Recent Improvements in the CUED Diarisation System

Sue Tranter, Rohit Sinha,
Mark Gales & Phil Woodland

March 19th 2005

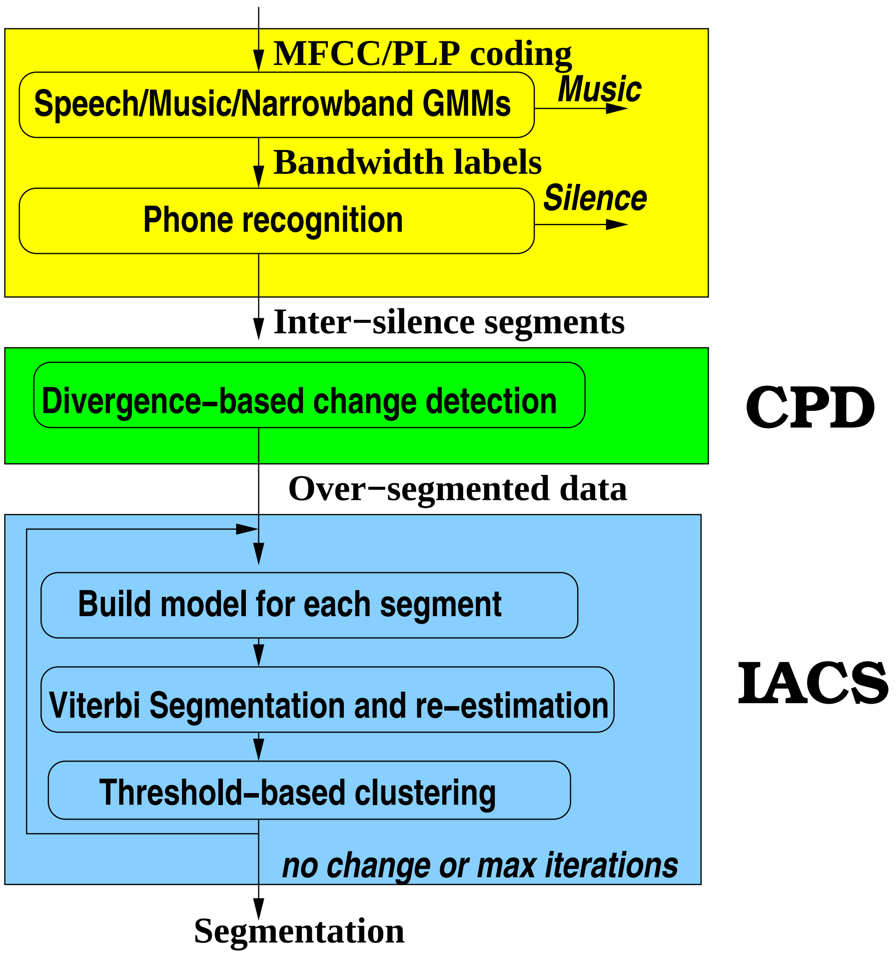
Cambridge University Engineering Department

Progress Since RT-04 Workshop

System	Dev-24 DER	Dev-12 DER	Eval DER
RT-04 Eval System (Oct 2004)	17.7%	17.2%	24.0%
RT-04 Workshop (Nov 2004) (without topdown clustering)	19.2%	20.2%	17.9%
MDE Tech Meeting (Mar 2005)	9.0%	7.7%	6.9%

Dev-24 data = 24 shows (eval03, didev03, dev04f2, sttdev04)
Dev-12 data = 12 shows (eval03,dev04f2 = RT-04 diarisation dev data)
Eval data = 12 shows (eval04 - reference 22nd Dec 2004)

System Architecture - RT04 Workshop



Iterative Agglomerative Clustering Stage (IACS)

- Run in two stages, first with diagonal (PLP_0_D_A) and second with full covariance (PLP_0).
- Each stage runs up to 6 iterations.
- RT-04 system used a constant threshold on the likelihood for merge decisions (no BIC penalty weight)
- The method of updating when clusters were combined was changed from centroid clustering to forming the stats from the concatenated data.
- Options for using a ('local') BIC criterion for ordering the merges and/or merge decision were added
- A furthest neighbour scheme (which didnt need distance recomputation after each merge) was also added.

IACS - Distance Metrics

The diagonal covariance step was run conservatively to oversegment the data, and fixed for these experiments on the full covariance stage.

