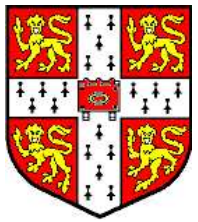
Interactions Between Diarisation and STT

Sue Tranter[1], Kai Yu[1], Doug Reynolds[2], Do Yeong Kim[1],
Gunnar Evermann[1], Phil Woodland[1] and the HTK STT team[1]

May 20th 2003



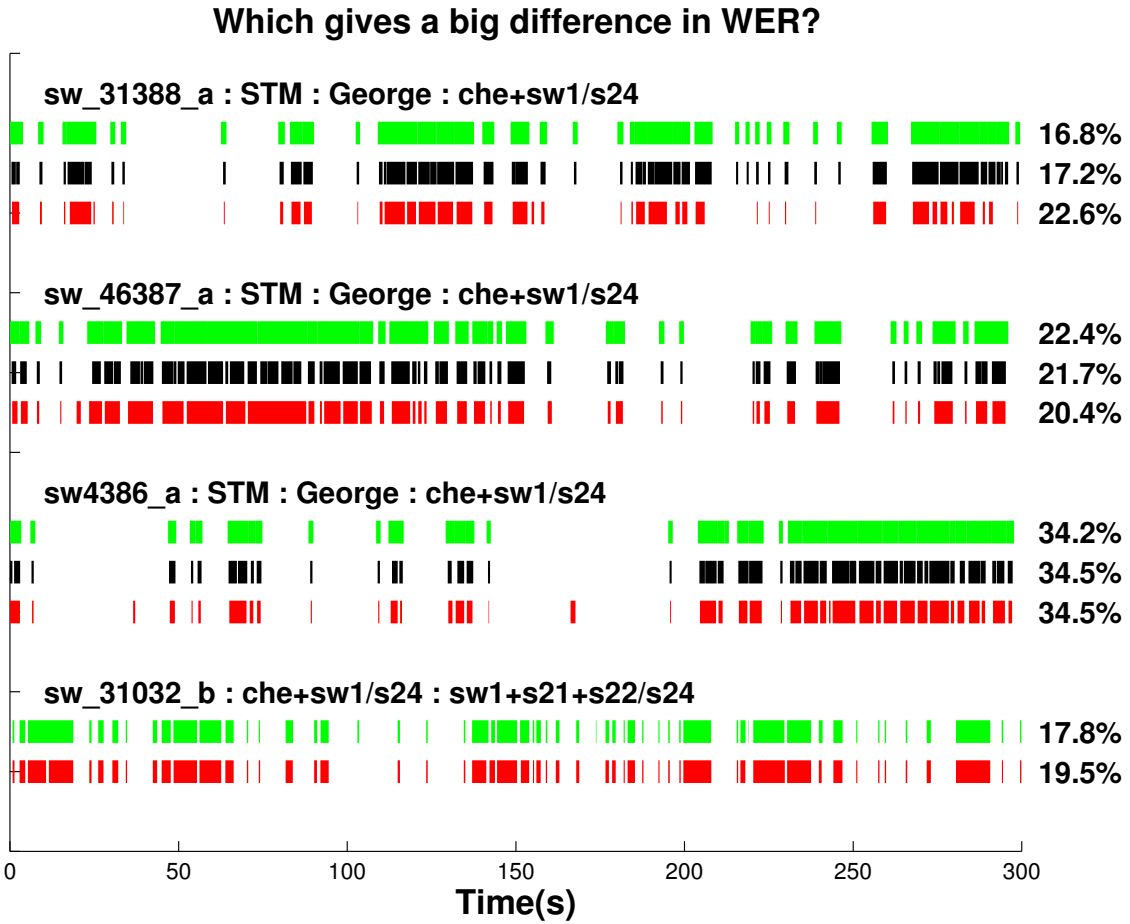
Cambridge University$^{(1)}$ and MIT-Lincoln Labs$^{(2)}$

