introduction

Introduction

- The IBM primary submission is a score-level linear combination of 6 core systems as follows:
  - 1 Phonetically inspired NAP-GMM system
  - 3 Discriminatively trained NAP-GMM systems
    - Varied by choice of front-end features, data subsets and system configurations.
  - 2 Factor analysis based systems
    - Trained on different frontend features: LPCCs and MFCCs
- The IBM alternate submission (a single system):
  - 1 Phonetically inspired NAP-GMM system (as above)

Discriminatively Trained System

Discriminative Regularization of Maximum Likelihood Estimation

- Estimating the UBM parameters using ML training is not related to reducing the speaker verification errors.
- We add a regularization term to the ML objective function to reduce the value of the imposter scores and increase the value of the target scores.
- The UBM parameters are updated using an EM-like algorithm.

The Objective Function

$$O = L - \lambda_t \sum_{v=1}^{T} e^{a_t \ominus b_t s_{tv}} - \lambda_i \sum_{j=1}^{J} e^{a_i \oplus b_i s_{ij}},$$

To increase the target scores and reduce the imposter scores

λ_t is the target regularization parameter,
λ_i is the imposter regularization parameter,
a_t, b_t are the parameters of the target function,
a_i, b_i are the parameters of the imposter function,
T is the number of target scores,
and J is the number of imposter scores.

Discriminative Regularization System & Data

System

- **Classifier:**
  - Dot-Product scoring of the GMM-based supervector representation of the enrollment and verification utterances developed from ASR front-end features.
- **Session Variability Compensation:**
  - NAP and ZT score normalization.

Data

- **UBM Training Database:**
  - The NIST08 interview development data. (6 speakers, 6 sessions, 9 mics)
  - For NIST 2010 systems, also includes the NIST 2008 evaluation data.
- **NAP Training Database:**
  - 11400 conv-sides from SWB2, Dev08, NIST04, and NIST06 databases.
  - For NIST 2010 systems, also includes the NIST 2008 evaluation data.
- **ZT score normalization Database:**
  - Same as NAP but divided into 2 gender-dependent subsets.

Interview Only Tasks of NIST08 – Discriminative Regularization

Table: (Norm. Min. DCF, EER (%)) for Baseline and Discriminative Regularization (DR) Systems

System	Int-Int-All	Int-Int-S	Int-Int-D
Baseline	(0.194, 4.2)	(0.029, 0.91)	(0.193, 4.1)
DR	(0.159, 2.7)	(0.019, 0.64)	(0.161, 2.8)

Task Description:

Int-Int-All: Interview speech in training and test.

Int-Int-S: Interview speech from the same (lapel) microphone in training and test.

Int-Int-D: Interview speech from different microphones in training and test.

Telephone Tasks of NIST08 - Discriminative Regularization

Table: (Norm. Min. DCF, EER (%)) for Baseline and Discriminative Regularization (DR) Systems

System	Int-Tel	Tel-Mic	Tel-Eng	Tel-US
Baseline	(0.375, 7.9)	(0.323, 7.7)	(0.156, 3.5)	(0.164, 4.1)
DR	(0.329, 7.0)	(0.256, 7.2)	(0.140, 3.4)	(0.141, 4.1)

Task Description:

Int-Tel: Interview speech in training & telephone speech in test.

Tel-Mic: Telephone speech in training and telephone microphone speech in test.

Tel-Eng: English language telephone speech in training and test (any variety).

Tel-US: English language telephone speech spoken by a native US English Speaker in training and test.

Interview Tasks of NIST10 – Discriminative Regularization

Table: (Min. DCF, EER (%)) for Baseline and Discriminative Regularization (DR) Systems

System	Int-S	Int-D	Int-NTel	Int-NMic
Baseline	(0.39, 3.4)	(0.52, 5.1)	(0.38, 4.1)	(0.45, 3.4)
DR	(0.37, 3.1)	(0.43, 4.7)	(0.32, 3.4)	(0.42, 3.2)

Task Description:

Int-S: Interview speech (same-microphone) in training and test.

Int-D: Interview speech (different-microphones) in training and test.