ID	Clustering	Ordering	Decision	Opt-dev	Eval	(Opt Eval)
1	Centroid	0	constant	19.4	17.7	(17.6)
2	Concat.	0	constant	19.1	17.1	(17.1)
3	Concat.	0	BIC	18.9	20.3	(16.8)
4	Concat.	BIC	BIC	18.6	17.9	(17.7)
5	Furthest N	0	constant	18.5	19.3	(19.0)

- Furthest Neighbour (5) performed best on the dev but not the eval data.
- Using a constant in the decision (2) gave the best (non-tuned) eval score.
- The more standard BIC method (4) did reasonably on both data sets.
- The results are often sensitive to relatively small parameter changes.

IACS - Summary

The baseline system uses:

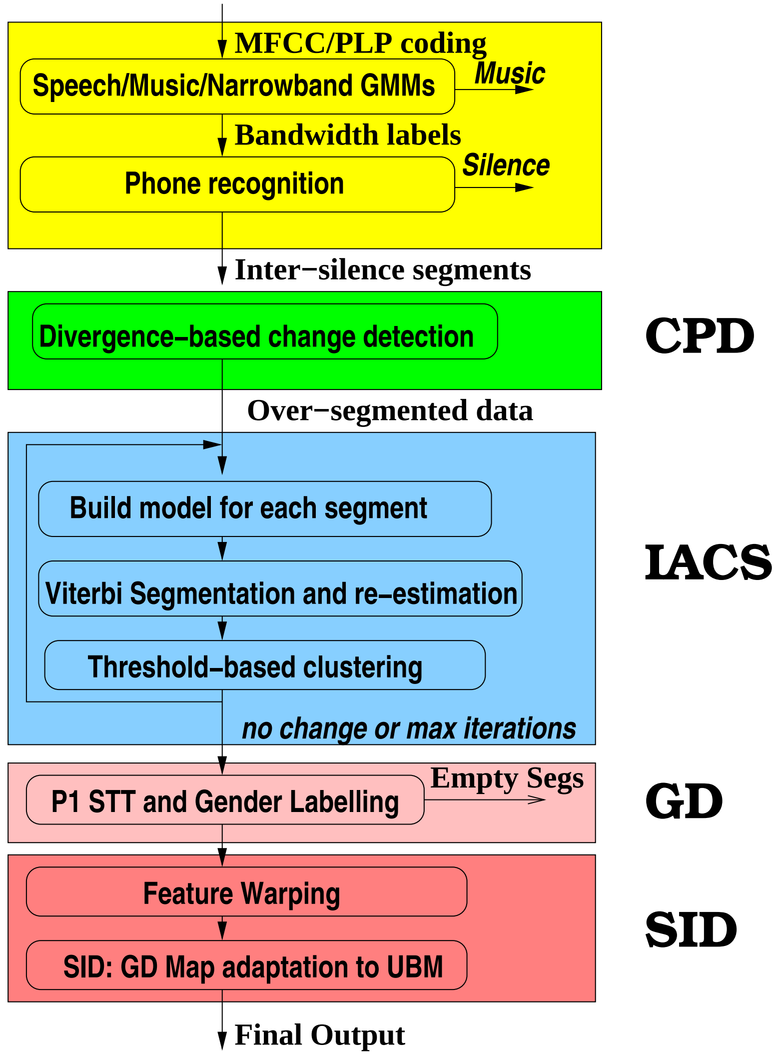
- 'local' BIC for both ordering and decision in merging stage
- Multiple iterations at optimal (dev) α
- Phasing of 1 iteration of $\alpha = 1$.
- This underclusters the data for subsequent SID stage.

The results are:

DataSet	MS/FA/SPE/DER	Cluster Imp †	Seg Imp †
Dev-24	1.2/1.1/17.9/20.17	6.59 @ 718	4.36 @ 2363
Dev-12	1.0/1.3/20.2/22.47	5.32 @ 321	4.22 @ 1078
Eval (12)	0.3/1.1/17.4/18.75	5.04 @ 336	3.63 @ 1072

† Seg/Cluster Imp = DER with oracle clustering of segments/clusters. (including MS/FA)

