Cambridge STT Overview

P.C. Woodland, H.Y. Chan, G. Evermann, M.J.F. Gales, T. Hain,
B. Jia, D-Y. Kim, X. Liu, D. Mrva, K.C. Sim, S.E. Tranter, L. Wang

February 4th 2004

Cambridge University Engineering Department

Outline

- Broadcast News
- Lightly supervised discriminative training
- Training on Fisher Data
- Mandarin
- HTK
- PhD Research projects
  - Discriminative SAT
  - Structured Transforms
  - Model Complexity Control

Broadcast News: Segmentation and Clustering

- Running our evaluation system on different segmentations and clusters
- STM-based manual seg/clust was 0.8% abs better than our automatic seg/clust
- Potential for improvement in segmentation rather than clustering

SEG	CLUST	sub(%)	del(%)	ins(%)	WER(%)
STM	STM	6.7	1.9	1.2	9.8
CU	CU	7.0	2.2	1.4	10.6
STM	CU	6.8	1.9	1.1	9.8
CU	STM*	6.7	1.9	1.2	9.8

2003 CU-HTK 10xRT with minor fixes on `eval03` for various seg/clust.
* assign speaker labels to produce maximum overlap with the ref speakers.

- Modified clustering yielded better diarisation score but did not improve WER

Broadcast News VTLN Experiments

- VTLN is done at cluster-level (with min.occ.=500 frames)
- Cluster-based mean and variance normalisation was used
- ML search for warp-factor done using parabolic search in the range [0.80,1.20]
- The warp-factors for test-data are estimated using the VTLN-MLE model.

	dev03		eval03	
	Baseline*	VTLN	Baseline*	VTLN
MLE	19.7	18.4	17.8	16.5
HLDA	17.9	16.9	15.9	14.9
HLDA+MPE	15.2	14.6	13.7	13.2
HLDA+MPE-MAP (GD)	14.9	14.5	13.4	13.0

%WER with VTLN models (tg LM). * Acoustic models with segment-based CMN.
All narrow-band results are borrowed from the corresponding baseline results.

- Gains reduced after adaptation

Lightly supervised discriminative training on TDT data

- Improve the English Broadcast News system by adding large amounts of TDT data
  - 144 hours of accurately transcribed data
  - 370 hours of wideband TDT2 data
  - 230 hours of TDT4 data
  - Only closed-caption transcripts are available for TDT data
- Lightly supervised discriminative training
  - Construct biased language model
  - Automatically recognize the TDT data with a 5xRT P1-P2 CU-HTK system
  - Use all the recognized transcripts for ML and MPE training
  - Compare with closed-captions filtering (BBN/LIMSI approach)

CU-HTK P1-P2 System WER - dev03/eval03

Acoustic model	dev03		eval03	
	P1	P2	P1	P2
bnac (144h)	16.2	12.5	14.8	11.5
bnac+370h wb TDT2	15.1	11.9	14.0	11.3
bnac+230h TDT4	14.5	11.8	13.6	10.9
bnac+370 wb TDT2 +230h TDT4	14.5	11.4	13.3	10.6

MPE training, 4-gram LM, adapted

- Adding TDT data reduce recognition WER
- Adding 600h TDT data (370h wb TDT2, 230h TDT4) to bnac
  - 1.1% (dev03) and 0.9% (eval03) WER reduction in P2 output
- Full CU-HTK 2003 10xRT system WER: 11.6% (dev03) and 10.7% (eval03)

Data selection: unadapted single pass decoding - dev03

Wide-band data	ML	MPE
bnac (144h)	17.8	15.0
bnac+80h TDT4 CC match	17.0	14.4
bnac+115h TDT4 CC match	16.9	14.2
bnac+115h TDT4 CC mismatch	17.1	14.3
bnac+115h TDT4 random	16.9	14.3
bnac+230h TDT4	16.8	13.8

ML/MPE training, trigram LM, unadapted

- Closed-captions filtering removes large amount of data (only 80h remains)
- Minor difference in performance between CC match , CC mismatch or random selection in the three 115h TDT4 data sets
- Using all the recognized transcripts is the best for MPE

CTS segmentation and WER

- What is the 'best possible' segmentation in terms of minimising WER?
- What do we lose in WER from doing segmentation automatically?

