EPSRC
Engineering and Physical Sciences Research Council

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Highlights from NST Systems for the MGB Challenge

Phil Woodland (Cambridge)

+ NST MGB Teams @
Cambridge, Edinburgh, Sheffield

http://www.natural-speech-technology.org

Overview

Natural Speech Technology

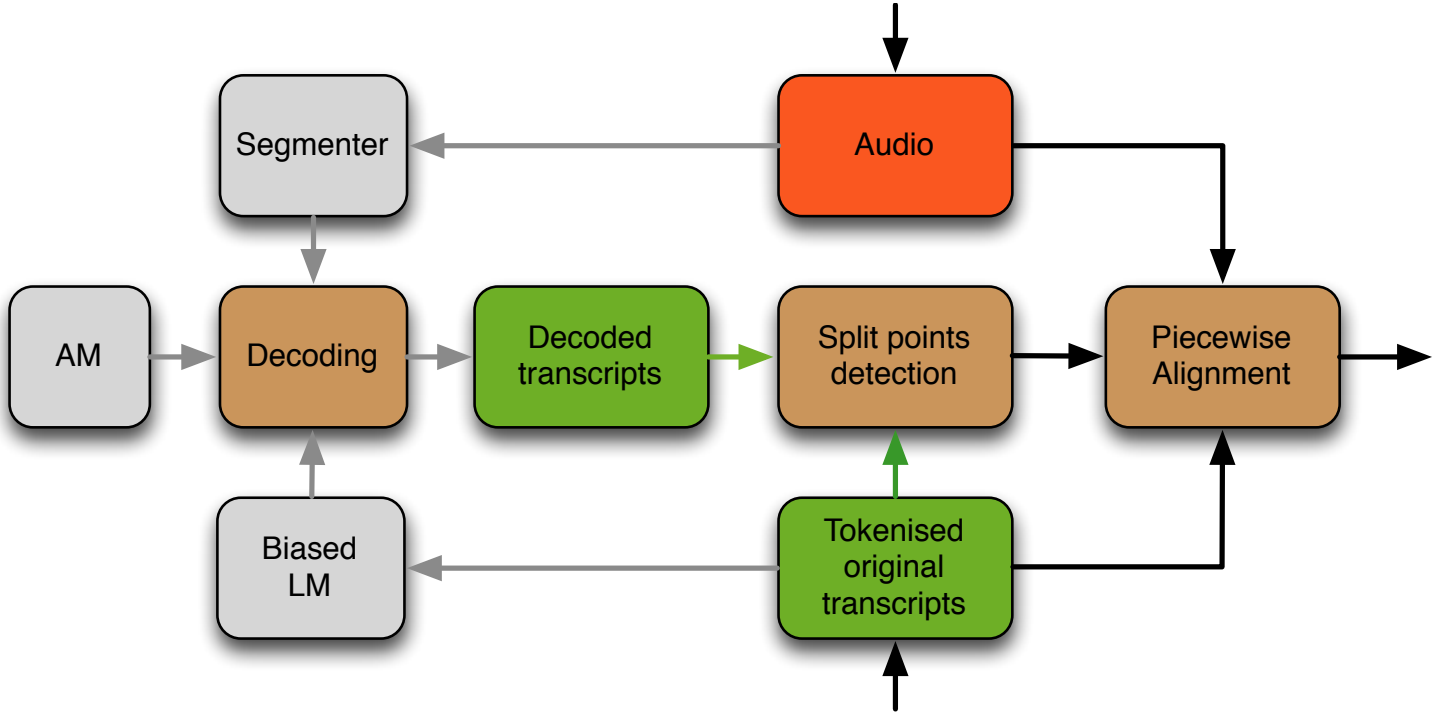
Edinburgh – Cambridge – Sheffield

- Lightly Supervised Alignment
  - reprocessing training data
- DNN-based Segmentation Methods
- Transcription
  - Multiple Acoustic Model Types & Acoustic Adaptation
  - RNNLMs and LM adaptation
  - Combination approaches
- Approaches to alignment
  - use of biased language model vs flexible alignment with WFST
- Diarisation Highlights