System Architecture - Adding SID

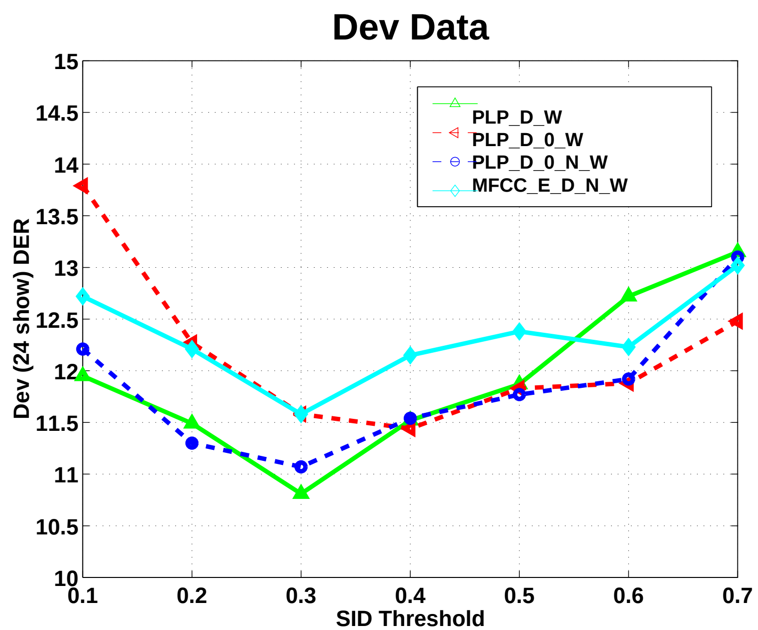


SID stage - Description

- Based on LIMSI's "SID-like" stage in their RT-04 evaluation system.
- Perform agglomerative speaker clustering using the cross log-likelihood ratio (CLR) between clusters.
- Clustering is done separately for each bandwidth and gender.
- Each cluster model is derived by MAP adapting (mean only) a universal background model (UBM).
- The stopping criterion used is a threshold on global CLR, θ_{CLR}.
- Feature warping is applied to reduce the effect of acoustic environment.

SID stage - Effect of Features and Feature Warping

Different warped features were investigated in particular the inclusion of c0 (_0), energy (_E), or just the differentials thereof (_N).



- PLP with deltas and no energy performed the best.
- Feature warping improved the DER from 18.1% to 10.8%.

SID stage - Variable Prior (VP) MAP

For MAP adaptation the mean, μ, is changed depending on a prior model p and the data d:

$$\hat{\mu}^{(1)} = \frac{\gamma_d^{(1)} \mu_d^{(1)} + \tau \mu_p}{\gamma_d^{(1)} + \tau}$$

- A small τ makes the mean stick to a few speakers in the data and thus is robust to the SID threshold, θ_{CLR}, but may get 'misled' by the data.

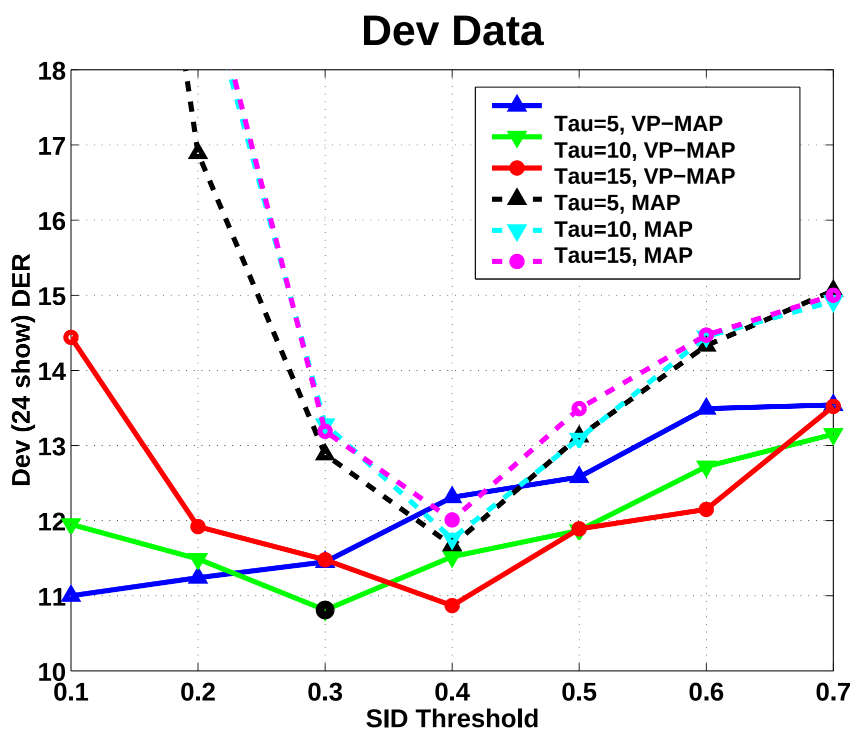
Variable Prior (VP) MAP uses $\hat{\mu}^{(N)}$ instead of μ_p for iteration N+1.

