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** Pavel Matějka, Ondřej Glembek, 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

- Try to survive the new DCF.
- Try some new i-vector solutions.
- Try to derive benefit from quality measures.
- Try to 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. See system description for details.
- CRIM also made their own submission for non-core tasks, which will be presented by Patrick Kenny.

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

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.

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

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.

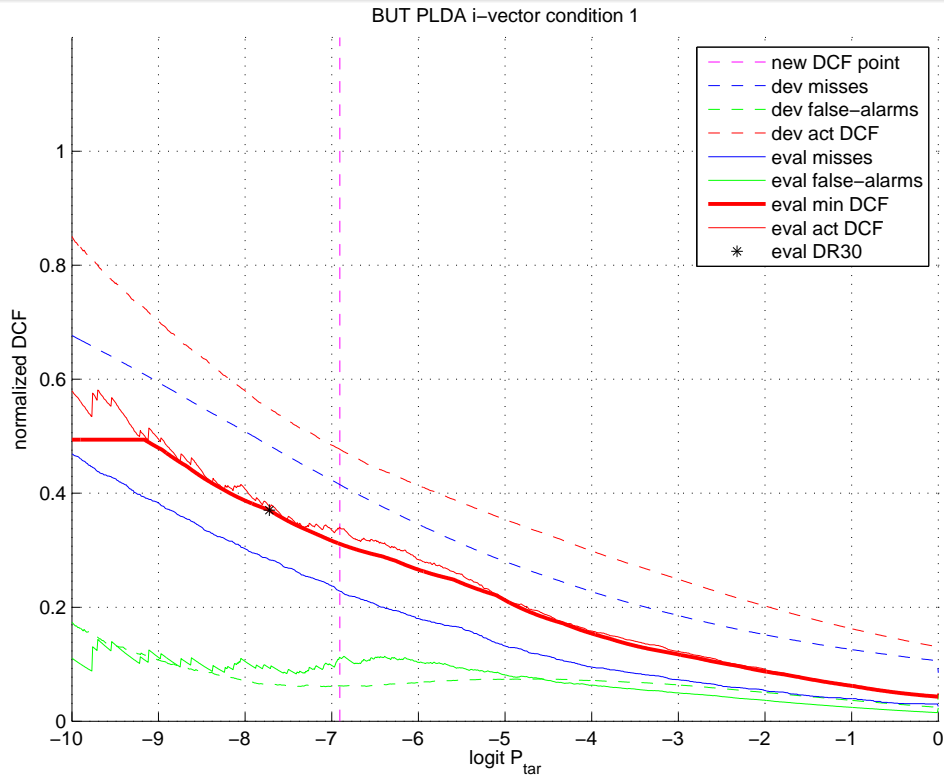Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

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.
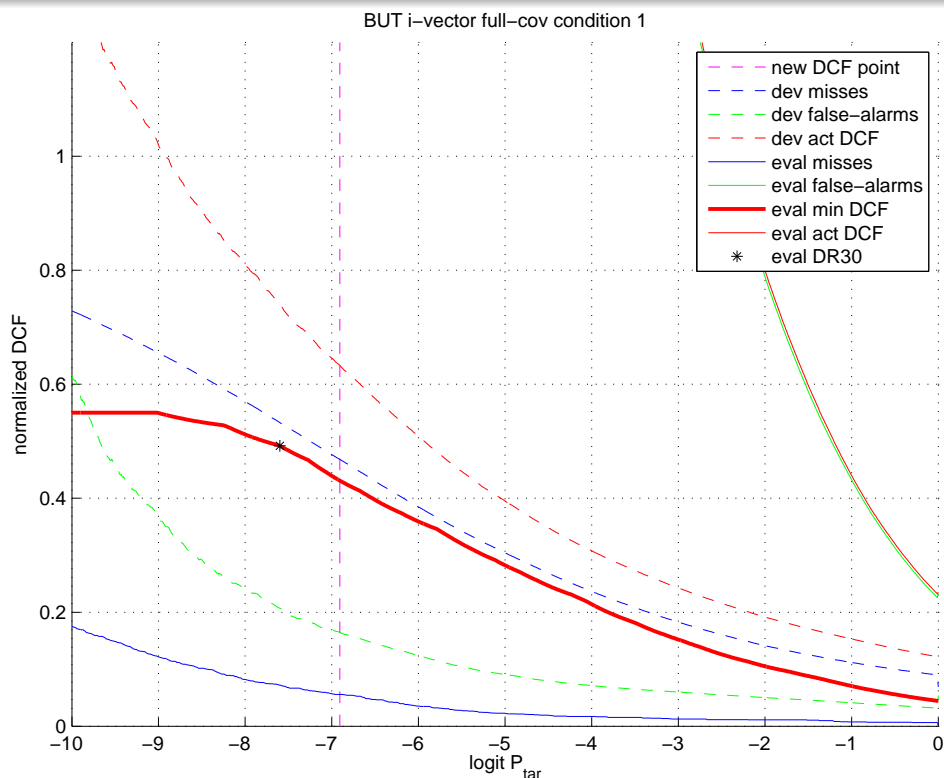- X-axis: The operating point is parametrized by the prior.