	Recogniser	Dec 2002 Dryrun 10xRT			RT-03 10xRT
	System	dry03	eval02	eval03	eval03
Auto	CUED Pre-ASR	28.1	27.3	26.3	22.2
Auto	CUED Post-ASR-187xRT	28.2	27.1	26.0	22.0
Ref	Manual word times	27.8	—	—	—
Ref	STM (unknown smth/pad)	27.7	26.7	25.6	21.6
Ref	CUED FA word times	27.4	26.2	25.4	21.3

- Best WER with segments from CUED FA of reference. (BBN found similar)
- Around 1% absolute WER degredation for automatic (Pre-ASR) segmentation.
- Diarisation score/WER highly correlated if reference generated appropriately

Experiments with Fisher Data

- Initial experiments on using large amounts of Fisher data

- Acoustic training data

  h5train03b 360h data set. 290h LDC data with MSU careful transcriptions. 70h BBN data with quick transcriptions
  fisher3896 520h Fisher data set, 3896 conversations
  fisher3896+h5 880h data set, the combined set of h5etrain03b and fisher3896

- Fisher data processing
  - Normalize the text, joining, padding
  - Apply about 2000 replacement rules (Abbreviations, typos, non-speech, ...)
  - Produce pronunciations for unknown words with frequency greater than 2
  - aligning the segments and fixing silence boundaries

Acoustic modelling, Language modelling and Testing

- Acoustic model
  - cross-word triphone, 6200 tied states, vtln, HLDA front-end
  - 28 variable Gaussian mixture components per state
  - Gender Independent MPE models
- Language model
  - LM03: LMs/training texts used for 2003 eval
  - LM03+Fsh3896: LM03 + Fisher3896
  - Built separate LMs for each component data source, then interpolate/merge
  - Full models also interpolate with 03 eval class-based model (no Fisher data)
- CU-HTK P1-P2 system
  - P1, P2 architecture of CU-HTK 2003 10xRT evaluation system
  - Trigram decoding, fourgram lattice rescoring
  - overall $\sim$ 5xRT include adaptation

Eval03 with CU-HTK P1-P2 System

		Overall	Swbd	Fisher	Male	Female
h5train03b	LM03	24.6	28.7	20.2	25.7	23.5
h5train03b	LM03+Fsh3896	23.9	28.2	19.3	25.0	22.8
fisher3896	LM03+Fsh3896	23.1	27.0	18.9	24.6	21.6
fisher3896+h5	LM03+Fsh3896	22.7	26.6	18.5	24.2	21.1

MPE training, eval03, 4-gram LM, adapted

- h5train03b: compare with using LM03, using LM03+Fsh3896 gives 0.7% overall improvement
- fisher3896: performs 0.8% better than h5train03b (LM03+Fsh3896)
- fisher3896+h5: performs 0.4% better than fisher3896 (with LM03+Fsh3896)
- Total 1.9% overall improvement by adding fisher3896 to h5train03b for both acoustic model and LM training

Comparing CTS Quick Transcription Approaches

- 20h of Swbd1 data with several transcriptions

- Acoustic models were trained for each of the transcriptions

	dev01		eval03	
	MLE	MPE	MLE	MPE
MSU	43.4	40.5	43.5	40.5
LDC QT	43.6	41.2	43.8	41.2
BBN WWave1	43.6	41.2	44.0	41.4
BBN WWave3	43.4	40.8	43.6	40.8

%WER, unadapted, tg LM, MLE and MPE (6it) acoustic models

- Discriminative training more sensitive to transcription differences

- Only small gap from BBN WWave3 to MSU transcripts

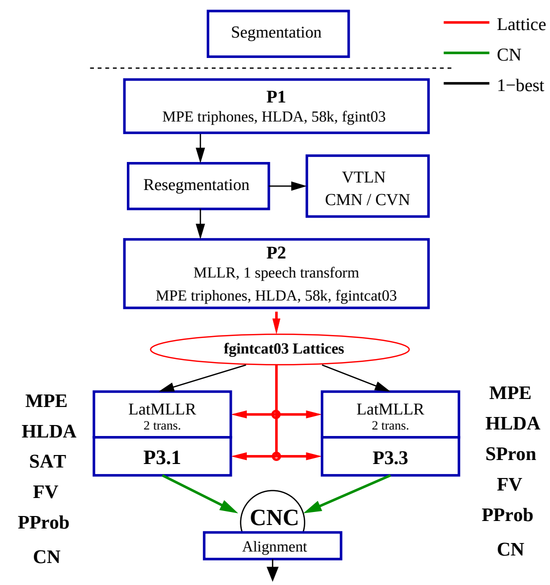
Mandarin CTS Progress

- CER was improved by 4.1% absolute since RT-03
- Improvement mainly due to:
  - Multiple rescoring branches and system combination
  - Fixed problems in training data setup
- Other issues investigated (without significant WER gains):
  - character-to-word segmentor
  - pronunciation variants
  - pitch smoothing
  - HLDA-SAT
  - full/block transform for pitch in adaptation