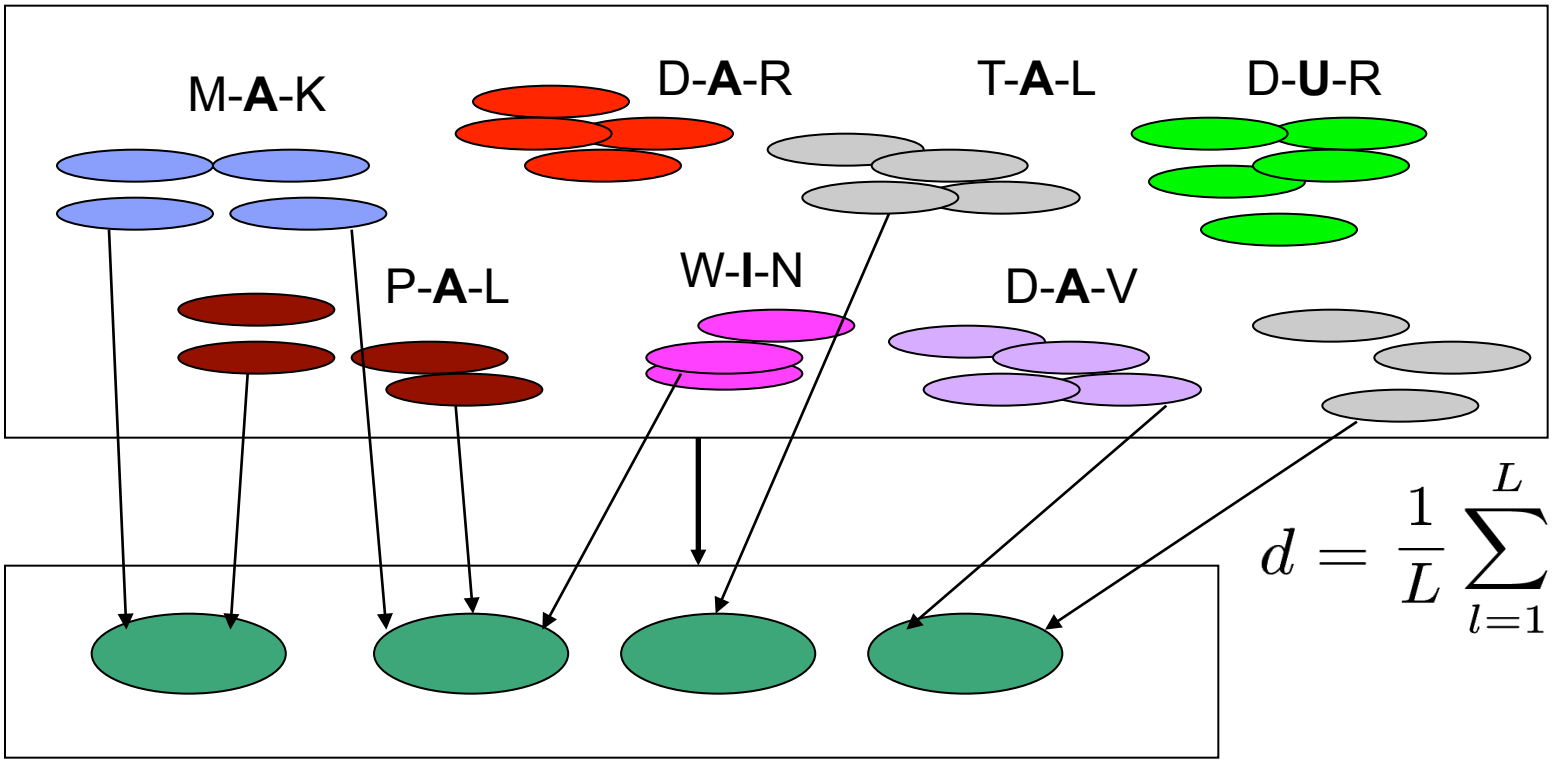
Int-NTel: Interview speech in training and normal vocal effort telephone speech in test.

Int-NMic: Interview speech in training and normal vocal effort microphone telephone speech in test.

Phonetically Inspired System

UBM Construction From ASR Models

L Gaussian components $G = (g_1, g_2, \ldots, g_L)$.



$$d = \frac{1}{L} \sum_{l=1}^{L} \min_{k=1}^{K} d(u_k, g_l)$$

K Gaussian components, $U = (u_1, u_2, \ldots, u_K)$.

$$d(u_k, g_l) = KL(u_k, g_l) + KL(g_l, u_k),$$

and $KL(u_k, g_l)$ is the KL divergence between u_k and g_l.

