Diarisation for RT-03s at Cambridge University

Sue Tranter, Kai Yu and the HTK STT team

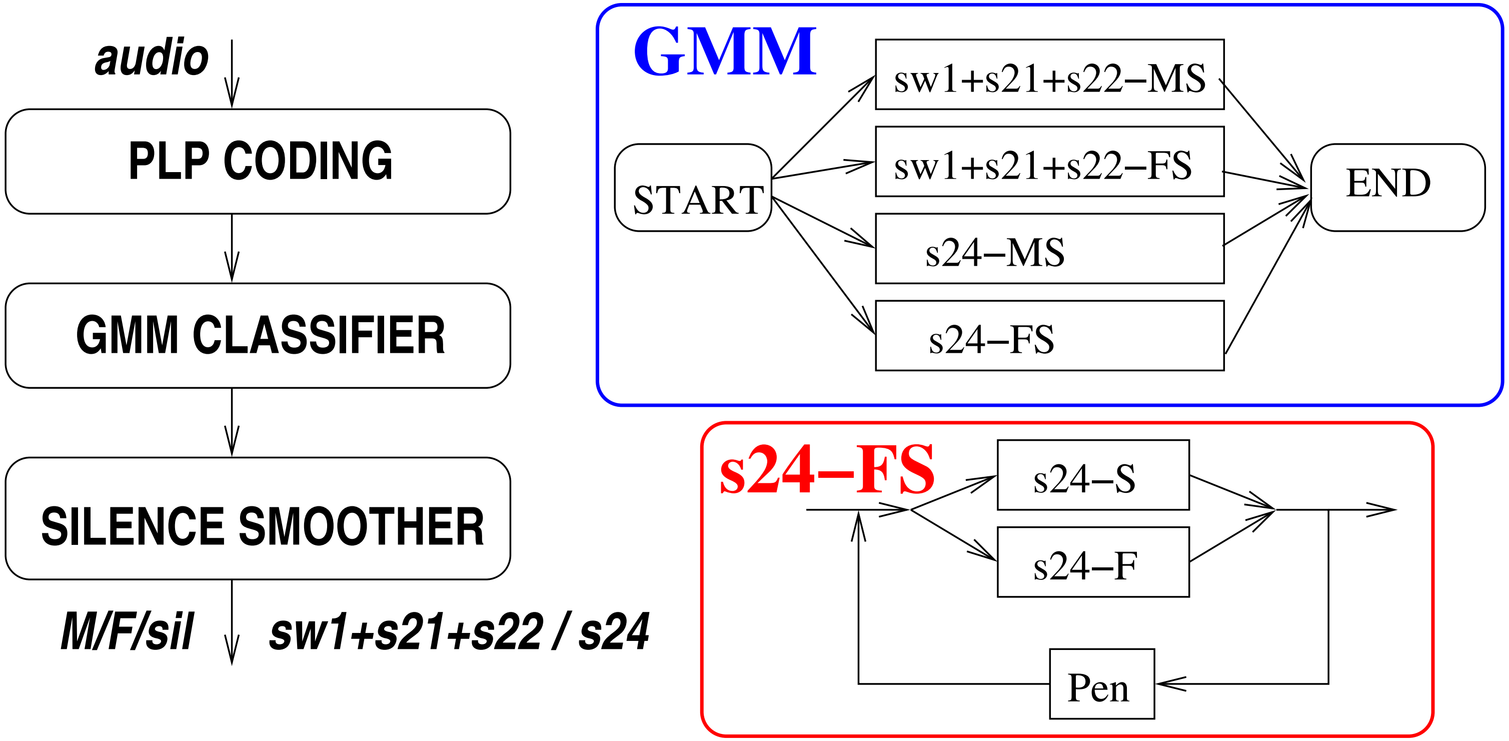
May 20th 2003

Cambridge University Engineering Department

Overview

- Diarisation for CTS
  - System Description
  - Development Work
  - Results
- Diarisation for BNEWS
  - System Description
  - Development Work
  - Results
- Conclusions

Diarisation for CTS - System Description (1)



audio

PLP CODING

GMM CLASSIFIER

SILENCE SMOOTHER

M/F/sil

sw1+s21+s22 / s24

GMM

START

sw1+s21+s22–MS

sw1+s21+s22–FS

s24–MS

s24–FS

END

s24–FS

s24–S

s24–F

Pen

Diarisation for CTS - System Description (2)

Data

sw1	SWB-I data from final MS-state transcripts of hub5train (1hr M/F/S)
s21	STT-eval97 SWB-II subset (0.63h-M/0.58h-F/1.69h-S)
s22	SWB-II phase2 Rapid transcription data from BBN (2hrs M/F/S)
s24	cell1 data from LDC transcripts of hub5train (3hrs M/F/S)

Models

Areas labelled with noise or laughter were rejected.