Development of Fast LVCSR Systems

In 10xRT or faster systems, the structure has to be designed & tuned carefully

- General system structure:
  - generate lattices with adapted models
  - rescore with multiple models
  - system combination



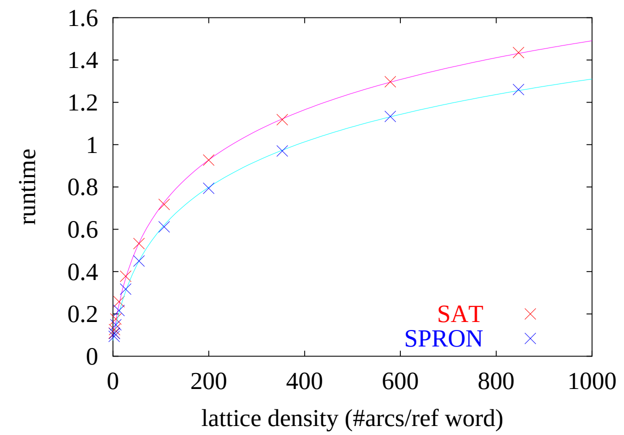
- Size of lattices has impact on speed (both generation and rescoring)

- System Combination is expensive but gives extra accuracy and robustness

Predicting Rescoring Time

Prune large lattices at successively tighter thresholds and rescore:

- Runtime can be predicted from lattice size
- Runtime grows logarithmically in lattice density
- SPron is consistently faster than SAT



Runtime vs. lattice density (CTS)

Pruning Rescoring Branches

- Often system combination doesn't change result relative to first branch $\Rightarrow$ could skip later rescoring branches
- Predict these segments; best features: min. confidence score, segment length
- Results on cts-eval03: skip 66% segments, $\Rightarrow < 0.1\%$ WER change

HTK Software Development

Aims:

- Technology Transfer: Document details of techniques used in CU-HTK systems
- Allow smaller groups to work on LVCSR

Status & Progress since RT-03:

- Public web site: `http://htk.eng.cam.ac.uk` (>20,000 registered users)
- Active discussion on support mailing lists about HTK and ASR in general
- Release of new public version (3.2.1) with many code fixes, many in response to user feedback
- Improved new adaptation framework about to be released

Discriminative Adaptation & Adaptive Training

- Investigate linear transform parameter estimation for:
  - Adaptive training
  - Unsupervised/supervised adaptation

- Discriminative speaker adaptive training.
  - Use consistent MPE/MMI criterion for linear transform generation & canonical model re-estimation.
  - Only gives very small improvement over using ML-estimated transforms

- Discriminative speaker adaptation
  - Use MPE/MMI optimization
  - Unsupervised adaptation with confidence scores

Unsupervised Adaptation with DLT: CTS Results

Adaptation	rescoring	+CN decoding
lattice MLLR	27.5	27.0
MLLR	27.7	27.0
MMI-DLT	27.5	26.8
MPE-DLT	27.3	26.9
MMI-DLT(conf)	27.3	26.6
MPE-DLT(conf)	27.1	26.7

%WER on *dev01sub* for MPE system.

- Supervision: outputs after lattice MLLR adaptation and CN decoding
- Confidence scores: from CN decoding
- MMI/MPE-DLT: 2 transforms
- MMI/MPE-DLT get 0.3%-0.4% gains after CN decoding over the supervision.

Structured Transforms

- Found data may be highly non-homogeneous:
  - multiple acoustic factors (e.g. gender/channel/style);
  - effects on acoustic signal of each factor varies;
- Multiple transforms:
  - a separate transform for each kind of unwanted variability;
  - nature of transform (should) reflect factor;
  - (possibly) more compact systems.
- Form examined in this work:
  - constrained MLLR (CMLLR) transforms;
  - interpolation weights in cluster adaptive training (CAT);
  - no explicit association of transform with factor.