PIUBM System

System

- **Classifier:**
  - Dot-Product scoring of the GMM-based supervector representation of the enrollment and verification utterances based on ASR features.
- **Session Variability Compensation:**
  - NAP and ZT score normalization.
- **ASR System:**
  - ~2000 hours of SWB and Fisher databases.
  - 13 PLP features compensated by CMVN+VTLN; 9-frame window + LDA [X] 40 dimensions.
  - Speaker adaptation using FMLLR; FMPE transform estimated in the FMLLR feature space.

Data

- **NAP Training Database:**
  - 11400 conv-sides from SWB2, Dev08, NIST04, and NIST06 databases.
  - For NIST 2010 system, also includes the NIST 2008 evaluation data.
- **ZT score normalization Database:**
  - Same as NAP but divided into 2 gender-dependent subsets.

Telephone Tasks of NIST08 - System Comparison

Table: (Norm. Min. DCF, EER (%)) for MFCC & ASR-frontend baselines, and PIUBM System

System	Int-Tel	Tel-Mic	Tel-Eng	Tel-US
Baseline	(0.375, 10.3)	(0.288, 7.4)	(0.156, 3.5)	(0.154, 4.4)
ASR frontend	(0.318, 7.6)	(0.218, 6.4)	(0.164, 3.4)	(0.154, 4.4)
PIUBM	(0.307, 8.6)	(0.221, 6.7)	(0.127, 2.7)	(0.116, 3.0)

Task Description:

Int-Tel: Interview speech in training & telephone speech in test.

Tel-Mic: Telephone speech in training and telephone microphone speech in test.

Tel-Eng: English language telephone speech in training and test (any variety of English).

Tel-US: English language telephone speech spoken by a native US English Speaker in training and test.

Telephone Tasks of NIST10 - System Comparison

Table: (Min. DCF, EER (%)) for Baseline and PIUBM Systems

System	NTel-NTel	NTel-HTel	NMic-HTel	NTel-LTel	NMic-LTel
Baseline	(0.67, 4.0)	(0.57, 2.9)	(0.43, 4.8)	(0.19, 1.1)	(0.28, 1.5)
PIUBM	(0.31, 2.9)	(0.58, 2.9)	(0.34, 5.8)	(0.25, 1.0)	(0.23, 1.5)

Task Description:

NTel-NTel: Normal vocal effort telephone speech in training & test.

NTel-HTel: Normal vocal effort telephone speech in training and high vocal effort telephone speech in test.

NMic-HTel: Normal vocal effort microphone telephone speech in training and high vocal effort telephone speech in test.

NTel-LTel: Normal vocal effort telephone speech in training and low vocal effort telephone speech in test.

NMic-LTel: Normal vocal effort microphone telephone speech in training and low vocal effort telephone speech in test.

Factor Analysis System

Factor Analysis

System

- Factor analysis point estimates used
- Log-likelihood ratio approximation to scoring
- ZT-Norm Applied + Symmetric scoring