e.g. 2nd iteration $$\hat{\mu}^{(2)} = \frac{\gamma_d^{(2)} \mu_d^{(2)} + \tau \left(\frac{\gamma_d^{(1)} \mu_d^{(1)} + \tau \mu_p}{\gamma_d^{(1)} + \tau} \right)}{\gamma_d^{(2)} + \tau}$$

- The numerator μ_p term becomes weighted by $(\frac{\tau^2}{\gamma_d^{(1)} + \tau}) \leq \tau$
- More iterations decreases the prior's influence as the new models improve.

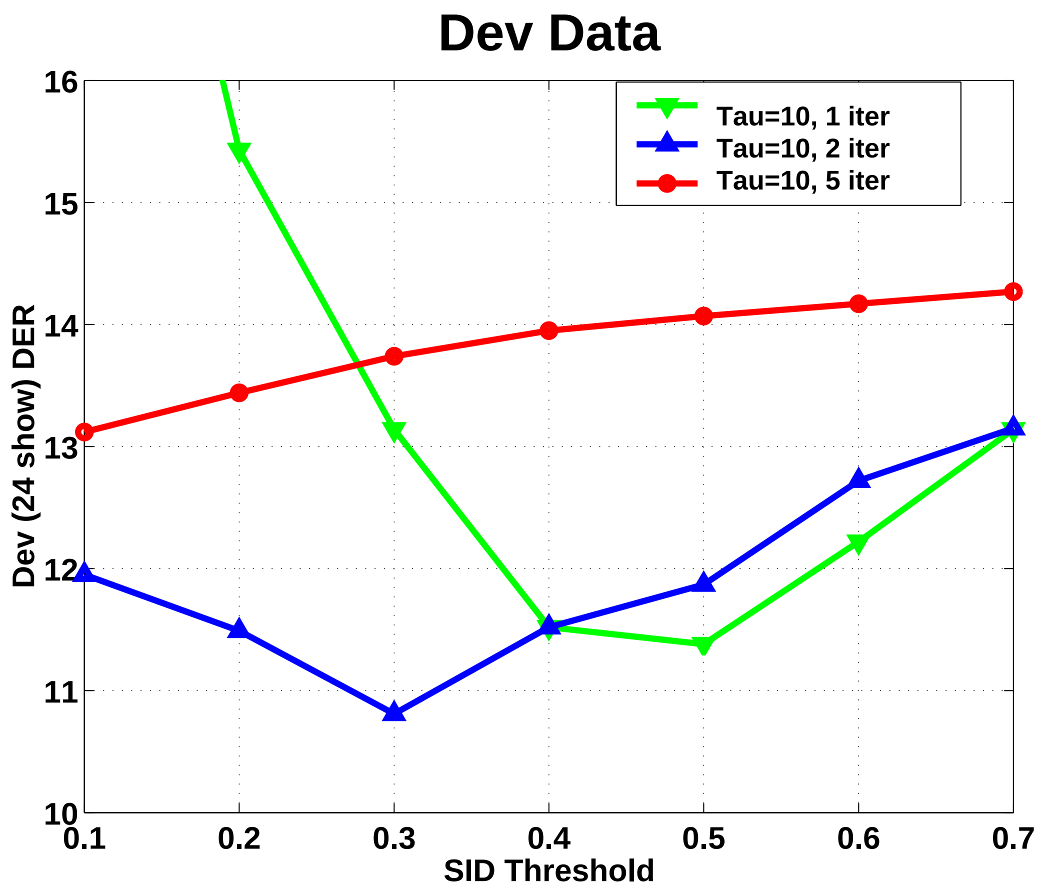
SID stage - Type of MAP and τ

We compared MAP and VP-MAP for different τ values and 2 iterations.



- VP-MAP outperforms MAP for every value of τ.
- Use VP-MAP, $\tau = 10$, giving 10.8% on dev and 9.9% on eval. (opteval=9.2%)

SID stage - Number of Iterations of MAP



- Use 2 iterations of VP-MAP
- This gives 10.8% on dev and 9.9% on eval.

