Metadata Extraction at Cambridge University

Sue Tranter, Marcus Tomalin, Kai Yu and the HTK STT team

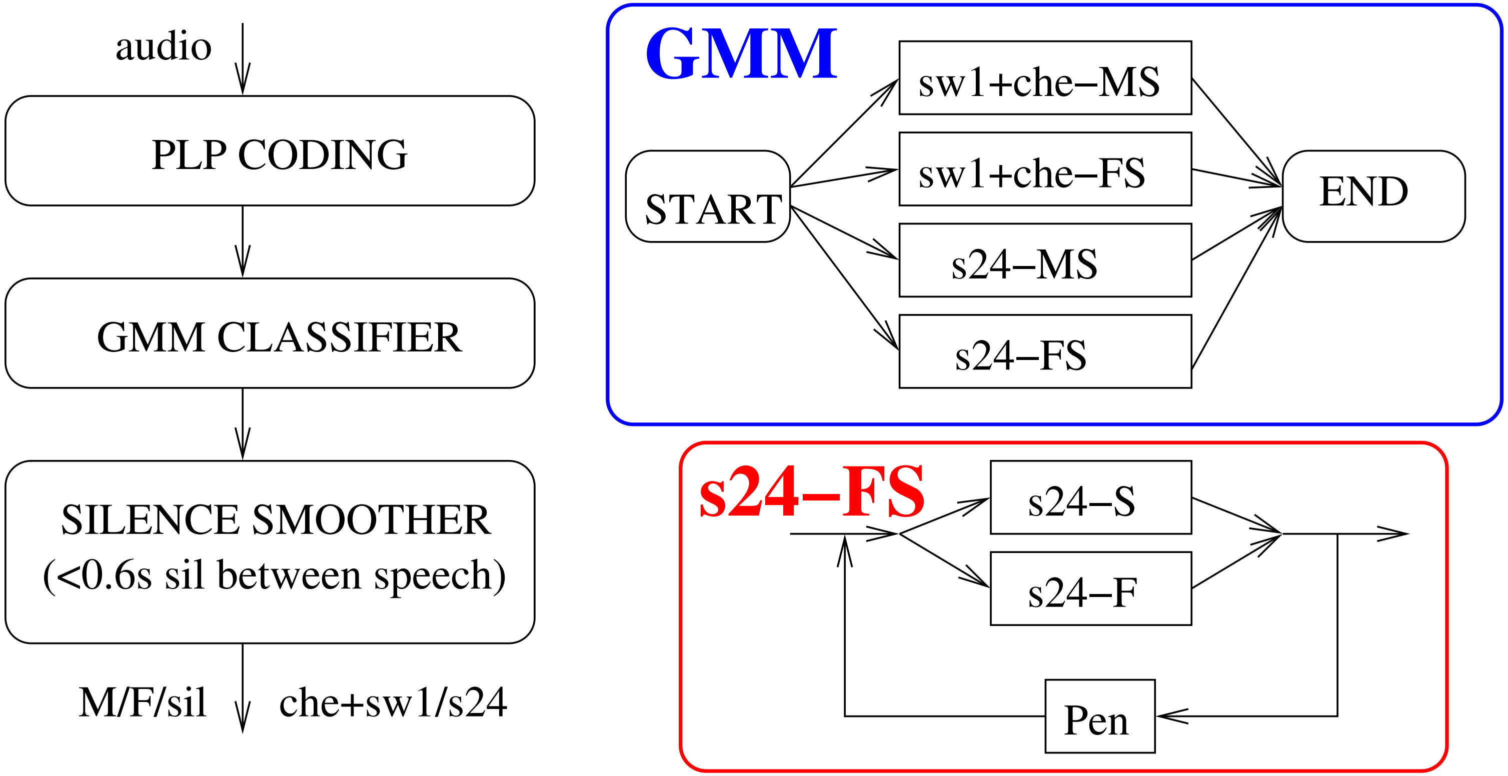
Cambridge University

January 22nd 2003

Overview

- Diarisation for CTS
- Diarisation for BNEWS
- Changes in STT output for MDE
- Disfluency labelling
- SU labelling
- Conclusions

Diarisation for CTS - System



Diarisation for CTS - Model Selection

- Final MS-State transcripts used to extract portions for silence models, and reject all areas with noise/laughter in training data.
- Phone-level forced-alignment used to extract areas of speech containing no silence (or noise).
- Simple 1-mixture Gaussian model built for male, female and silence for cell1 (s24) and for che/sw1 for 3 hour (random) subset.
- CHE data weighted by a factor of 5 in data selection due to problems with crosstalk in SW1.
- More mixtures (8) used for speech models for final submission.