- LPCC and MFCC features + feature warping

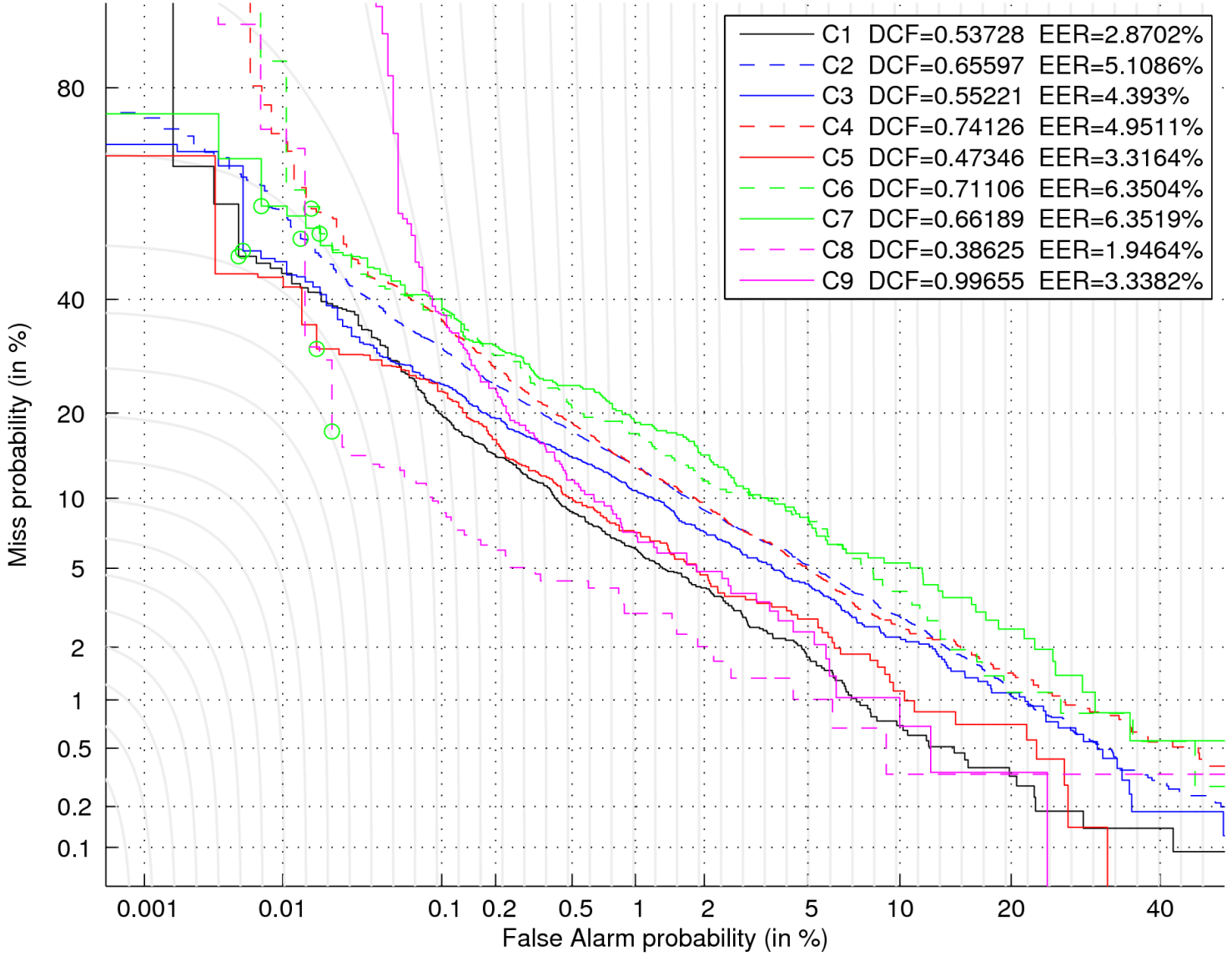
Data

- **FA Training Database:**
  - 11400 conv-sides from SWB2, Dev08, NIST04, and NIST06 databases.
  - For NIST 2010 systems, also includes the NIST 2008 evaluation data.
- **ZT score normalization Database:**
  - Same as NAP but divided into 2 gender-dependent subsets.