EARS Workshop: May 2003

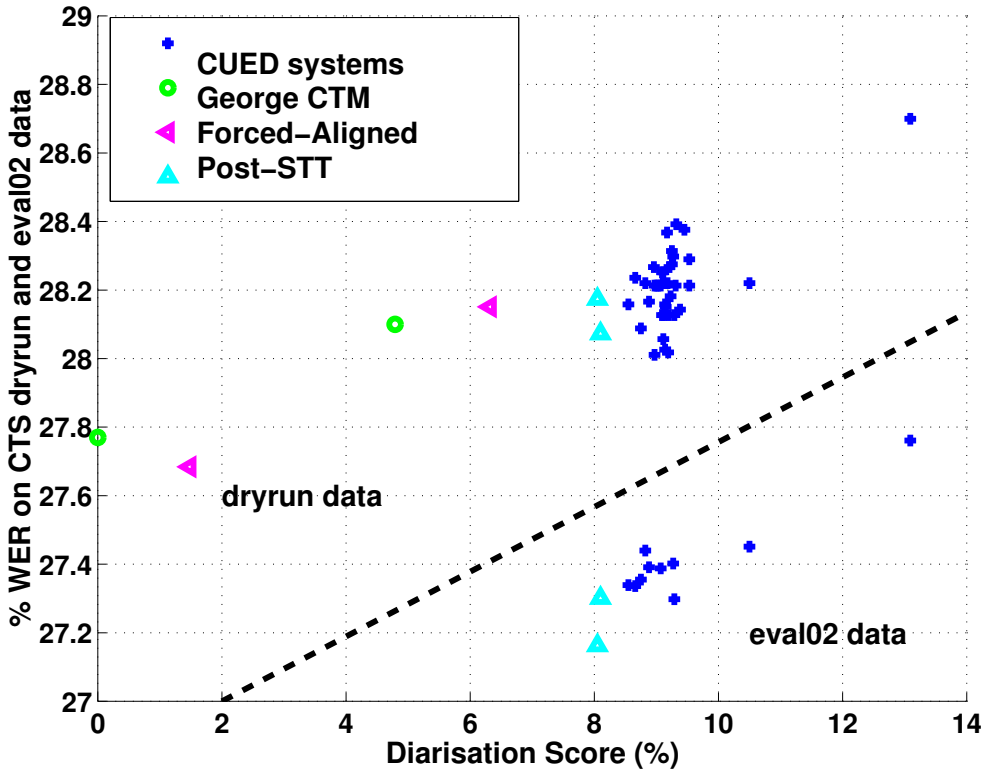
Overview

- CTS
  - Are diarisation scores correlated to WER ?
  - How should we use diarisation output for STT ?
  - Can we improve diarisation scores using info from STT ?
- BNEWS
  - Can removing commercials help diarisation or STT ?
  - Are diarisation scores correlated to WER ?
  - Can we use diarisation output for STT ?
  - Can we mix-and-match components of diarisation systems ?
- Conclusions

CTS - Guess the WER difference game



CTS - Relationship between diarisation score and WER



- Some correlation between diarisation score and eval02 WER.

[NB - all runs have 0.6s smoothing, plus 0.2s padding for WER - STT 10xRT dryrun 2002 system]