Examples follow.

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

Normalized DCF Curve: Example of Good Calibration

BUT PLDA i−vector condition 1

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

Normalized DCF Curve: Example of Bad Calibration

BUT i−vector full−cov condition 1

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

Normalized DCF: Example of Mediocre Calibration

BUT JFA10 condition 3

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation
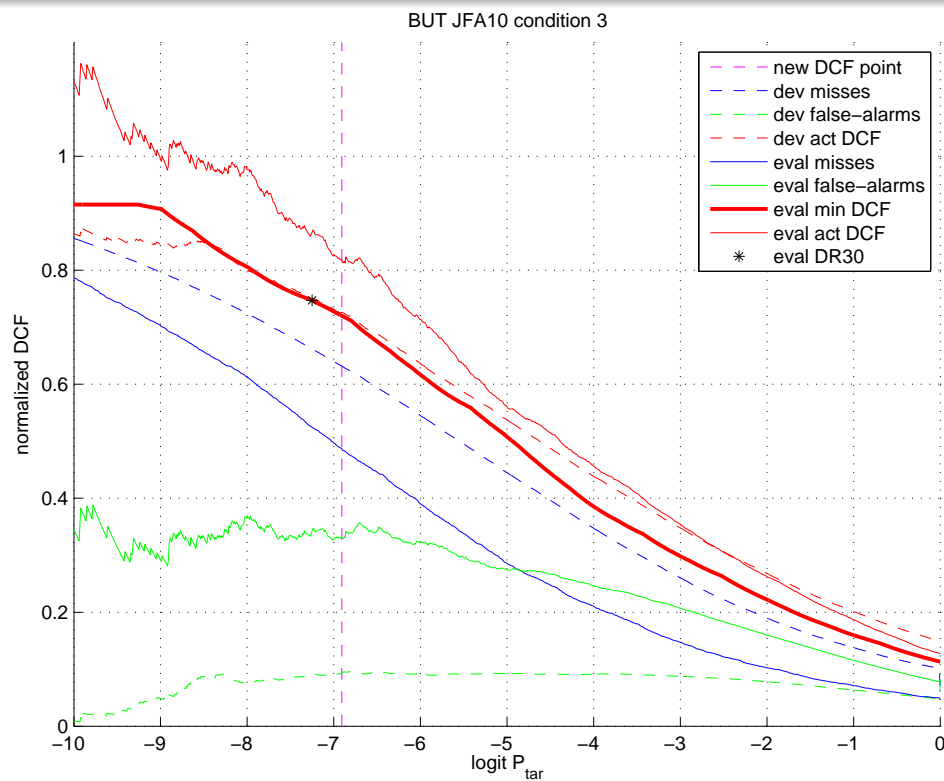
Calibration
Fusion
Quality

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.)

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

ABC-1 Calibration
Condition 5: Tel-Tel

	System	act DCF	min DCF
1	ABC-1	1.73	0.32
2	without broken sub-system	0.81	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 removed.

3. Broken sub-system fixed and included. Corrected some ill-advised development trial index pruning.

4. As 3, but also replaced non-linear *s-cal* fusion with plain linear fusion.

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

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 ...

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

Un-normalized PLDA
Robust against calibration mismatch?

The 'BUT PLDA' sub-system used:

- BUT's i-vectors (as did two other BUT systems)
- AGNITIO's PLDA model training and scoring. (This PLDA was Gaussian, not heavy-tailed like CRIM's PLDA.)
- The same development data as all other ABC sub-systems
- Some careful tuning by Lukas
- No score normalization

This system got good calibration for all conditions despite being very similar to—and using the same resources as—the other i-vector systems. Is this because it has no score normalization?