Diarisation for CTS - Results

	against-nist-ref-2			against-bbn-ref			WER (10xRT)	
system	MISS SPCH	FA SPCH	Σ SPKR ERRORS	MISS SPCH	FA SPCH	Σ SPKR ERRORS	eval02 (old)	dev03 (new)
nist-ref-1	0.0	0.6	1.3	11.9	0.9	23.0	-	28.38
bbn-ref	0.5	12.0	28.2	-	-	-	-	-
mit-base	2.0	12.5	32.9	4.9	3.9	15.7	-	30.00
cu-base	2.9	2.6	12.4	12.2	0.3	22.5	-	-
cu-sub	2.2	3.4	**12.6**	10.9	0.5	20.4	28.5	**29.10**
cu-stt2	1.4	7.4	20.0	6.2	0.8	12.6	28.9	28.99
manual	-	-	-	-	-	-	26.7	29.43

- The optimum parameters for reducing WER are not the same as those for reducing segmentation errors.

Diarisation for CTS - Future Work

Things to try which *might* improve the system:

- Add models for sw2 data.
- Remove the constraint of only 1 speaker per side.
- Clean up the models (remove background speech from silence model).
- Include noise and/or laughter models.
- Incorporate information from the STT word times.
- Add echo-cancellation or similar for removing crosstalk.
- Add more mixtures and different prior probabilities.
- Add a stage to reclassify the gender, using alignments with GD models.

Diarisation for BNEWS - System

Our BN diarisation system consisted of the first two stages (segmentation and clustering) of our $<$10xRT STT system used for TREC-8. (see Refs)

- 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.
- A first-pass STT run is aligned against GD models to determine the most likely gender of each segment.
- The segments are then clustered together (subject to minimum and maximum length constraints for subsequent adaptation) using the divergence between the covariance matrices of the coded segments.

Diarisation for BNEWS - Results

	SPEECH		GENDER	SPEAKER			
	MISS	FA	ERROR	MISS	FALARM	ERROR	TOTAL
cu-stt	0.0	5.0	**0.7**	0.4	9.3	53.9	**63.6**
MIT-base	0.0	7.0	2.5	0.3	13.2	18.5	32.0

The results show that the system designed for STT speaker adaptation does not perform well for diarisation, although the gender-detection works well.

Improving the system will focus on

- Removing the occupancy constraints needed for STT
- Joining the segmentation/clustering processes into a single stage

Changes in STT output for MDE

- Phone-level alignment used to remove inter-word silences and modify end-times of words correspondingly.

- SENT_START and SENT_END tags from segmentation boundaries added under 'MISC' category.

- Fillers which were previously deleted (optionally deletable) now retained under 'FP' category.

Disfluency labelling

Define some categories and a set of rules:

```
FP  = { AH EH HA HM MM UH UM }
BC  = { HM MM-HMM OH OKAY REALLY RIGHT SURE YEAH YEP YES UH-HUH MHM UM-HMM }
DM  = { "LET'S SEE NOW" "LET'S SEE" "I MEAN" "YOU KNOW" "SEE" "SO"
        "ACTUALLY" "ANYWAY" "BASICALLY" "LIKE" "NOW" "YOU SEE" "WELL"}
EET = {"I GUESS"}

RULE1: filler = filled pause if (word == FP)
RULE2: filler = discourse_marker if (word(s) == DM)
RULE3: filler = explicit_editing_term if (words == EET)
RULE4: edit   = repetition if (word == word+1)
RULE5: edit   = repetition if ((word word+1) == (word+2 word+3))
```