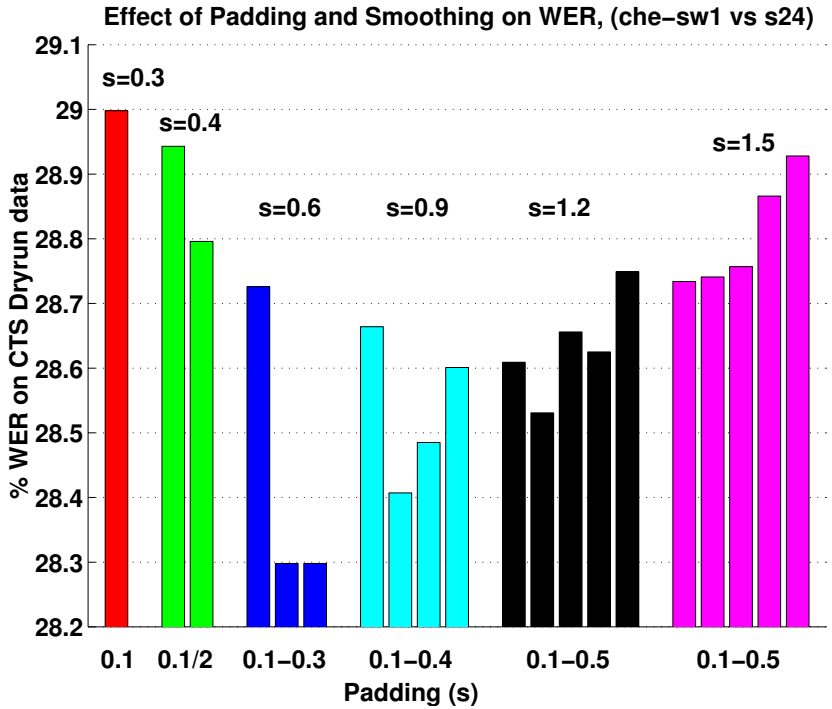
CTS - Correlation between diarisation and WER

		MISS	FALARM	DIARY	WER(dry)
N=12 eval02 data	MISS	1.0000	-	-	-
	FALARM	-0.4219	1.0000	-	-
	DIARY	-0.0239	0.9163	1.0000	-
	WER(dry)	0.3029	0.6872	0.8914	1.0000
	WER(eval02)	-0.1756	0.8981	0.9145	0.8619
N=44 dryrun data†	MISS	1.0000	-	-	-
	FALARM	0.4421	1.0000	-	-
	DIARY	0.7246	0.9383	1.0000	-
	WER(dry)	0.5573	0.7211	0.7676	1.0000

- FALARM highly correlated with diarisation score.
- eval02 WER more correlated to dryrun DIARY score than dryrun WER.

† - Numbers changed slightly since workshop due to the inclusion of a run with non-lex stripped from LDC forced alignment

CTS - How should diarisation output be used for STT?



- Optimal smoothing/padding for WER = 0.6s/0.2 or 0.3s.
- Same minima found with different numbers of Gaussians in models.
- WER on CTS dryrun data reduced from 30.5% to 28.3%.

CTS - Can STT info help improve diarisation scores?

		CTS DryRun[1]			CTS-eval03s[2]			
		MS	FA	DIARY	MS	FA	DIARY	GE
Pre-STT	RT03 (0.05xRT)	2.2	6.3	8.55	8.8	1.9	10.65	1.0
Post-STT	dryrun (10xRT)	1.9	6.2	8.10	-	-	-	-
Post-STT	RT03 (187xRT)	4.0	4.1	8.05	10.0	2.0	11.95	0.0

[1] Reference derived from George's times, removing misc+non-lex, with 0.6s smoothing.

[2] Official RT03 results - reference derived from LDC forced alignments, 0.3s smoothing.

- Using wordtimes from STT output helped on the development data.
- This improvement did not occur on the eval03s data.
  (Note different accuracy of reference data due to method of generation.)
- The gender error was eliminated using the STT output on eval03s.

CTS - Summary of Key Results

		DryRun[1]				eval02
		MS	FA	DIARY	WER[2]	WER[2]
Pre-STT	dryrun system	2.8	10.3	13.09	28.7	27.8
Pre-STT	eval03 system	2.2	6.3	8.55	28.2	27.3
Post-STT	dryrun-10xRT o/p	1.9	6.2	8.10	28.1	27.3
Post-STT	RT03-187xRT o/p	4.0	4.1	8.05	28.2	27.2
Ref	George-CTM (all)	0.0	4.8	4.79	28.1	N/A
Ref	as above (no-nonlex)	0.0	0.0	0.00	27.8	N/A
Ref	LDC Forced-alignment	0.7	5.7	6.32	28.2	N/A
Ref	as above (no-nonlex) †	0.9	0.5	1.48	27.7	N/A
Ref	STM-file [3]	0.0	39.9	39.89	27.7	26.7

[1] Diarisation reference derived from George's CTM, removing misc+non-lex, with 0.6s smoothing.

