ABC and CRIM

AGNITIO, BUT, CRIM

SRE Worskshop, 24 June 2010

Outline

ABC = AGNITIO + BUT + CRIM

The ABC submission is a collaboration between:

- Agnitio Labs, South Africa
- Brno University of Technology, Czech Republic
- CRIM, Canada

(In alphabetical order.)

Contributors

- **AGNITIO** Niko Brümmer, Luis Buera, Edward de Villiers
- **BUT** Ondřej Glembek, Pavel Matějka, Lukáš Burget, Doris Baum, Marcel Kockmann, Oldřich Plchot, Valiantsina Hubeika, Martin Karafiát
- **CRIM** Patrick Kenny, Pierre Ouellet, Gilles Boulianne, Mohammed Senoussaoui

(Presenters are highlighted.)

ABC Collaboration Goals

We tried to:

- survive the new DCF.
- use some new i-vector solutions.
- improve MLLR and prosodic recognizers.
- derive benefit from quality measures.
- ignore vocal effort variation.

Challenges induced by the new DCF

- We had to redefine our own development trial indices to maximize the number of non-target trials in our development database, rather than just re-using SRE 2008 trial lists.
- Duplicate PIN errors in SRE'08 tel-tel answer key caused false false-alarms. We will defer this issue to the discussion session later.

Submissions

- ABC submitted a fusion of multiple sub-systems for the core-core task, some analysis of which follows.
- CRIM also made their own submission for non-core tasks, which will be presented by Patrick Kenny.

Results Analysis

- We analyse some of our results, to examine calibration, fusion and quality measures.
- We analyse only conditions 1–5, since we did no special development for vocal effort variation. If we got good results there, those are accidental.
- We analyse results only for the extended evaluation, the main evaluation results having been shown already by NIST.

Calibration Goal

- We pursued log-likelihood-ratio calibration, rather than point calibration.
- We optimized our calibration transformation to minimize cross-entropy, rather than just setting a decision threshold.
- The cross-entropy was biased with prior = 0.001, to focus on an area centred around the new DCF.

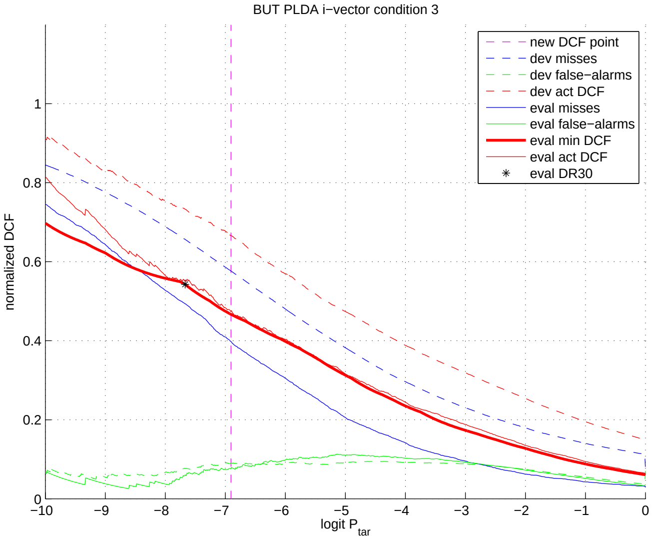
Calibration Analysis

We use the normalized DCF curve to analyse development and evaluation calibration:

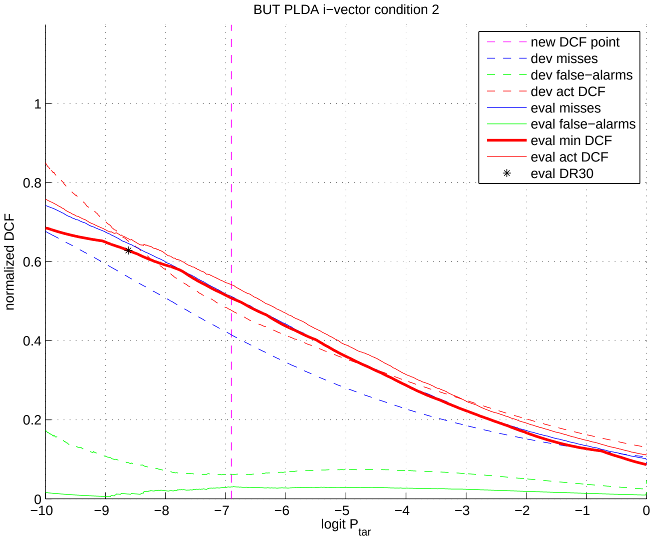
- Y-axis: Normalized minimum and actual DCF against the operating point.
- X-axis: The operating point is parametrized by the prior.

Examples follow.