Portions of silence were extracted from gaps in the STM.

Phone-level forced-alignment gave areas of speech with no silence.

256 mixture GMM models built for male and female, 128 for silence.

Parameters

Insertion penalty to prevent rapid oscillation between models.

Pruning threshold to speed up search (removes unlikely paths.)

Diarisation for CTS - Improving the Results

- Better models → more mixture components.
- Better parameters → lower insertion penalty.
- Better data → add in SWB-II data, remove CHE data.
- Contrast run - Incorporate STT info → word times, gender relabelling.

Diarisation for CTS - Development Results

	DryRun[1]				eval02
	MS	FA	DIARY	WER[2]	WER[2]
CUED-dryrun	2.8	10.3	13.09	28.7	27.8
New params/models	3.0	6.3	9.27	28.3	27.4
New training data	2.2	6.3	8.55	28.2	27.3
Post-STT - 187xRT RT-03 o/p	4.0	4.1	8.05	28.2	27.2
LDC Forced-alignment†	0.9	0.5	1.48	27.7	N/A
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.

- Pre-STT diarisation score reduced by 35% relative since the dryrun.
- Diarisation score further reduced by 6% relative using STT word-times.

† This number is not that reported at the workshop, due to having 'non-lex' tokens removed.

Diarisation for CTS - Evaluation Results

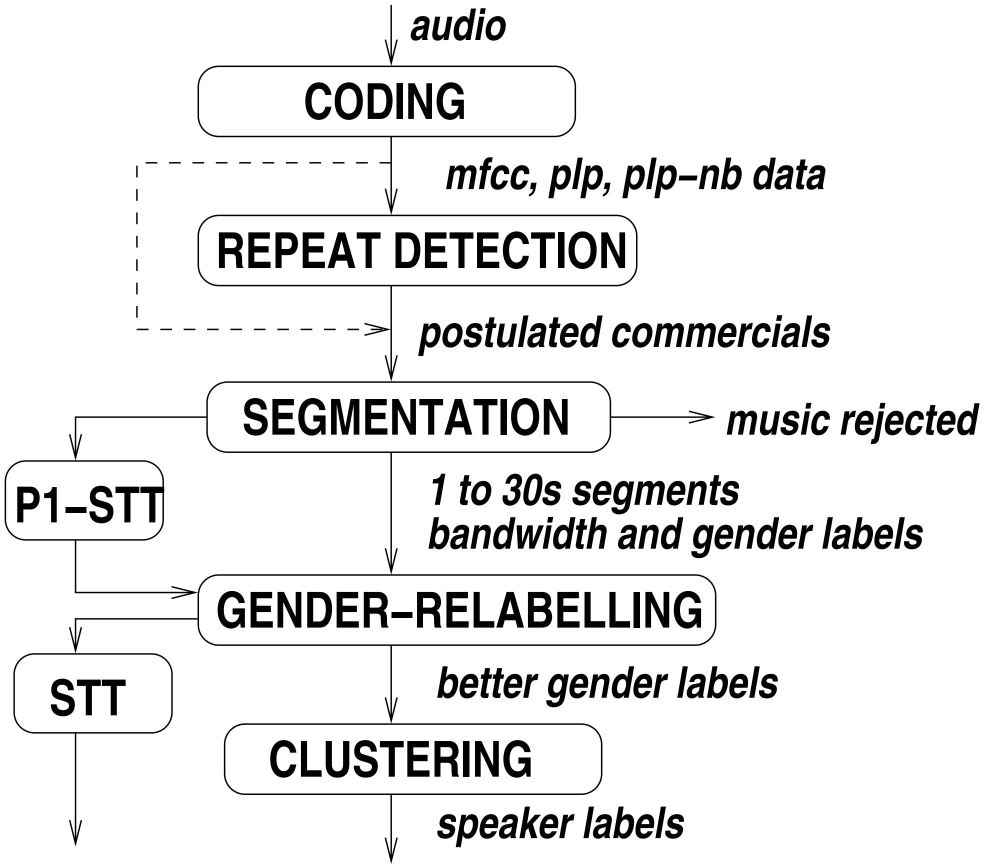
		CTS DryRun[1]			CTS-eval03s[2]			
		MS	FA	DIARY	MS	FA	DIARY	GE
Pre-STT	RT03 (0.05xRT)	2.2	6.3	8.55	8.8	1.9	10.65	1.0
Post-STT	RT03 (187xRT)	4.0	4.1	8.05	10.0	2.0	11.95	0.0

[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 STT word times did not help on the eval03s data.
- The balance between MS and FA speech has completely changed from the dev data. (Note different method of producing the reference.)