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

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

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

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

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

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.
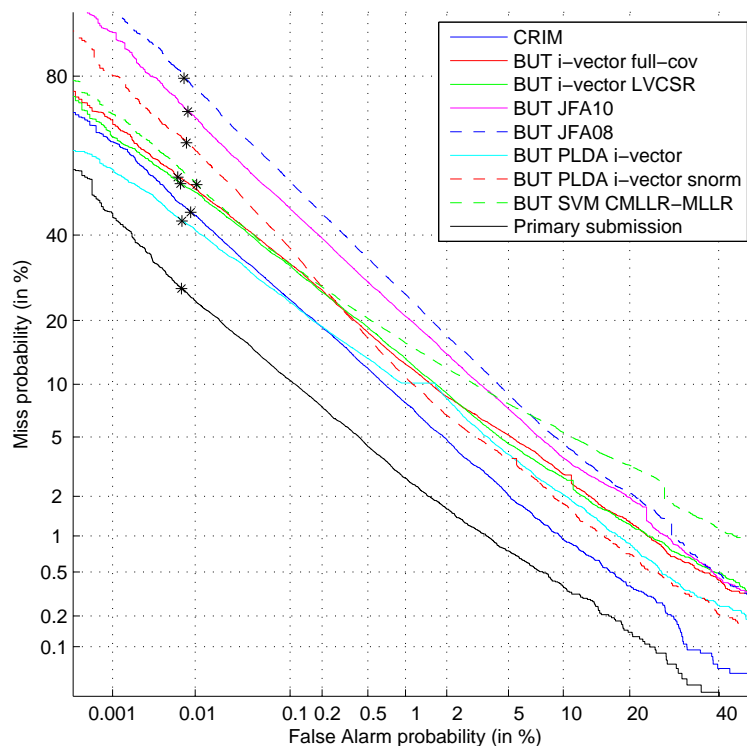- For conditions 1-4, these fusions included quality measures.

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

ABC-1 Extended Core-Core Condition 1

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

ABC-1 Extended Core-Core Condition 2

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

ABC-1 Extended Core-Core Condition 3

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

ABC-1 Extended Core-Core Condition 4

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

ABC-1 Extended Core-Core Condition 5

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality
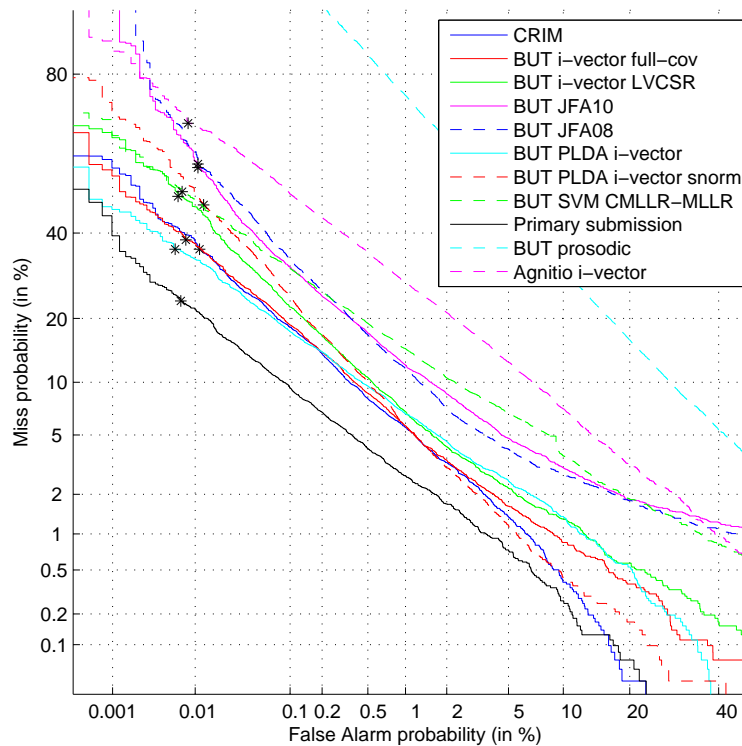
Quality Measures