[2] Recogniser used for WER is 10xRT from dryrun Dec 2002.

[3] The default 0.6s smoothing (+0.2s padding for recognition) was not done on the STM-file.

BN - Issues with Scoring

- The forced alignment is $>$ 1 second off in many places. (use collars?)
- Some foreign speakers are missing from bndidev03 reference MDTM.
- Not enough shows (3 in bneval03s data, 6 in bndidev03).
- Is splitting a 5-minute speaker into 2 clusters really 50% worse than missing 10 speakers of 10s ? (not for STT)
- Is FALARM equally as important as MISS speech ? (not for STT)
- Are there other useful spkr attributes ? (e.g. bandwidth for STT)
- Is it desirable to be able to remove unwanted audio e.g. commercials ?

BN - Removing Commercials for Diarisation

Diary clustering		NONE	CU_EVAL[3]	CU_TDT4[2]	PERFECT
Audio Removed		0.0%	6.75%	18.41%	19.19%
DIARY[1]	MS	0.2 (0.2)	0.5 (0.5)	1.0 (1.0)	0.3 (0.3)
	FA	9.1 (29.7)	9.1 (23.5)	8.9 (12.6)	9.0 (10.0)
	DIARY	33.3 (53.9)	33.6 (48.0)	34.1 (37.8)	36.9 (37.9)

All results are on BN diarisation development data (bndidev03)

[1] numbers in () dont use .spkreval.uem file in scoring, i.e. commercials and vocal noise treated as FA.

[2] CU_TDT4 uses all TDT-4 data (except the dev shows) for the library of commercials.

[3] CU_EVAL does not use any data from the same month as the dev broadcast.

- Untranscribed contemporaneous data helps remove commercials.
- Main diarisation score fairly unaffected by removing commercials.
- Contrast with commercials scored as FAs almost reduced to usual score.

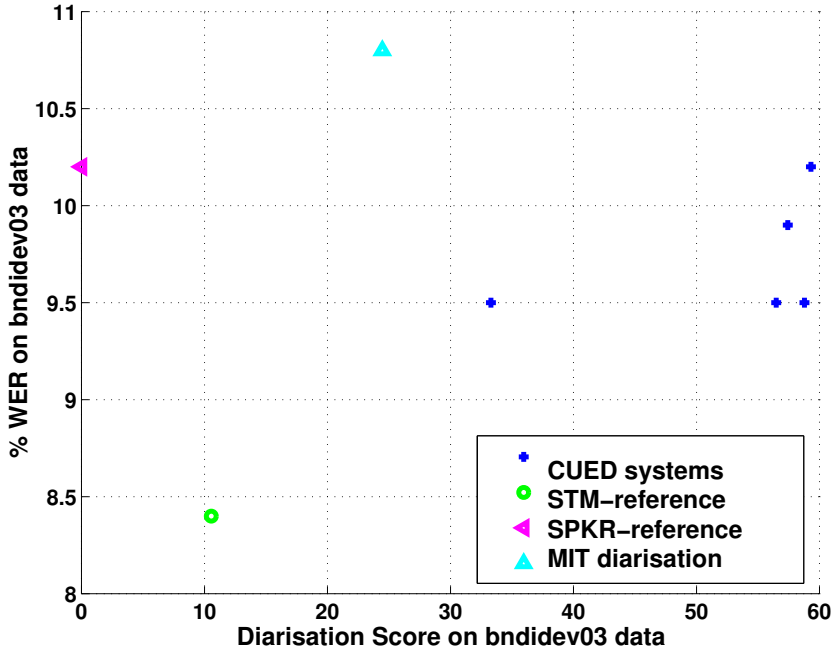
BN - Removing Commercials for STT

STT clustering		Technique for Commercial Removal NONE	CU_EVAL	CU_TDT4	PERFECT
Audio Removed		0.0%	6.75%	18.41%	19.19%
DIARY	MS	0.2 (0.2)	0.5 (0.5)	1.0 (1.0)	0.3 (0.3)
	FA	9.1 (29.8)	9.1 (23.6)	8.9 (12.7)	9.0 (10.1)
	DIARY	56.5 (77.1)	57.4 (72.0)	59.3 (63.1)	58.8 (59.8)
STT[1]	Del	2.0	2.4	2.9	2.1
	Ins	1.6	1.6	1.5	1.6
	Sub	6.0	6.0	5.8	5.9
	WER	9.5	9.9	10.2	9.5

[1] 10xRT BN RT-03 system with new LM excluding days from bndidev03 data set.