Diarisation for BNEWS - System Description



BNEWS - Postulating Commercials

- Data coded into PLP coefficients with 1st and 2nd derivatives.
- TDT-4 and eval shows split into overlapping windows 5s@1s shift.
- Windows represented by diagonal correlation matrix.
- Arithmetic Harmonic Sphericity (AHS) distance found between eval and library (TDT-4) windows and thresholded.
- Eval windows which match 2 library windows labelled as repeats.
- Smoothing between labelled repeats to give postulated commercials.
- Boundaries refined to take into account original window granularity.

BNEWS - Postulating Commercials - Results

	bndidev03 data		bneval03s data
Scheme	CU_TDT4[2]	CU_EVAL[3]	CU_TDT4
Audio Removed	18.41%	6.75%	2.55%
Commercials Removed	86.33%	28.19%	16.1%
'News' Removed	2.28%	1.66%	1.00%

[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.
- Gap between bneval03s and TDT-4 data too large to work effectively.
- Some inconsistency in the way pre-recorded announcements (e.g. 'This is the news from ABC') are transcribed in the reference.

BNEWS - Segmentation - System Description

Segmentation is based on the CUHTK Hub-4 1998 10xRT STT system. Modifications include: new music models (including TDT-4 data) and clustering/merging parameter changes to increase homogeneity of segments on bneval02 data.

- A GMM classifier divides the coded audio into wideband-speech/ telephone-speech / [music|noise] / speech + [music|noise].
- A phone recogniser is run to locate silence portions to help split these regions into smaller segments.
- Clustering and merging of similar adjacent segments is used with the GMM output to produce the final bandwidth-labelled segments.
- A first-pass STT run is aligned against GD models to determine the most likely gender of each segment.

BN - Segmentation - Results on bneval02 (=dryrun) data

Segmentation	Segments		Perfect-clustering			
	N	GE	MS	FA	DIARY	GE
dryrun (+bugfix)	248	2.4	0.1	12.8	17.90	1.5
+ new final-clustering	276	1.6	0.1	12.8	15.30	0.4
+ new music model	266	1.6	0.1	12.5	14.74	0.5
+ new smooth-clustering	282	0.7	0.1	12.5	14.31	0.7
+ new final-clustering	276	0.5	0.1	12.5	14.44	0.5

Reference derived from George's CTM times with 0.3s smoothing. Scoring also used .spkreval.uem file

- Gender Error (GE) reduced from 2.4% to 0.5%.
- Perfect clustering score reduced from 17.9% to 14.4%.
- Number of segments roughly equal.

BN - Clustering - System Description

- Clustering is done bandwidth and gender-dependently.
- Segments represented by full correlation matrix of (static only) PLPs.
- Distance metric is Arithmetic Harmonic Sphericity (AHS).
- Clustering is top down growing between 2 and 4 children at each stage.
- Stopping criteria consists of :
    - Minimum allowable occupancy constraint.
    - Gain from splitting must exceed a proportion of global cost.
    - Ratio of inter:intra child node cost must exceed threshold.
    - Special case for clusters containing a single segment.

BNEWS - Diarisation Results

System	bndidev03	bneval03s DIARY			
	DIARY	VOA	PRI	MNB	TOTAL
DryRun [STT] system (occ=25s)	64.65	-	-	-	-
bneval03 STT system (occ=40s)	54.98	-	-	-	-
best occupancy (occ=150s)	50.07	-	-	-	-
bneval03s diary system (allaudio)	33.29	42.65	14.01	10.67	23.61
ditto with CU_EVAL adv-removal	33.58	-	-	-	-
ditto with CU_TDT4 adv-removal	34.06	31.63	28.80	20.22	27.44
perfect clustering	11.60	-	-	-	-

- Clustering improved by 49% relative from STT-dryrun.
- Clustering is not very robust to changes in segmentation.

Conclusions

CTS

- Markedly different effects on dev and eval data.
- (is the accuracy of the reference a problem?).
- Pre-STT segmentation improved by 35% relative.
- Using STT word times helped diarisation on dev data.

BNEWS

- Too little eval data to draw reliable conclusions.
- Detecting commercials can help but needs (untranscribed) contemporaneous audio to work effectively.
- Clustering is too sensitive to initial segmentation.