Lightly Supervised Alignment

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Reduce the text/audio alignment of complete shows (>6h) to the alignment of small segments;
- original and lightly supervised decoded transcripts are compared to detect reliable split points;
- improved acoustic models (AM) and segmenter can be trained on data considering the obtained alignment and confidence measures

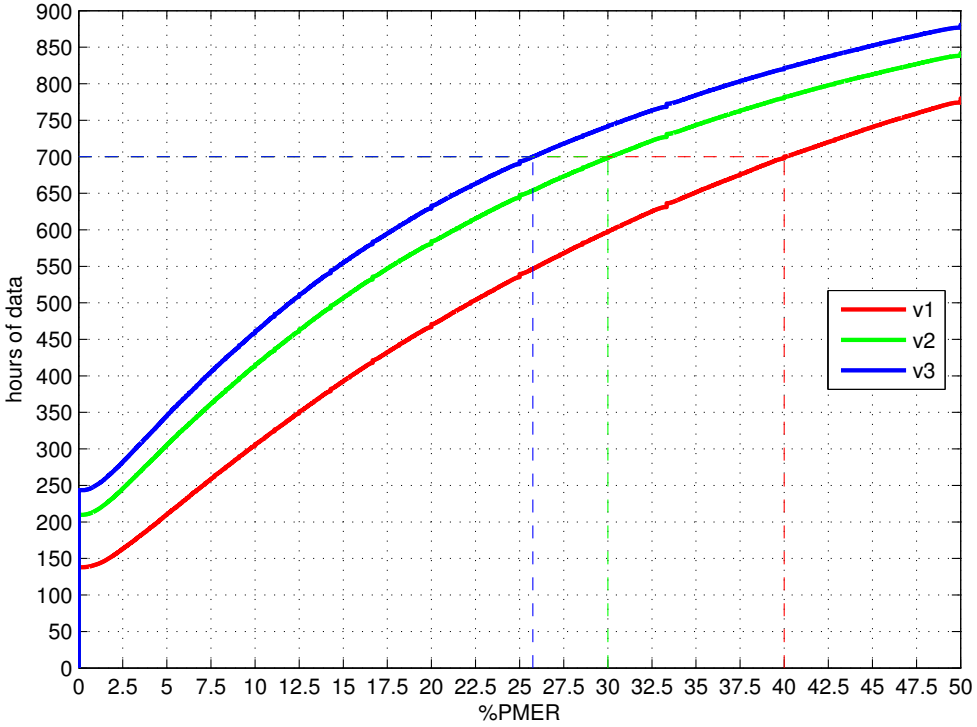


Training Data 700h-v2

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Re-run lightly supervised decoding on complete training set with
  improved DNN-based segmenter;
  better acoustic models (hybrid 700h-v1)
  increased LM bias (show)

System	LM	segmenter	%WER(del/ins)
T200	week biased	base-seg	35.0(16.9/4.5)
H200.v1	week biased	base-seg	29.1(16.4/3.3)
H200.v1	episode biased	base-seg	26.9(16.2/3.1)
H200.v1	episode biased	DNN-seg.v1	23.1(10.7/3.4)
H700.v1	episode biased	DNN-seg.v1	22.1(9.3/4.0)