Disfluency Labelling with Context

```
DIG   = {ONE TWO THREE FOUR FIVE SIX SEVEN EIGHT NINE ZERO OH}
LET   = {A. B. C. D. E. F. G. H. I. J. K. L. M. N. O. .... Z.}
RE    = {REALLY VERY HAD GREAT $DIG $LET $FP $BC}
DM_NL = {[I|YOU|WE|THEY]-LIKE}
DM_NR = {LIKE+[THIS|THAT|ME|YOU|HER|HIM|IT|US]}
DM_NL = {[RIGHT]-NOW}
DM_NR = {SO+[THAT|THEN]}

RULE4/5c: NO repetition if (word == RE)
RULE2c:   NO discourse_marker if ((word(s)+context) == (DM_NL || DM_NR))
```

Disfluency Results

System (Context ?)	Edit (BN=84,CTS=521) Miss	FA	Error	Filler(BN=143,CTS=840) Miss	FA	Error	Total Error
BN-ASR (×)	91.67	28.57	120.24	48.95	87.41	136.36	130.40
BN-ASR (✓)	91.67	15.48	107.14	48.95	82.52	131.47	122.47
BN-REF (×)	83.33	20.24	103.57	6.29	66.43	72.73	84.14
BN-REF (✓)	83.33	7.14	90.48	6.29	62.24	68.53	76.65
CTS-ASR (×)	90.60	10.17	100.77	27.26	55.95	83.21	89.93
CTS-ASR (✓)	91.17	9.02	100.19	29.76	47.02	76.79	85.75
CTS-REF (×)	86.95	13.82	100.77	10.36	41.90	52.26	70.83
CTS-REF (✓)	87.14	12.28	99.42	11.43	33.33	44.76	65.69

- Performance is much better for CTS than BN, and reference than ASR.
- Adding context reduces the error by 6.7% relative on average.

SU Labelling

N : gap of N seconds in transcriptions $\rightarrow$ SU
SENT: SENT_START or SENT_END tag in ASR output $\rightarrow$ SU

Classify SU type using the following linguistic groups and rules:

```
QUES     = {WHAT WHY WHERE WHEN HOW DO ARE IS HAVE DID HAS REALLY}
CO-CONJ = {AND BUT OR}
SUB-CONJ= {IF HOWEVER THEREFORE}
ART      = {THE A AN}
QUANT    = {ANY ALL MOST EVERY}
INCOMP   = {$CO-CONJ $SUB-CONJ $ART $QUANT}

RULE1: su = question if ( su-initial word == QUES )
RULE2: su = incomplete if ( su-final word == INCOMP )
RULE3: su = backchannel if ( su == BC+ )
RULE4: su = statement if (su not already classified)
```

* SU Results

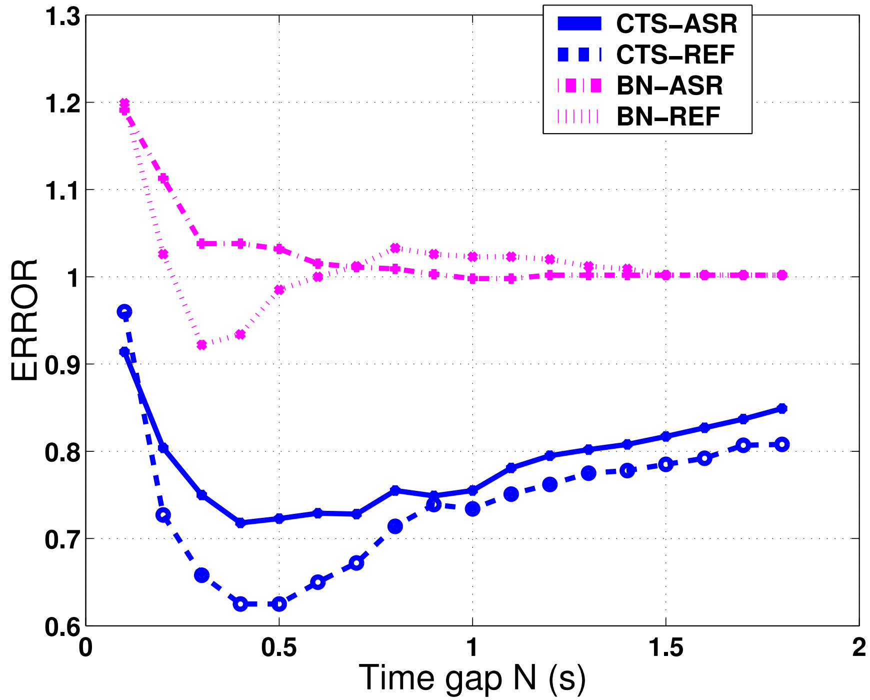
	N(s)	Extent,Type ✓,✓	✓,×	×,✓	×,×	Unmapped Hyp	Ref	%Error
*BN-REF	0.4	98	7	118	4	184	426	93.4
*BN-ASR	SENT	384	29	69	7	458	164	95.3
*BN-ASR	0.4	6	2	17	0	50	628	103.8
*CTS-REF	0.4	442	235	292	80	492	438	62.5
*CTS-ASR	SENT	496	170	124	37	305	660	64.9
*CTS-ASR	0.4	414	132	89	33	248	819	71.8