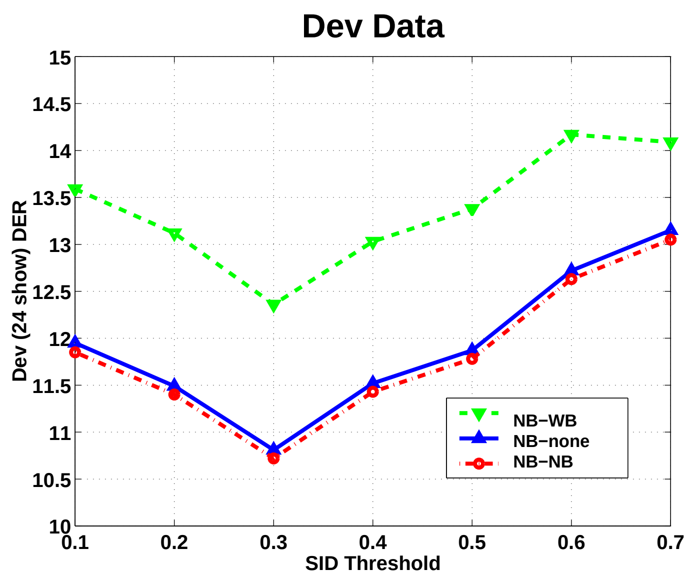
SID stage - Narrowband Data

How to deal with the automatically labelled NB data in the SID stage:

NB-none Pass NB clusters directly to output (default)

NB-WB Run NB clusters through SID using WB coding

NB-NB Run NB clusters through SID using NB coding



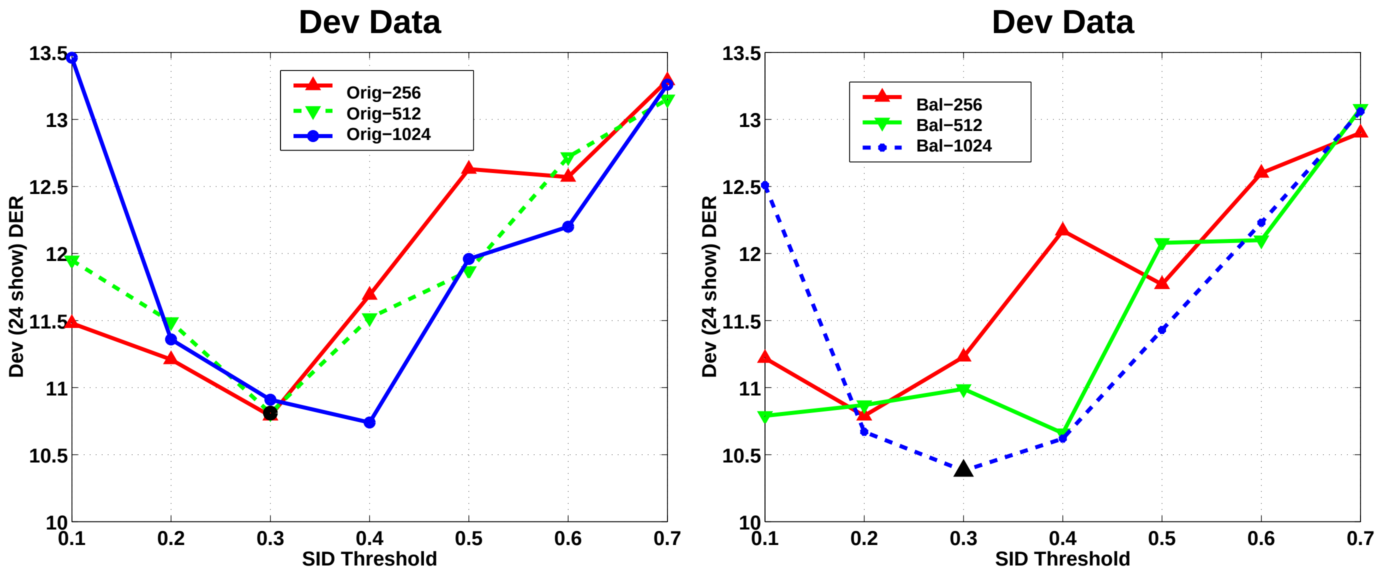
- Use NB coding for NB clusters. (Using WB makes worse - add a NB θ ?)
- Hardly any data classified as NB for eval04 makes this less worthwhile.
- This gives 10.7% on dev and 9.9% on eval.