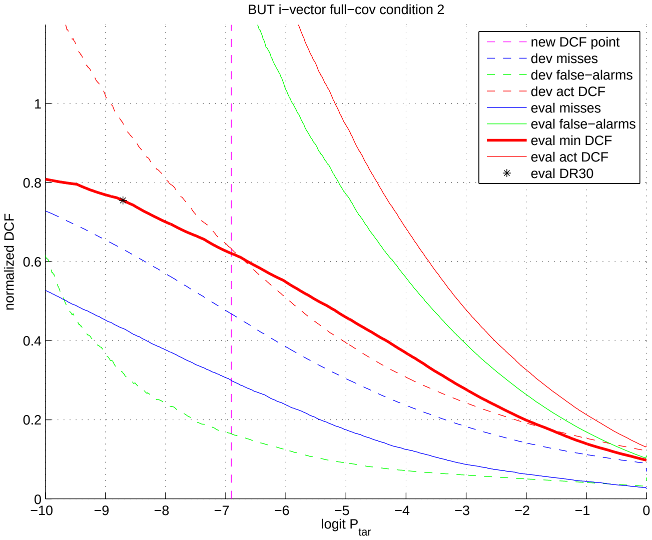
Normalized DCF: Example of Excellent Calibration



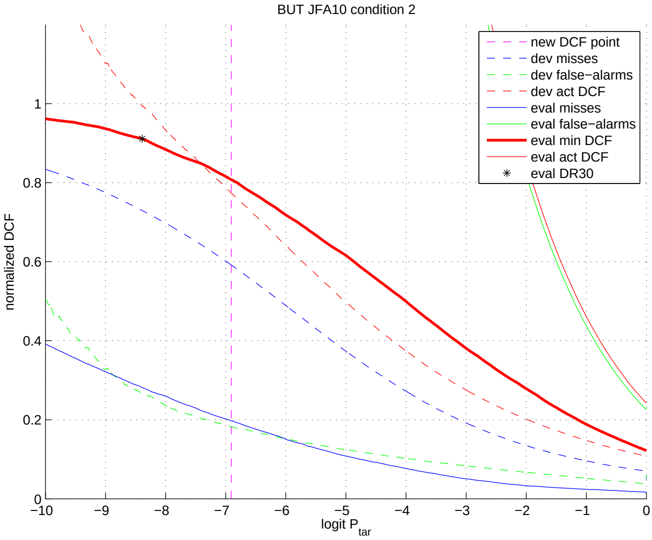
Normalized DCF Curve: Example of Good Calibration



Normalized DCF Curve: Example of Bad Calibration



Normalized DCF Curve: Example of Worse Calibration



Calibration Bottom Line

We had mixed success:

- Calibration failed for tel-tel, but we fixed it post-eval.
- Calibration was OK for int-tel.
- Calibration failed for int-int and int-auxmic. And we still can't explain or fix it.

(See printed notes for detailed DCF numbers.)

ABC-1 Calibration

Condition 5: Tel-Tel

	System	act DCF	min DCF
1	ABC-1	1.73	0.32
2	sub-system fixed	0.62	0.31
3	alt. dev. key	0.51	0.30
4	alt. dev. key & alt. fusion	0.36	0.30

1. Had a broken sub-system.
2. Broken sub-system fixed.
3. As 2, but corrected some ill-advised development trial index pruning.
4. As 3, but also replaced non-linear s-cal fusion with plain linear fusion.

ABC-1 Calibration

Conditions 1–4: Involving Microphones

- **Conditions 1,2,4:** Only one sub-system, the un-normalized PLDA got act. norm DCF < 1 and indeed, this system had very good calibration.
- **Condition 3:** All sub-systems and all fusions got mediocre to good calibration.

At present, we can offer no explanations, except for the ...

Un-normalized PLDA

Robust against calibration mismatch?

The ‘BUT PLDA’ sub-system used:

- The same development data as all other ABC sub-systems.
- Same i-vectors as one of the other BUT systems (which used cosine score).
- AGNITIO’s PLDA model training and scoring:
    - PLDA was Gaussian, not heavy-tailed like CRIM’s PLDA.
    - Improved on Prince’s PLDA training by including minimum-divergence[1].
- Some careful tuning by Lukas.
- No score normalization.

[1] http://niko.brummer.googlepages.com/EMandMINDIV.pdf

Minimum Divergence
Patrick's Envelope Explanation

data D

hidden H

$P(D,H)$

posterior $Q(H)$

$$\left\{ \begin{array}{l} P(H) \\ Q(H) \end{array} \right.$$

$$\mathcal{L} = E_Q\left[\ln \frac{P(D,H)}{Q(H)}\right]$$

$$= \underbrace{E_Q\left[\ln P(D|H)\right]}_{\text{EM auxiliary}} + E_Q\left[\ln \frac{P(H)}{Q(H)}\right]$$

$$- \mathrm{DIV}(Q(H) \| P(H))$$

Un-normalized PLDA

Robust against calibration mismatch?