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

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

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.
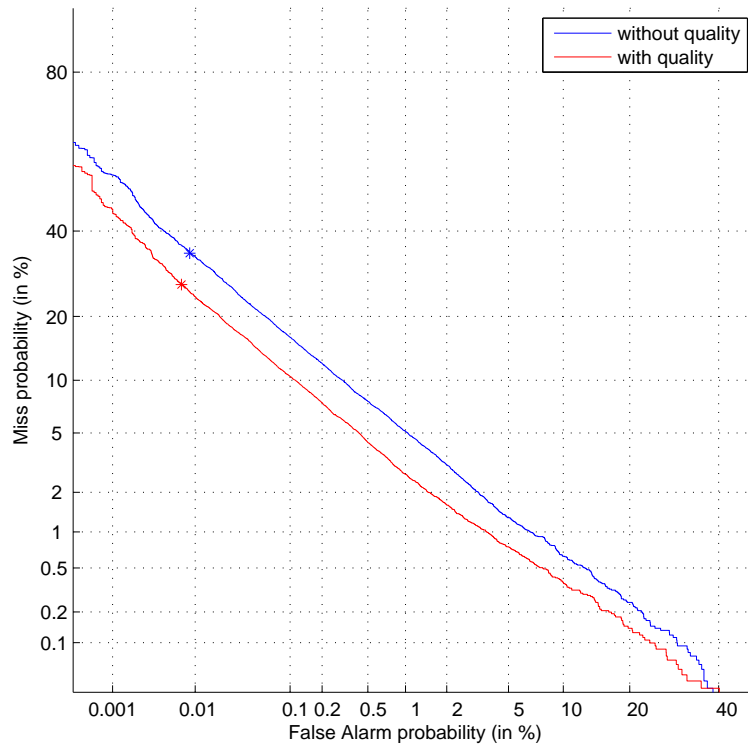- We compare fusions, with and without quality measures.
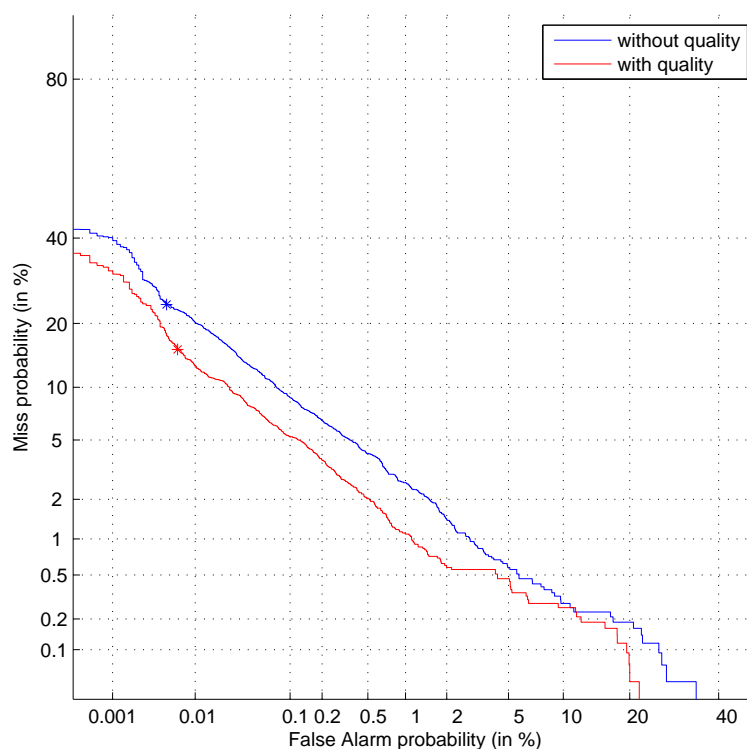
Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