- Increase in MS leads to increase in deletions and hence WER for STT.

BN - Are diarisation scores correlated to WER ?



- No correlation between diarisation score and WER.
- Correlation coefficient = 0.08 (-0.47 without STM point).

BN - Can we use diarisation output for STT?

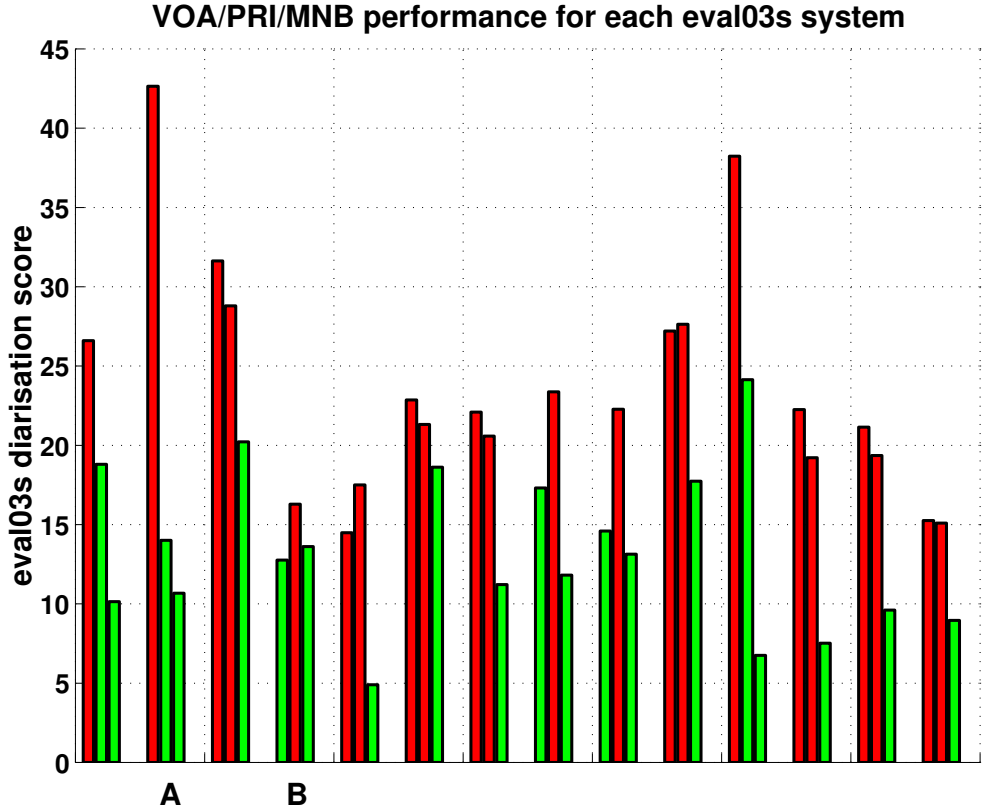
Clusters	MS	FA	SPKR	DIARY	DEL	INS	SUB	WER
CU-STT	0.2	9.1	47.2	56.48	2.0	1.6	6.0	9.5
CU-DIARY	0.2	9.1	24.0	33.29	2.1	1.5	5.9	9.5
MIT-DIARY*	2.7	5.6	16.1	24.46	2.8	1.7	6.3	10.8
DIARY-REF*	0.0	0.0	0.0	0.00	2.9	1.3	6.0	10.2
STM-REF*	1.4	9.2	0.0	10.56	2.0	1.0	5.4	8.4

* BW labels added by CUED *after* diarisation scoring.