SID stage - UBM generation

Different UBMs were built for experiments with 256, 512 and 1024 mixtures.

Orig 6 hrs per gender taken from hub4-train 96/7

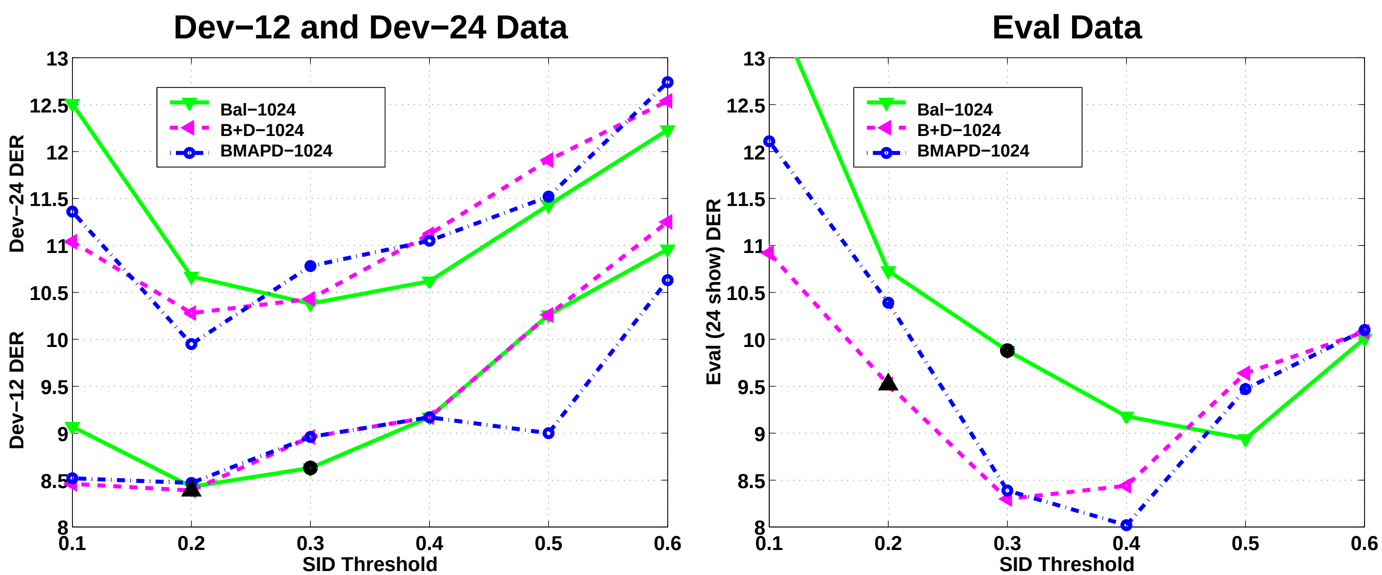
Bal 7.5 hrs per gender taken evenly across all sources in hub4-train 96/7



- The Balanced set performed best with 1024 mixtures and θ_{CLR}=0.3
- This gives 10.4% on dev and 9.9% on eval. (opteval=8.9%)