The ‘BUT-PLDA’ sub-system got good calibration for all conditions despite being very similar to—and using the same resources as—the other i-vector systems built by AGNITIO, BUT and CRIM.

- Is this because it has no score normalization?
- Did minimum-divergence help to stabilize calibration?

Extended vs Main for ABC-1 Core-Core

DCF Details

condition	norm act dcf	norm min dcf	prbep	%eer
main 1	3.57	0.26	95.52	1.15
ext 1	9.29	0.22	331.92	1.03
main 2	0.67	0.37	543.65	1.98
ext 2	1.16	0.34	1 867.60	1.77
main 3	0.49	0.30	83.22	1.34
ext 3	0.39	0.27	362.31	1.74
main 4	0.80	0.49	288.32	3.05
ext 4	1.93	0.36	528.86	1.94
main 5	0.83	0.27	49.71	1.60
ext 5	1.73	0.32	628.19	1.90
main 6	0.67	0.49	46.22	1.98
ext 6	0.75	0.68	858.22	2.76
main 7	0.72	0.63	86.93	3.92
ext 7	1.08	0.65	109.00	3.98
main 8	0.12	0.12	10.25	0.76
ext 8	0.50	0.34	333.17	1.12
main 9	0.54	0.39	34.84	2.50
ext 9	0.89	0.20	33.00	1.73

Core-Core Extended vs Main

Counts of Models, Segments and Trials

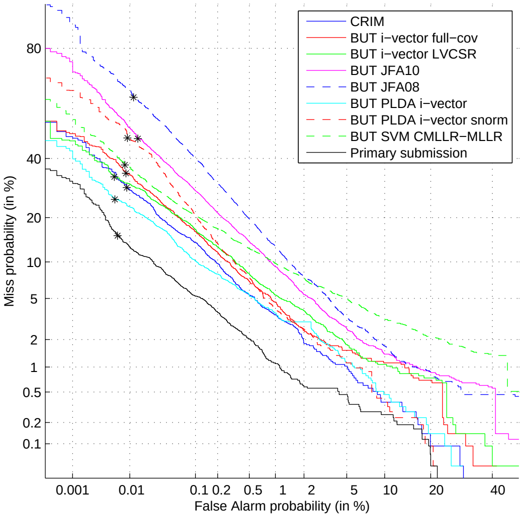
condition	male mods	male segs	male tar	male non	fem mods	fem segs	fem tar	fem non
main 1	990	991	989	28 114	1 169	1 170	1 163	32 598
main 2	990	2 974	3 463	98 282	1 169	3 516	4 072	114 025
main 3	750	239	837	26 178	859	285	796	30 232
main 4	731	432	1 225	39 166	789	407	1 141	44 370
main 5	290	355	353	13 707	290	357	355	15 958
main 6	181	147	178	12 825	184	185	183	15 486
main 7	180	149	179	12 786	180	185	180	15 211
main 8	119	116	119	10 997	181	184	179	17 309
main 9	117	115	117	10 697	176	181	173	16 533
ext 1	1 108	1 108	1 978	346 857	1 283	1 283	2 326	449 138
ext 2	1 108	3 328	6 932	1 215 586	1 283	3 858	8 152	1 573 948
ext 3	1 126	384	2 031	303 412	1 347	430	1 958	334 438
ext 4	1 108	440	1 886	364 308	1 283	409	1 751	392 467
ext 5	1 906	388	3 465	175 873	2 361	379	3 704	233 077
ext 6	2 096	181	1 816	191 784	2 598	210	2 321	269 654
ext 7	219	183	179	39 898	203	211	180	42 653
ext 8	2 096	137	1 447	144 982	2 598	205	2 374	259 866
ext 9	219	136	117	29 667	203	202	173	40 833

Fusion Analysis

Ignoring calibration, our fusions worked well for all conditions.
Below, we analyse our primary fusions for conditions 1–5:

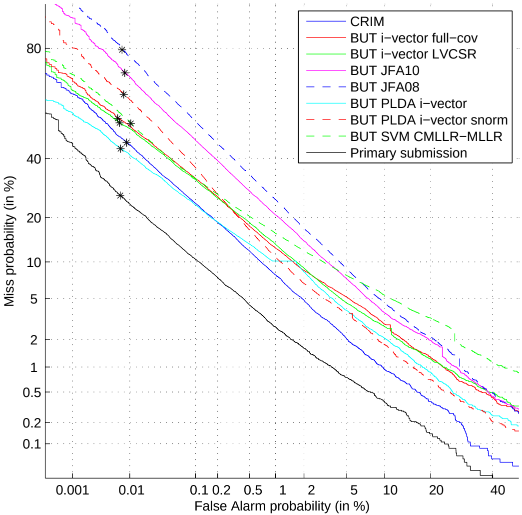
- We use DET-curves to ignore calibration.
- We show the primary fusions, compared to the sub-systems that were fused.
- For conditions 1–4, these fusions included quality measures.