Factor Analysis – NIST 2008 Results

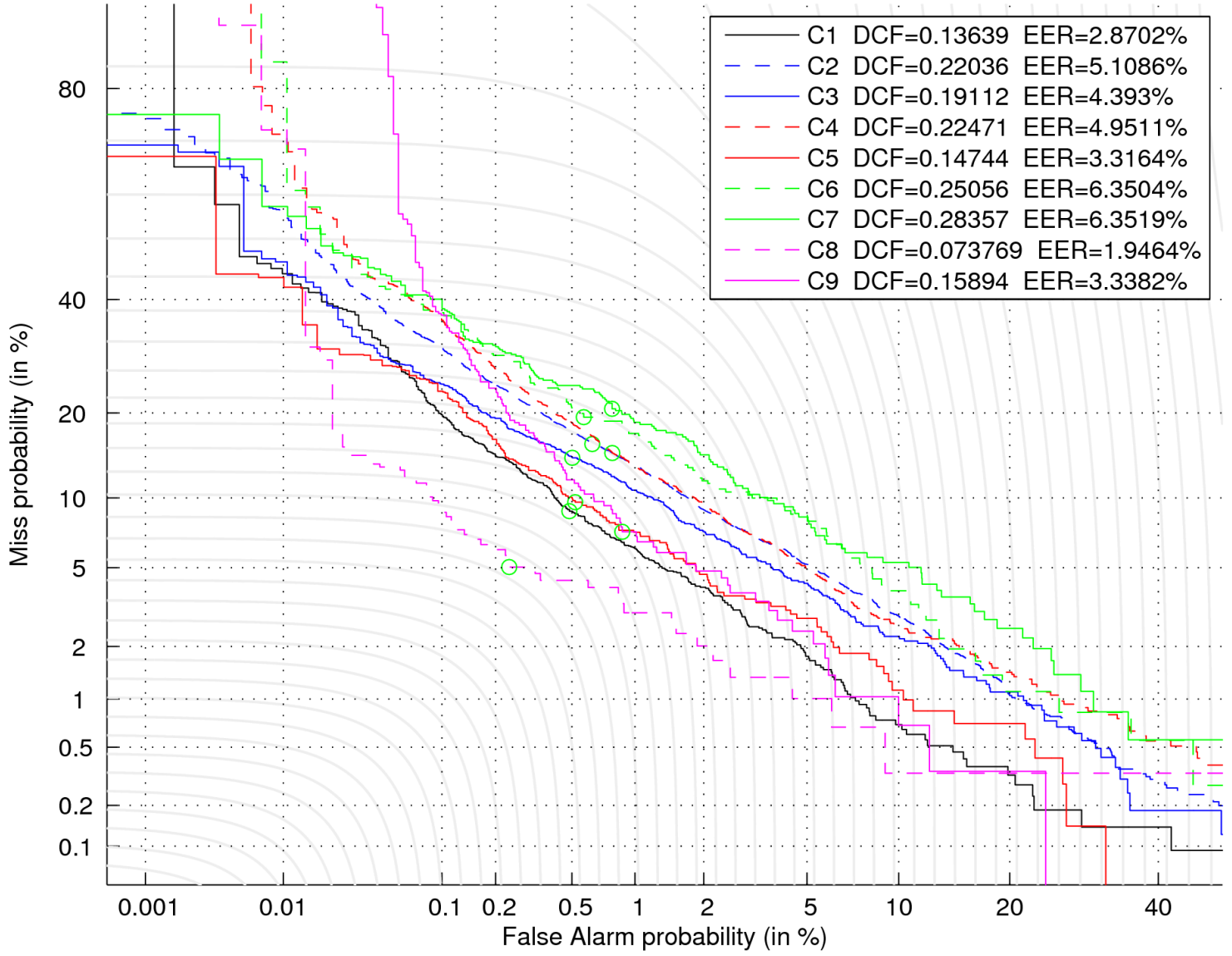
Task	LPCC (norm mDCF)	MFCC (norm mDCF)	Equal Weight Linear Fusion (norm mDCF)
1	0.098	0.143	**0.077**
2	0.019	0.029	**0.017**
3	0.101	0.149	**0.079**
4	0.223	0.225	**0.171**
5	0.175	0.166	**0.128**
6	0.326	0.306	**0.303**
7	0.114	**0.079**	0.080
8	0.125	0.093	**0.088**

Factor Analysis – MFCC system (NIST 2010 Result)