* Some problematic segments (e.g. very short ones) were removed before recognition.

- Changing the CUED clustering made no difference to WER (!!)
- CUED segmentation was developed for the STT not diarisation task.
- Padding (and smoothing) segments might improve the results.

BN - How similar are the different systems ?



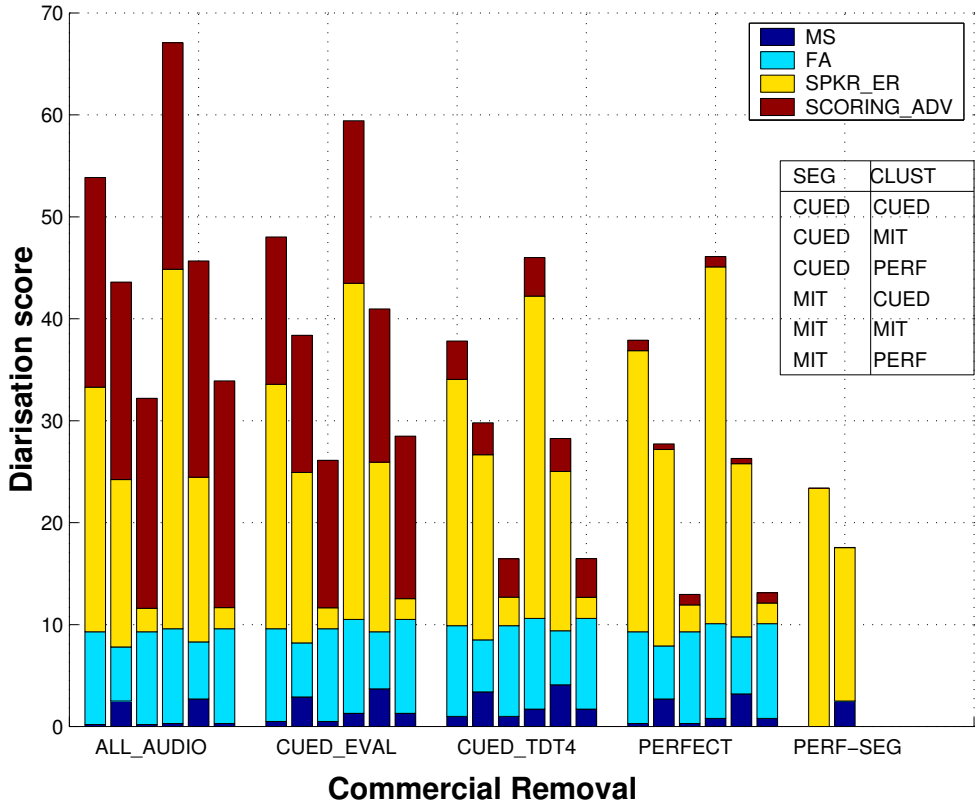
- site B performs consistently across all shows, A (=CUED) does not!
- 9 out of 14 submissions did worst on VOA data

BN - Mixing-and-matching - the details

- Removing Commercials
  - (NONE / CUED_EVAL / CUED_TDT4 / PERFECT)
  - use the output as the new UEM input to the segmentation
- Segmentation
  - (CUED / MIT / PERFECT)
  - gender-labelling and music/non-speech removal done here.
  - CUED ran a GMM bandwidth-labeller over MIT/PERF segments.
- Clustering
  - (CUED / MIT / PERFECT)
  - MIT did speech-activity gating and gender-labelling of speakers here.