ABC-1 Extended Core-Core Condition 1



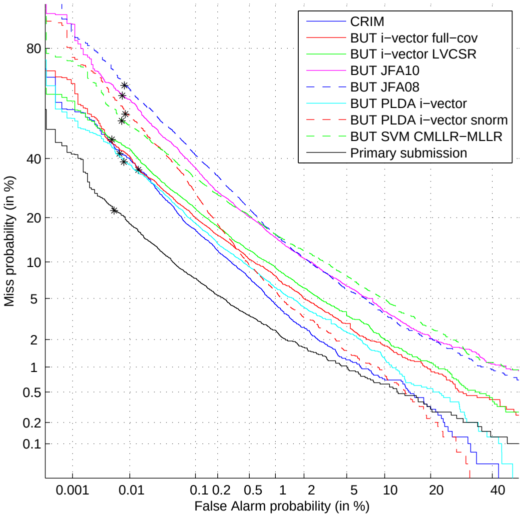
int-int, same microphone

ABC-1 Extended Core-Core Condition 2



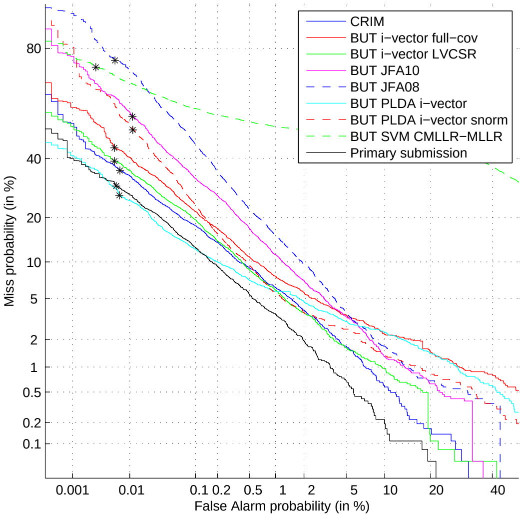
int-int, different microphone

ABC-1 Extended Core-Core Condition 3



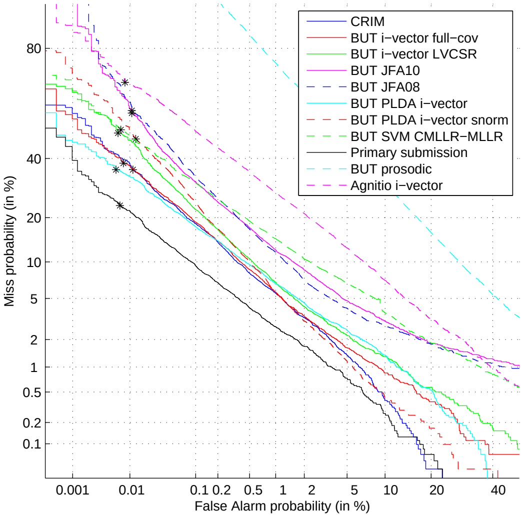
int-tel

ABC-1 Extended Core-Core Condition 4



int-auxmic

ABC-1 Extended Core-Core Condition 5



tel-tel, different number

Quality Measures

Our quality measures, computed for every test and every train segment, included:

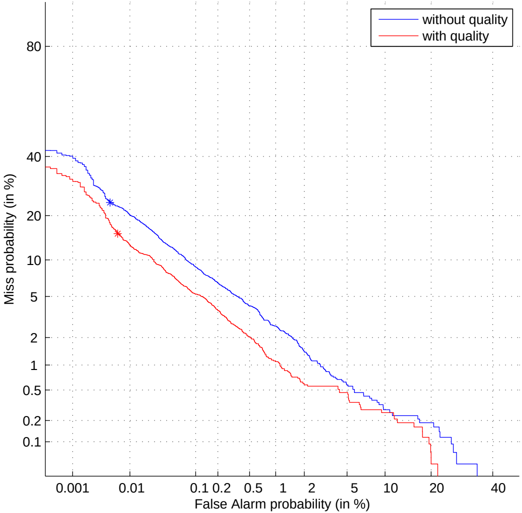
- log number of frames
- gender recognizer score
- SNR
- speech vs silence detector score

Quality Measures
Results

Ignoring calibration, quality measures contributed to better discrimination in all conditions (1–4) involving microphones, but was not helpful for tel-tel.