Fusion with Quality for Ext. Core-Core Condition 1

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

Fusion with Quality for Ext. Core-Core Condition 2

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

Fusion with Quality for Ext. Core-Core Condition 3

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality

Fusion with Quality for Ext. Core-Core Condition 4

Introduction
Analysis
AGNITIO Conclusion
CRIM Presentation

Calibration
Fusion
Quality
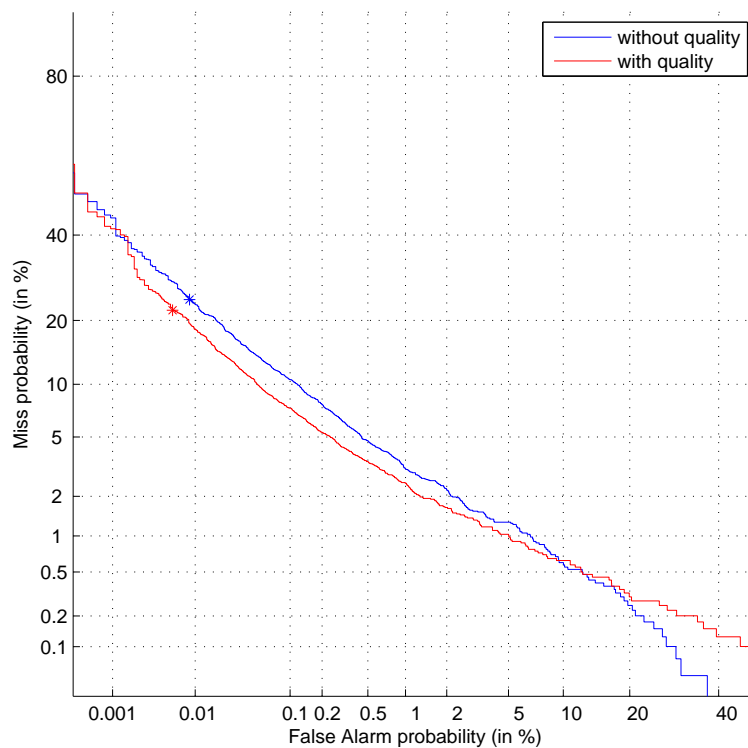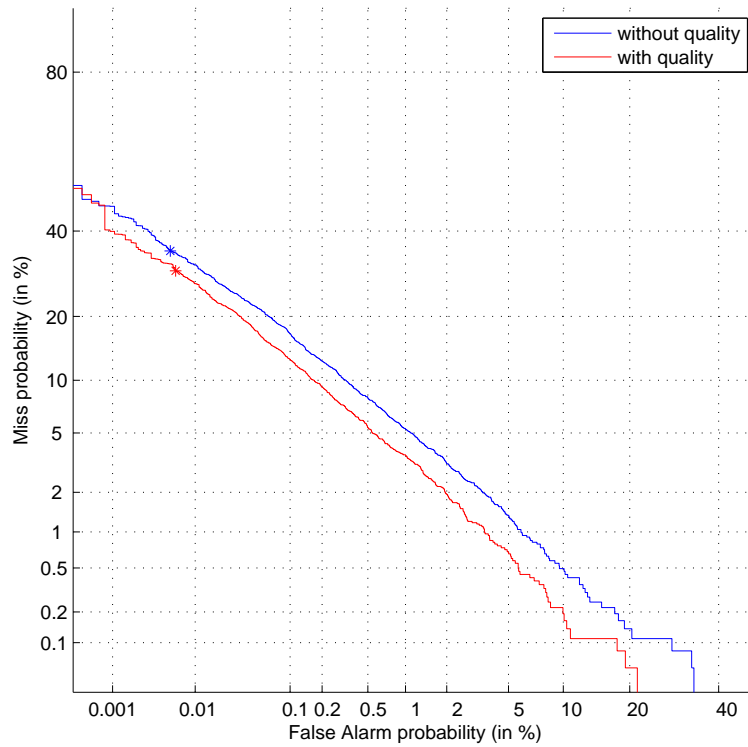
Fusion with Quality for Ext. Core-Core Condition 5
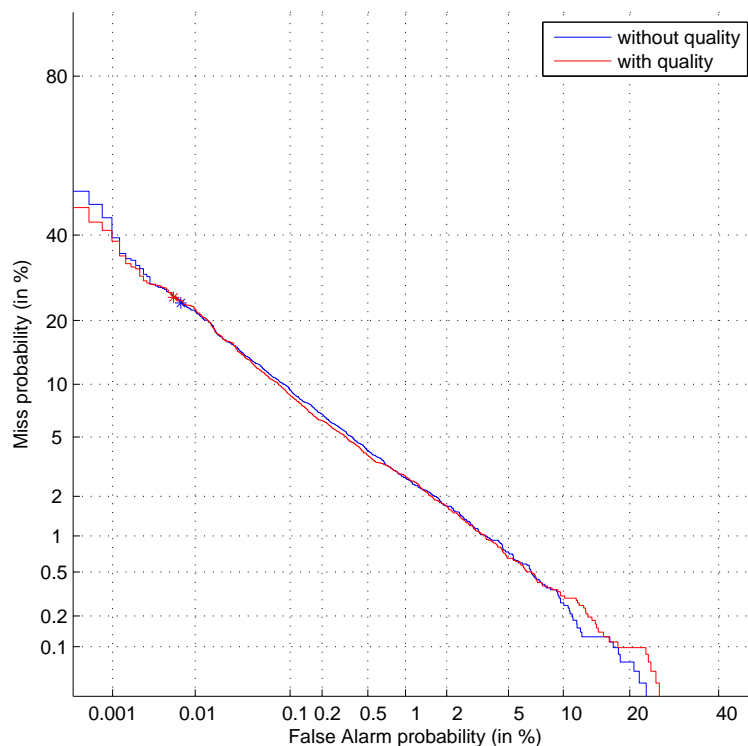
AGNITIO's Conclusion

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

CRIM Presentation
Bayesian Speaker Verification with Heavy-Tailed Priors

Presenter: Patrick Kenny

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.

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 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

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.