Comms	Seg	Clust	Using .spkreval.uem file				Not using .spkreval.uem file				STT scores			
			MS	FA	GE	DIARY	MS	FA	GE	DIARY	Sub	Del	Ins	WER
NONE	CUED	CUED	0.2	9.1	1.9	33.29	0.2	29.7	1.9	53.86	5.9	2.1	1.5	9.5
		MIT	2.5	5.3	2.1	24.23	2.6	24.7	2.1	43.60				
		PERF	0.2	9.1	0.4	11.60	0.2	29.7	0.4	32.20				
	MIT	CUED	0.3	9.3	2.5	44.86	0.3	31.5	2.5	67.09				
		MIT	2.7	5.6	2.2	24.46	2.7	26.8	2.2	45.68	6.3	2.8	1.7	10.8
		PERF	0.3	9.3	0.6	11.67	0.3	31.5	0.6	33.91				
CUED	CUED	CUED	0.5	9.1	2.0	33.58	0.5	23.5	2.0	48.02				
EVAL		MIT	2.9	5.3	1.7	24.92	2.9	18.7	1.7	38.38				
		PERF	0.5	9.1	0.5	11.65	0.5	23.5	0.5	26.11				
	MIT	CUED	1.3	9.2	2.5	43.48	1.3	25.1	2.5	59.42				
		MIT	3.7	5.6	2.3	25.93	3.7	20.6	2.3	40.96				
		PERF	1.3	9.2	0.6	12.54	1.3	25.1	0.6	28.49				
CUED	CUED	CUED	1.0	8.9	2.3	34.06	1.0	12.6	2.3	37.82				
TDT4		MIT	3.4	5.1	1.8	26.67	3.4	8.2	1.8	29.80				
		PERF	1.0	8.9	0.8	12.69	1.0	12.6	0.8	16.46				
	MIT	CUED	1.7	8.9	2.4	42.22	1.7	12.7	2.4	46.00				
		MIT	4.1	5.3	1.7	25.02	4.1	8.5	1.7	28.26				
		PERF	1.7	8.9	0.6	12.67	1.7	12.7	0.6	16.48				
PERF	CUED	CUED	0.3	9.0	2.0	36.88	0.3	10.0	2.0	37.90				
		MIT	2.7	5.2	2.4	27.18	2.7	5.8	2.4	27.73				
		PERF	0.3	9.0	0.6	11.93	0.3	10.0	0.6	12.96				
	MIT	CUED	0.8	9.3	2.5	45.10	0.8	10.3	2.5	46.10				
		MIT	3.2	5.6	2.2	25.78	3.2	6.1	2.2	26.30				
		PERF	0.8	9.3	0.6	12.12	0.8	10.3	0.6	13.12				
	PERF	CUED	0.0	0.0	0.0	23.38	0.0	0.0	0.0	23.38				
		MIT	2.5	0.0	2.3	17.55	2.5	0.0	2.3	17.57				
		PERF	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.00	6.0	2.9	1.3	10.2
STM	STM	STM	1.4	9.2	0.0	10.56	1.4	10.2	0.0	11.61	5.4	2.0	1.0	8.4

BN - Mixing-and-matching diarisation components



BN - Mixing-and-matching diarisation components

- MIT clustering is better than CUED, and **much** more robust to changes in segmentation
  - σ^2(CUED-clust) $= 27.5, \sigma^2$(MIT-clust) $= 1.1$
  - μ (CUED-clust) $= 39.2,\ \mu$(MIT-clust) $= 25.5$

- CUED and MIT segmentation is comparable
  - average perfect-clustering score = 11.97 CUED-segs, 12.25 MIT-segs
  - average MIT-clustering score = 25.75 CUED-segs, 25.30 MIT-segs
  - MIT adds SAD-gating to reduce FA at expense of increased MS.

- Removing commercials consistently reduces the (comm=FA) score.
  - μ: NONE=46.1, CU_EVAL=40.2, CU_TDT4=29.1, PERF=27.4

Conclusions

- CTS
  - Diarisation score (and FA) can be used to help predict WER.
  - There are noticeable exceptions (e.g. using STM segmentation).
  - We found it best to smooth@0.6s and pad@0.2s before STT.
  - Using word-times and gender relabelling can improve diarisation.

- BNEWS
  - Removing commercials can speed up systems, and has little effect on the primary diarisation score, but increases MS and thus WER.
  - Diarisation score does not help predict STT performance.
  - We can learn about and improve diarisation systems by trading stages