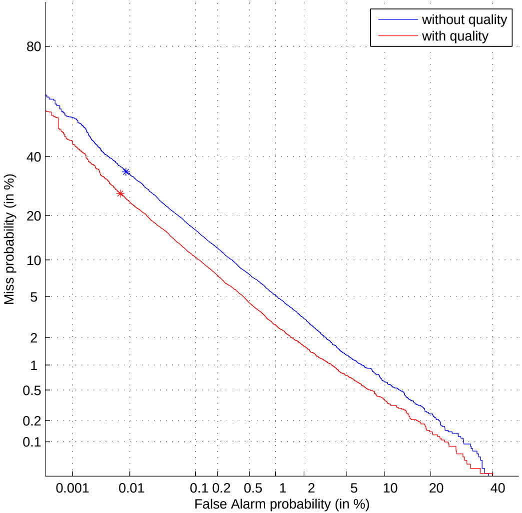
- We use DET-curves to ignore calibration.
- We compare fusions, with and without quality measures.

Fusion with Quality for Ext. Core-Core Condition 1



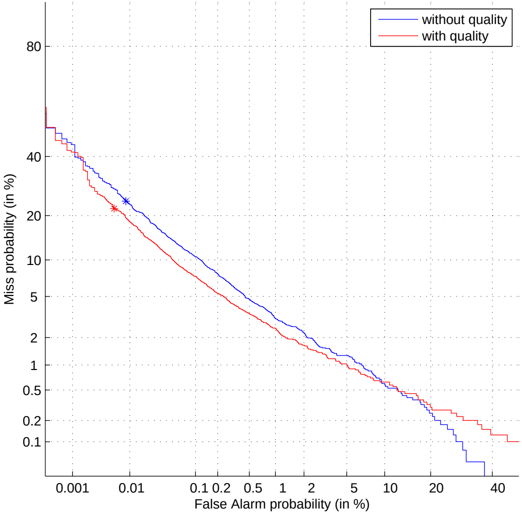
int-int, same microphone

Fusion with Quality for Ext. Core-Core Condition 2



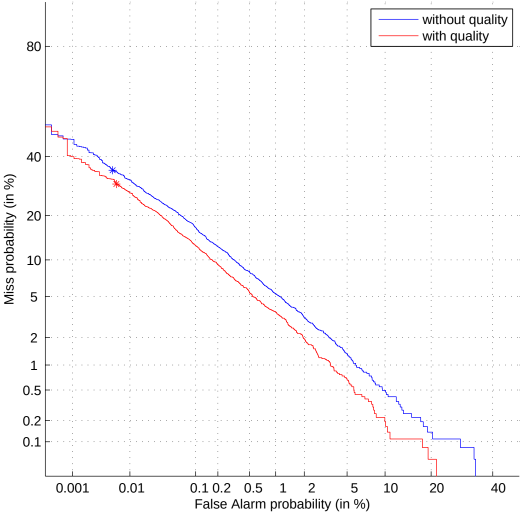
int-int, different microphone

Fusion with Quality for Ext. Core-Core Condition 3



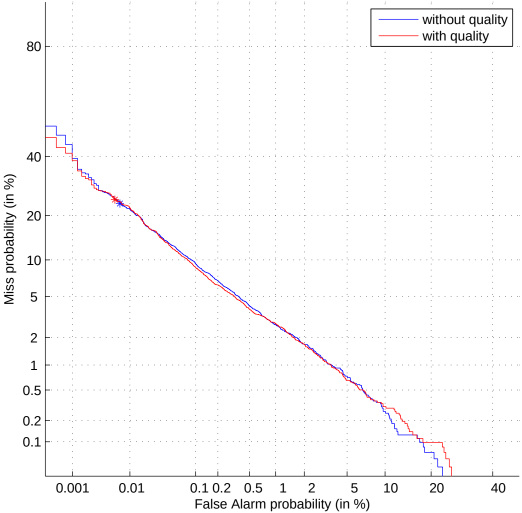
int-tel

Fusion with Quality for Ext. Core-Core Condition 4



int-auxmic

Fusion with Quality for Ext. Core-Core Condition 5



tel-tel, different number

AGNITIO's Conclusion

- There is life after JFA:
  - we improved on the 2008 state-of-the-art
  - i-vectors contributed significantly
- Fusion helped.
- Quality measures helped (a first for us).
- Farewell score normalization?

AGNITIO's Conclusion

- The new DCF is difficult, but do-able. It forced most of us—participants and evaluator—well outside of our comfort zones, but I think it was a worthwhile exercise.