Structured Transforms: Initial Results on CTS

- Initialisation of parameters:
  - Interpolation weights initialised using gender information;
  - CMLLR transforms initialised to identity transforms.

System	Training Adaptation	Test Adaptation	Estimation MLE	Estimation MPE
GI	—	—	33.4	30.4
		CMLLR	31.5	28.3
SAT	CMLLR	CMLLR	31.0	27.8
ST	ST	ST	30.6	27.3

%WER on dev01sub (3h), trained on h5train03 (290h), 16comp, HLDA, 4it. MPE

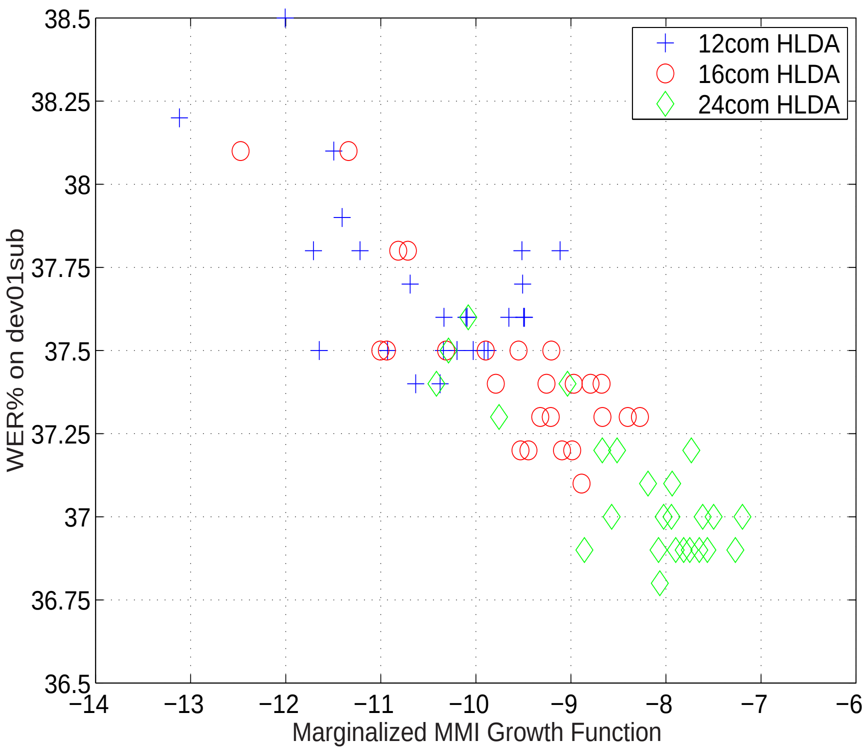
- ST refers to structured transforms (CAT+CMLLR)
- SAT refers to speaker adaptive training (CMLLR)
- Adaptive training with ST significantly outperforms SAT

Automatic model complexity control (1)

- LVCSR systems are highly complex, automatic criteria needed to quickly optimize complexity and minimize word error for unseen data.

- Bayesian likelihood based schemes unsuitable, weak correlation with WER.

- Discriminative criteria strongly related with recognition error, marginalized as complexity control schemes.



Marginalized MMI GFunc vs. WER

Automatic model complexity control (2)

Complexity Control	#Gauss	#Trans	#Dim	dev01sub WER% MLE	MPE	MLLR
Std	28	—	39	34.7	—	—
Fixed	28	1	39	33.4	30.1	28.5
			52	33.2	—	—
Fixed	28	65	39	33.3	29.8	28.4
			52	32.9	—	—
GFunc	25.6	65	41.5	32.7	29.6	28.0

Optimizing #Gaussians and retained HLDA dimensionality on 296 hour CTS `h5etrain03`

- 20% parameters reduction
- 0.5%∼0.7% abs improvement over global HLDA system.
- Gain retained after discriminative training and adaptation.

Conclusions

- Making progress on many fronts
- Use of large data resources with low-cost transcriptions works!
  - much more data becoming available
  - trying to find best ways to exploit it
  - costly to do experiments!
- Segmentation still not a solved issue
- Other PhD work hope to integrate into RT04 systems
  - complexity control
  - discriminative adaptation
  - structured transforms
  - precision matrix superposition
  - implicit-topic language models