- The SENT method is better than the method using N=0.4s
- The best results come on the CTS REF transcripts.

(NB CTS-ASR may have benefited from using the manual segmentation)

* These results are slightly different to those presented at the workshop. A more recent reference (dated 20030114) was used for scoring, and minor bug fixes relating to the first and last SU in a file and treating SENT_START and SENT_END tokens as outside SUs were made.

* SU Results - changing N



- The best N is 0.3s (BN-REF) 0.4s (CTS-REF/ASR).
- This method doesn't work well on the (fluent) BN speech, but works significantly better on the (disfluent) CTS data.

Conclusions

- Metadata Research is an interesting topic which is still finding its feet.
- Diarisation requires accurately defined reference data.
- CTS diarisation should benefit from improving models, reducing noise, allowing multiple speakers per side, and eliminating cross-talk.
- BN Diarisation is much harder. ASR speaker-adaptation systems have potential for significant improvement for the diarisation task.
- A simple rule-based system can be used to try to identify disfluencies and slash units. Automatic rule-learning will improve this method.
- Integration throughout the system should help improve performance, e.g. performing acoustic segmentation and clustering simultaneously, using word-times to modify speaker boundaries, or using acoustic phenomena and lingustic patterns to help recognise slash units.

References for Scoring

$NIST = ftp://jaguar.ncsl.nist.gov/rt/rt03/
$EARS = http://macears.ll.mit.edu/

Segmentation	
UEM defining data	$NIST/rt-03-dry-run-indices.20021206.tar.gz
NIST-ref-1 for CTS	$NIST/rt-03-dry-run-reference-expt-data.20021216.tar.Z
	$EARS/macears_mail/0534.html
NIST-ref-2 for CTS	$NIST/DryRunResults.20030114.b.tgz
BBN-ref for CTS	$EARS/macears_docs/volunteer-dev-data/hub5-bbn-v01.tgz
BBN-sub for CTS	$EARS/macears_docs/rt03-dry-run/bbn-rt03-early-dry-run.tgz
Scoring for CTS	$EARS/macears_docs/eval/SpkrSegEval-v11.pl
Scoring for BN	$EARS/macears_docs/eval/SpkrSegEval-v13.pl
Reference for BN	$NIST/DryRunResults.20030114.b.tgz
MIT-Baseline	$NIST/mitllbase-dry-run-segmentations.20021206.tar.Z
Structural MDE	
*Scoring AIF for SU	$NIST/DryRunResults.20030114.tgz
Scoring AIF for DISFL	$NIST/DryRunResults.20030114.tgz

References for BN Segmentation System

S.E. Johnson , P. Jourlin, K. Spärck Jones & P.C. Woodland
Spoken Document Retrieval for TREC-8 at Cambridge University
Proc. TREC-8, NIST SP 500-246, pp. 197-206 (2000)
`http://svr-www.eng.cam.ac.uk/reports/full_html/johnson_trec8.html/`

T.Hain, S.E.Johnson, A.Tuerk, P.C.Woodland & S.J.Young
Segment Generation & Clustering in the HTK Broadcast News Transcription System
Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, pp. 133-137 (Lansdowne, VA, Feb. 1998)
`http://svr-www.eng.cam.ac.uk/reports/full_html/hain_darpa98.html/`