JFA Systems

UBM

- GD, 2048G, Diag Cov, NO Varflooring applied

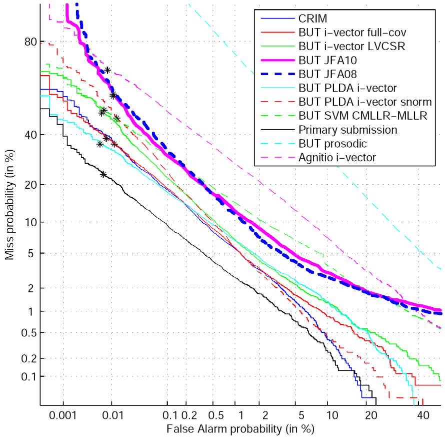
JFA Systems

JFA08 $V = 300$, $U_{\text{tel}} = 100$, $U_{\text{mic}} = 100$, $U_{\text{int}} = 20$,
$\mathbf{U}_{\text{allcond}} = (\mathbf{U}_{\text{tel}}\mathbf{U}_{\text{mic}}\mathbf{U}_{\text{int}})$

JFA10 $V = 300$, $U_{\text{tel}} = 100$, $U_{\text{mic}} = 100$, $U_{\text{int}} = 50$,
$\mathbf{U}_{\text{tel}-\text{tel}} = (\mathbf{U}_{\text{tel}}\mathbf{U}_{\text{mic}})$
$\mathbf{U}_{\text{int}-\text{tel},\text{int}-\text{int}} = (\mathbf{U}_{\text{tel}}\mathbf{U}_{\text{mic}}\mathbf{U}_{\text{int}})$

- Linear scoring was used
- ZT-norm score normalization was applied in both systems

JFA Systems - Extended Core-Core Cond 5



tel-tel, different number

I-vector LDA+WCCN

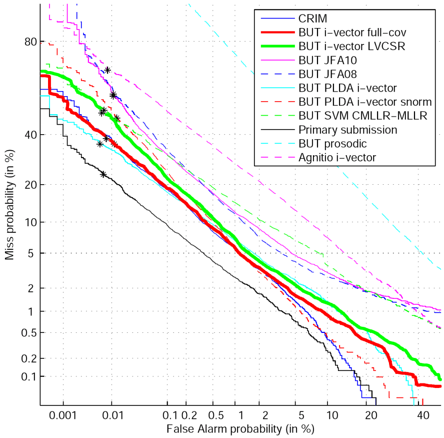
UBMs

- GI, 2048G, FullCov, VarFlooring applied
- GI, 2048G, LVCSR - Clustered phoneme GMMs

I-vector Extractor

- GD I-vector extractors trained on 1400 and 1000 hours of speech for females and males, respectively.
- Dimensionality reductors $2048 \times 60 = 122880 \rightarrow 400$
- Adopted Najim Dehak's concept:
  - Unwanted variability reduction using LDA+WCCN $400 \rightarrow 200$
  - Score computed as cosine distance
- Simplified S-norm score normalization applied

I-vector LDA+WCCN - Extended Core-Core Cond 5



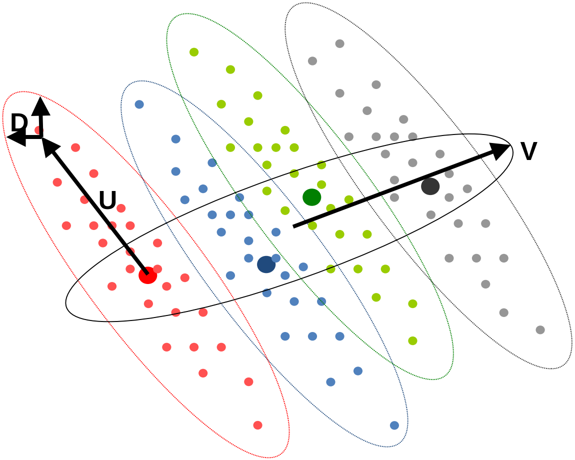
tel-tel, different number

I-vector PLDA

- Simplified version of Joint Factor Analysis (JFA) introduced for face verification (Prince ’07)
- LDA-like assumptions
    - Gaussian-distributed data
    - Gaussian-distributed data within each class
    - Shared within-class covariance matrix
    - Distributions pre-trained using large number of examples of speakers and conditions
- Modeling of variances makes use of sub-spaces, similarly to JFA.

$$\mathbf{o} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z}$$

I-vector PLDA



$$\mathbf{o} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \mathbf{D}\mathbf{z}$$