SID stage - Adding Dev Data to UBM

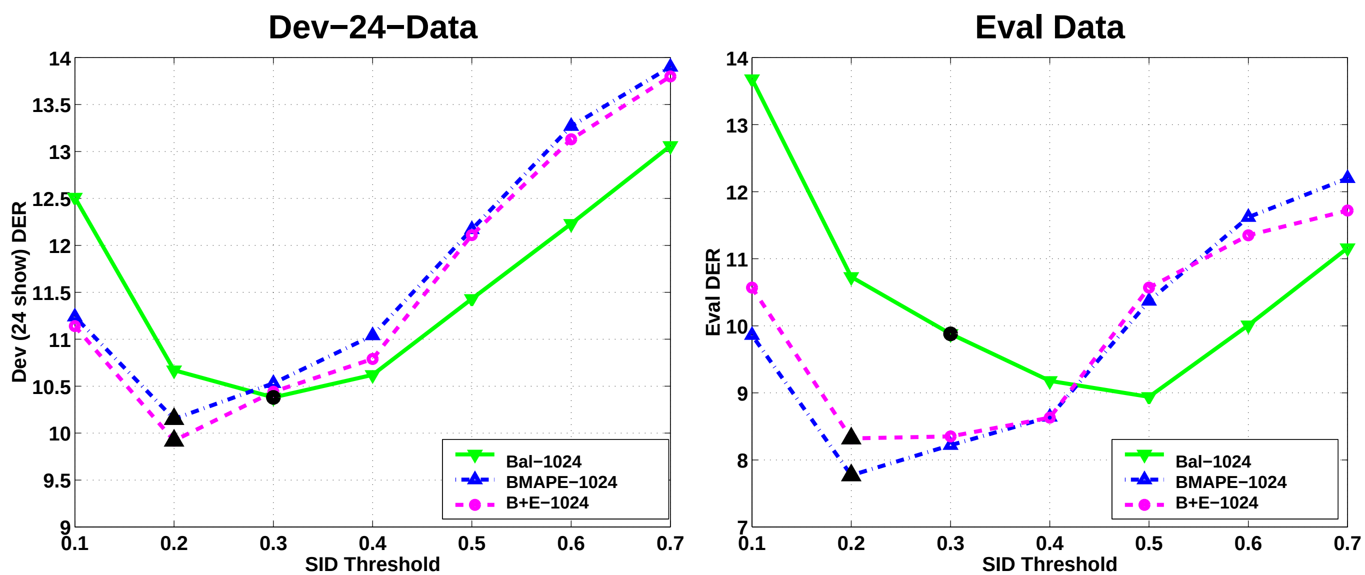
- `sttdev04` and `didev03` development data, D, included in UBM.
- B+D retrained UBM with Bal+D data, BMAPD MAPed Bal to D
- dev-12 remained 'uncontaminated' dev set.



- Using dev data made much larger improvements on eval than dev-12 DER.
- BMAPD-1024 gives 8.5% on dev-12 and 10.4% on eval (opteval=8.0%)
- B+D-1024, gives 8.4% on dev-12 and 9.5% on eval (opteval=8.3%)

SID stage - Adding Eval Data to UBM

- New UBMs, B+E and BMAPE, were created using the whole test data set.
- *System* gender labels were used. (no real cheating but against eval rules).
- Using just the target show (rather than whole test set) did not work.



- Using all the test data improved the best performance over the Bal UBM.
- BMAPE-1024, gives 10.2% on dev24 and 7.8% on eval (opteval 7.8%).
- B+E-1024 gives 9.9% on dev24 and 8.3% on eval (opteval 8.3%).

SID stage - Summary

- We built a successful SID-like stage using LIMSI's as a base model.
- Feature warping slashed our dev24 DER from 18.1% to 10.8%.
- VP-MAP was introduced and shown to outperform MAP.
- 2 iterations of VP-MAP using PLP_D and $\tau = 10$ worked best.
- Carefully adding (reference) dev data into the UBM helped eval performance.
- The final system gave a DER of 8.4% on dev12 and 9.5% on eval.
- Adding the test data itself into the UBM improved performance, giving 8.3% or 7.8% on the eval data depending on the method used.

SID stage - Using LIMSI's Segments

LIMSI were kind enough to provide us with the input they used for their SID stage in the RT-04 evaluation.

	SID input			SID DER*	
	Segment Impurity MS/FA/SPE/DER @ #Seg	Cluster Impurity DER @ #Spk	DER	θ_{dev} 0.2	(θ_{eval}) (0.3)
CUED	0.3/1.1/2.3/3.63 @ 1072	5.04 @ 336	18.8	9.5	(8.3)
LIMSI	0.2/1.8/1.0/3.05 @ 1110	4.02 @ 477	18.4	9.1	(7.6)

* B+D-1024 model used, $\tau = 10$, VP-MAP, 2 iterations

- DERs of 7% were obtained using LIMSI's SID input with different UBM models. (With a *further* gain of 0.6% by using the CUED SAD labels)
- We need to improve our 'pre-SID' segmentation/clustering !