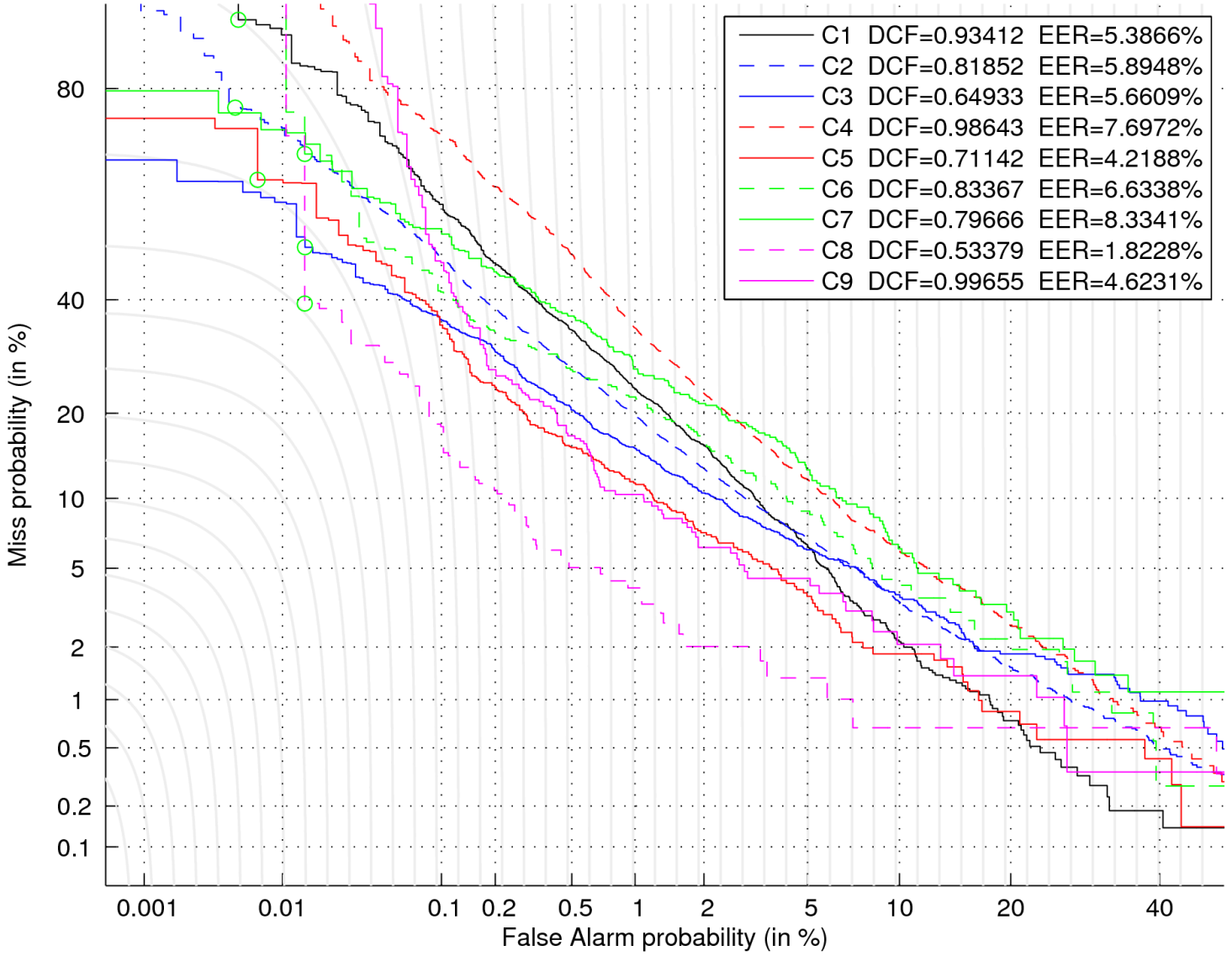


Factor Analysis – MFCC system (NIST 2010 Result*)

*2010 Result using NIST 2008 min. DCF



Factor Analysis – LPCC system (NIST 2010 Result)



Fusion Analysis

Data Fusion – NIST 2010 System Breakdown (min. DCF)

Task	Primary	Alternate	Best Single (of 6 systems)
1	0.498	0.398	**0.371** (D10)
2	0.463	0.521	**0.432** (D10)
3	**0.311**	0.383	0.325 (D10)
4	0.539	**0.401**	**0.401** (Alt)
5	**0.279**	0.312	0.312 (Alt)
6	**0.528**	0.582	0.572 (D10)
7	0.422	**0.342**	**0.342** (Alt)
8	0.259	0.258	**0.198** (D10)
9	0.683	**0.237**	**0.237** (Alt)

- D10 -
Discriminatively Trained, 10 iteration system
- Alt -
IBM's alternate system submission

Conclusions

Conclusions

- Discriminatively trained and Phonetically Inspired UBMs provided noteworthy contributions to overall system performance for NIST 2010.
- LPCC+MFCC Factor Analysis System components performed extremely well on NIST 2008 data. Performance did not carry across to the NIST 2010 data:
  - Suggests a possible implementation issue... additional follow-up required here
- The data set developed for fusion optimization and general development was not representative of the 2010 tasks.
  - A mixed bag of fusion results across the 9 conditions.