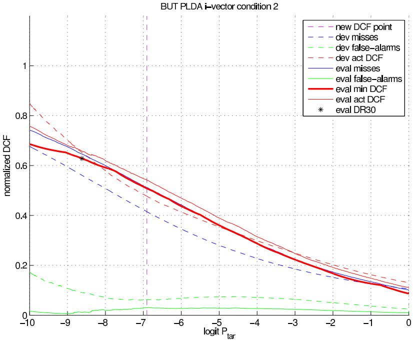
I-vector PLDA

- Simple probabilistic model allowing for fast symmetric scoring
  - Allows us to evaluate the probability of both segments in a trial being pronounced by the same speaker
  - Instead of the usual probability that the test segment is produced by model trained (or adapted) on the enrollment segment
  - Not suitable for modeling sequences (a segment has to be represented by a fixed-length vector)

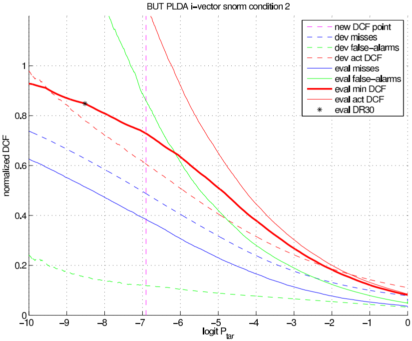
I-vector PLDA

- **tel-tel** – 90 eigenvoices, 400 eigenchannels (full rank). NO score normalization.
- **int-tel, int-int** – 90 eigenvoices and 1600 eigenchannels. After V and D are trained, 4 separate U (400) are trained: mic, tel, int, and all together. These are concatenated. NO score normalization.

I-vector PLDA - Extended Core-Core Cond 2

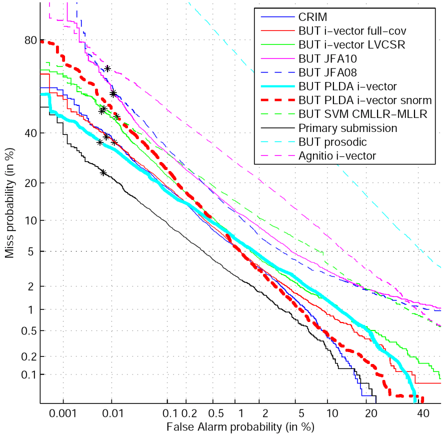


NO normalization



S-norm

I-vector PLDA - Extended Core-Core Cond 5

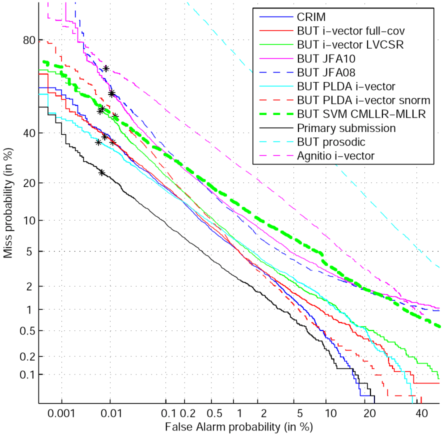


tel-tel, different number

SVM MLLR-CMLLR

- LVCSR - PLP12_0DAT, VTLN, HLDA, fMPE + MPE, xwrd triphones, WER 24% on NIST eval01 task
- CMMLR - 2 classes (speech, silence)
- MLLR - 3 classes (2 speech, 1 silence)
- The SVM input is a concatenation of vectorized $\text{CMLLR}_{\text{speech}}$, $\text{MLLR}_{\text{speech1,2}}$ matrices
- Rank normalization applied
- NAP
  - Trained on SRE04, SRE05
  - $U_{\text{tel}} = 20$, $U_{\text{mic}} = 10$, $U_{\text{int}} = 10$
    $\mathbf{U}_{\text{tel}-\text{tel}} = (\mathbf{U}_{\text{tel}}\mathbf{U}_{\text{mic}})$
    $\mathbf{U}_{\text{int}-\text{tel},\text{int}-\text{int}} = (\mathbf{U}_{\text{tel}}\mathbf{U}_{\text{mic}}\mathbf{U}_{\text{int}})$
- Linear kernel used
- NO score normalization

SVM MLLR-CMLLR - Extended Core-Core Cond 5



tel-tel, different number

Prosodic JFA System

Features

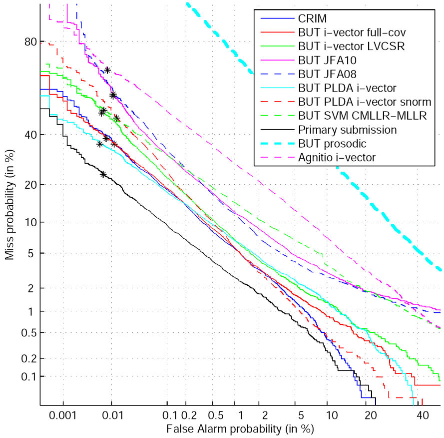
- based on Duration and short time Pitch & Energy
- 6 DCT coefficients of temporal trajectories of pitch and energy
- only voiced part within fixed 300ms window (50ms shift)
- duration is number of voiced frames within 30 frame interval