Altering the Change Point Detection

The change point detection was rewritten:

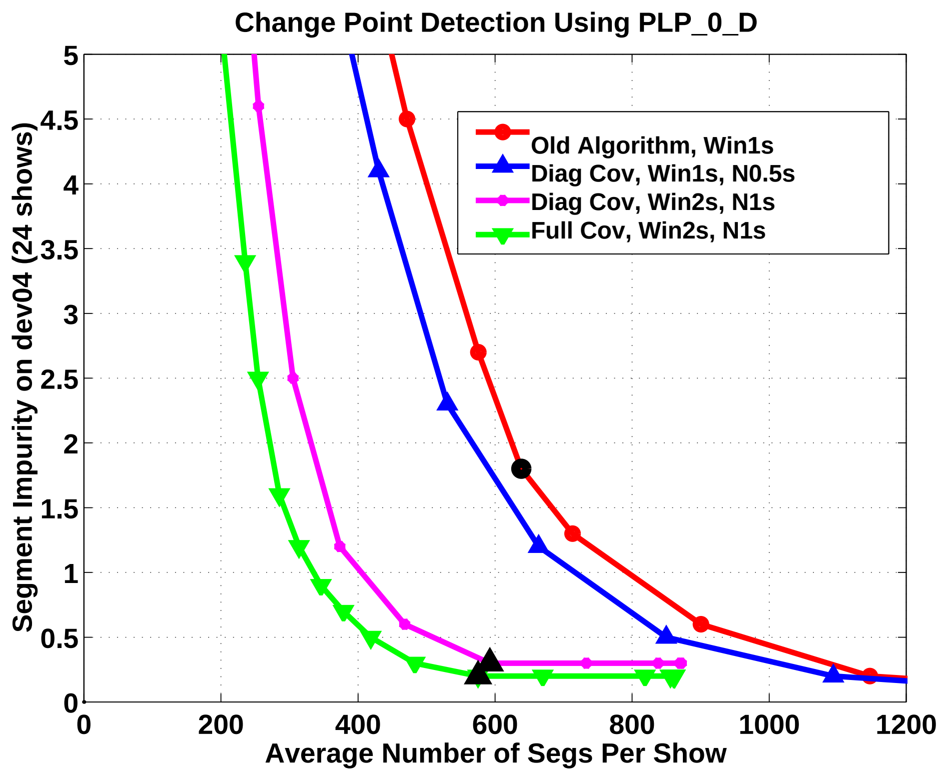
- Finding peaks directly from distance metric improved potential purity.
- New minimum length constraint enforced by removing the smaller of neighbouring peaks (L to R) reduced number of segments dramatically.

Results:

- Larger window size improved performance.
- Full covariance model worked better than diagonal for the larger window size.
- Switching features from PLP_0_D to MFCC_E_D_A_N did not help.
- Using feature warping degraded performance severely.

Change Point Detection - Results

The speaker error component from the ideal clustering of the segments is used to measure the segment impurity.



- Best Performance from 2s windows, 1s min length, full covariance.

Change Point Detection - Effect on Whole System

dev-24 data	CPD out	IACS out			dev-12*
	Seg Imp†	Seg Imp†	Clust Imp†	MS/FA/DER	SID DER
baseline	3.99 @10573	4.36 @ 2363	6.59 @718	1.2/1.1/20.2	8.4
newCPD-diagc	2.59 @11348	4.11 @ 2385	6.40 @722	1.2/1.1/19.2	8.6
newCPD-fullc	2.50 @11299	4.21 @ 2371	6.39 @720	1.2/1.1/20.3	8.1

* B+D-1024 model used, $\tau = 10$, VP-MAP, 2 iterations, $\theta_{opt}(dev12)$

† Segment/Cluster Imp = DER with oracle clustering of segments/clusters. (including MS/FA)

- Segment purity much better after CPD stage (˜0.25% SPE).
- Early promise not carried through. (no gain in DER seen on eval data)
- IACS should be adjusted (e.g. removing diag cov stage as segs now >1s).
- First results on retuned IACS give 7.4/8.8% on dev 12/24 and 8.6% on eval. (Results with B+E models give 7.7/9.0% on dev12/24 and 6.9% on eval.)

Future Work

Short Term Goals

- Re-tune IACS with new CPD output and try with B+E model.

Things to think about

- CPD: add smoothing such as a median filter or hamming window
- IACS: Try encorporating feature warping
- SID: Use BW-labelled data to build GMMs.
- FINAL: Post-process output with STT cues.

Conclusions

- The DER of the CUED Diarisation system on the eval04 data has been reduced from 17.9% to 6.9% with a similar drop in the dev data DER.
- Most of the improvement came from adding a SID-like stage in a similar style to LIMSI's.
- DERs of around 6.5% are possible on eval04 data with this method.
- These experiments will be written up for a Eurospeech 2005 submission.

Thanks are due to the *LIMSI speaker recognition team*, and in particular *Claude Barras* for helping us improve our system performance by providing intermediate files and helpful advice.