Model

UBM	GD, 512G, Diag Cov, Varflooring applied
JFA System	$V = 100$, $U_{\mathrm{tel}} = 40$

- Linear scoring was used
- ZT-norm score normalization applied

Prosodic JFA Systems - Extended Core-Core Cond 5



tel-tel, different number

BUT's Conclusions

- PLDA system with NO score normalization seems to be always well calibrated.
- The best performing system for 2010 is 2-times better than the best 2008 system (at least for the new DCF).

JFA with i-vectors as features

Assume that there are matrices U (eigenchannels) and V (eigenvoices) such that

$$\mathrm{i-vector} = m + Ux + Vy + \mathrm{noise}$$

where x (channel factors) and y (speaker factors) have standard normal distributions.

Because each speech segment is represented by a single i-vector, rather than by a sequence of cepstral vectors, the UBM drops out. This version of JFA is known as **Probabilistic Linear Discriminant Analysis** (PLDA).

Because i-vectors are of relatively low dimension (e.g. 400), a fully Bayesian treatment is feasible. This is difficult to do with JFA.

Heavy-tailed PLDA

Retain the assumption that speaker and channel effects are additive and statistically independent:

$$\mathrm{i-vector} = m + Ux + Vy + \mathrm{noise}$$

but assume that the priors on x and y have **power law** rather than Gaussian distributions.

Power law: There is an exponent $k > 0$ such that

$$P(x) = O(\|x\|^{-k})$$

as $\|x\| \to \infty$.

Heavy-tailed PLDA can be implemented in such a way that Gaussian PLDA is a limiting case.

Gaussian vs. heavy-tailed

Gaussian modeling is ill-equipped to handle exceptional speaker and channel effects (e.g. speakers whose native language is not English, severe channel distortions)

- The Gaussian assumption effectively prohibits large deviations from the mean
- Maximum likelihood estimation of a Gaussian (i.e. least squares) can be thrown off by outliers (and by mislabeled data in particular).

Heavy-tailed PLDA includes additional hidden variables to model outliers.

In the Gaussian case, posterior and likelihood calculations can be performed exactly.

In the heavy-tailed case, variational Bayes is needed to handle the additional hidden variables. See my Odyssey presentation, available at

`http://www.crim.ca/perso/patrick.kenny`

Outlier modeling in heavy-tailed PLDA seems to do away with the need for score normalization in general. (Score normalization is actually harmful.)

For **telephone speech** we found that on NIST 2008 SRE data

- Heavy-tailed PLDA without score normalization works better than Gaussian PLDA with score normalization
- Gaussian PLDA with score normalization is comparable to cosine distance scoring
- All three work better than traditional JFA
- Error rates measured by 2008 DCF, EER

For **microphone speech** heavy-tailed PLDA modeling breaks down if it is left to its own devices. Microphone transducer effects are so non-Gaussian as to be pathological. More development is needed.

Performance of heavy-tailed PLDA on the core condition

- See the CRIM det curves in the first part of the presentation
- The Agnitio-BUT Gaussian PLDA system was developed independently of the CRIM heavy-tailed system
- Heavy-tailed did well in development, less well in the eval
- More experimentation needed

Performance of heavy-tailed PLDA on the non-core conditions

Table: *Rankings of the CRIM stand-alone system on the non-core conditions. NDCF = normalized detection cost function.*

condition	rank	actual NDCF	min NDCF
core-10sec	5	0.372	0.365
8summed-core	1	0.045	0.041
8conv-10sec	4	0.270	0.258
core-summed	2	0.193	0.158
10sec-10sec	1	0.590	0.548
8summed-summed	2	0.092	0.077
8conv-summed	3	0.127	0.068
8conv-core[1]	5	0.411	0.253

[1] 2010 cost function

Cross-gender trials

The decision thresholds for the summed tests were poorly set.

The summed-tests involve **cross-gender** trials. These are tricky for systems that use score normalization, since the z-norm and t-norm imposter cohorts have to be chosen in a trial-dependent way.

We adopted a very simple strategy: for trials involving male targets we used a heavy-tailed PLDA model trained on male data (without score normalization) and similarly for females.

This is vulnerable to gender labeling errors. In the eyes of a male PLDA model, two female speakers may appear to be the same, resulting in a false alarm.

Parting thought

It may be better to design a system that does not make use of the gender labels.

Aside from its practical interest, this could pay off in the 4 wire tests as well.