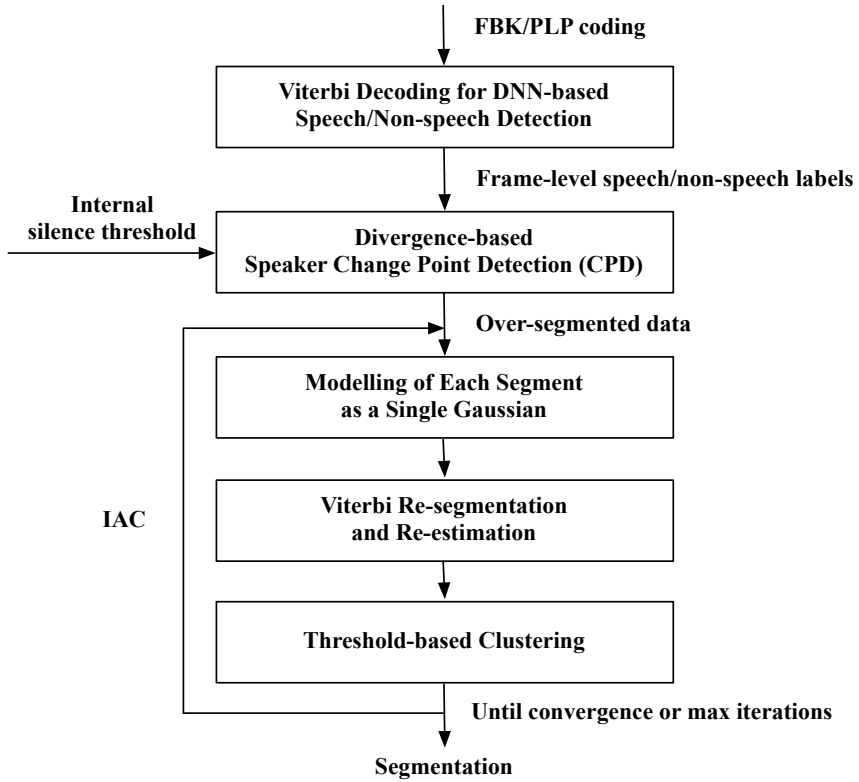


- v2 has PMER=30 for 700h (cf PMER=40 for v1).
- HTK MGB eval Cambridge models and segmenter trained on 700h-v2
- Can repeat for v3, but no improvement in trained models (but can train on ASR output)

DNN-Based Segmentation

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- About 4% absolute increase in WER if use baseline segmenter. (3.6% MS, 3.7% FA)
- New Speech/Non-speech DNN trained on PMER=0 v2 data
  - Optimised large window size (55 frames) and architecture.
  - Speech/Non-Speech followed by change-point detection and clustering to form segments
  - Resulting segmenter reduces WER by 1.9% absolute (2.5% MS, 1.9% FA)
  - Used for transcription, alignment & diarisation



Cambridge Acoustic Models

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Cambridge System tried to include a diverse set of acoustic models trained using both HTK and Kaldi
- HTK-based models:
  - sequence trained standard DNN hybrid
  - high performance Tandem system trained on improved DNN features
  - Tandem/Hybrid models combined in joint decoding (log-linear state level)
  - also adapted ReLU hybrid combined via CNC
- 1.3% abs reduction in WER over sequence-trained hybrid from joint decoding

System	Criterion	%WER
SI Hybrid	CE	28.4
SI Hybrid	MPE	25.9
SI Tandem	MPE	27.0
Joint: Tandem $\otimes$ Hybrid	MPE	24.6

Table 2: % WER on dev.full. $\text{LM1}_{\text{prune}}$, manual segmentation

Cambridge Kaldi-based Acoustic Models

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Kaldi acoustic models included to include more model types/diversity (all sequence trained)
  - DNN: std DNN
  - CNN: convolutional neural network - best individual model
  - LSTM: long short-term memory recurrent network (combines well with others)
  - For MGB evaluation trained on 500h v1 data (LSTM 250h). Since trained on 700h-v2
- Kaldi when combined with best HTK models (include adaptation)
  - reduce error by 0.5-0.6% abs in evaluation setup
  - reduces error by 1.1-1.2% abs with revised models
  - Combination with HTK uses common RNN language model and CNC

500h/250h-v1 system	WER	700h-v2 system	WER
CNN (K00)	26.4	CNN (K10)	25.4
DNN (K01)	27.7	DNN (K11)	26.4
LSTM (K03)	31.1	LSTM (K12)	26.8
K00:K01	26.0	K10:K11	24.9
K00:K01:K03	**25.7**	K10:K11:K12	**23.7**

Table 4: WER (%) for the Kaldi CNN, DNN and LSTM systems & MBR combination represented by ":" (Auto segmentation, LM2).

Cambridge Acoustic Models: Adaptation

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Various types of unsupervised acoustic adaptation used:
  - CMLLR adapted features at input for Tandem/SAT models/stacked models
  - i-vector adaptation
  - p-ReLU adaptation (alters slope of activation function)

- p-ReLU adapted models
  - applied layer-by-layer
  - **reduces WER by 1.1%** over CMLLR
  - combines well with other systems

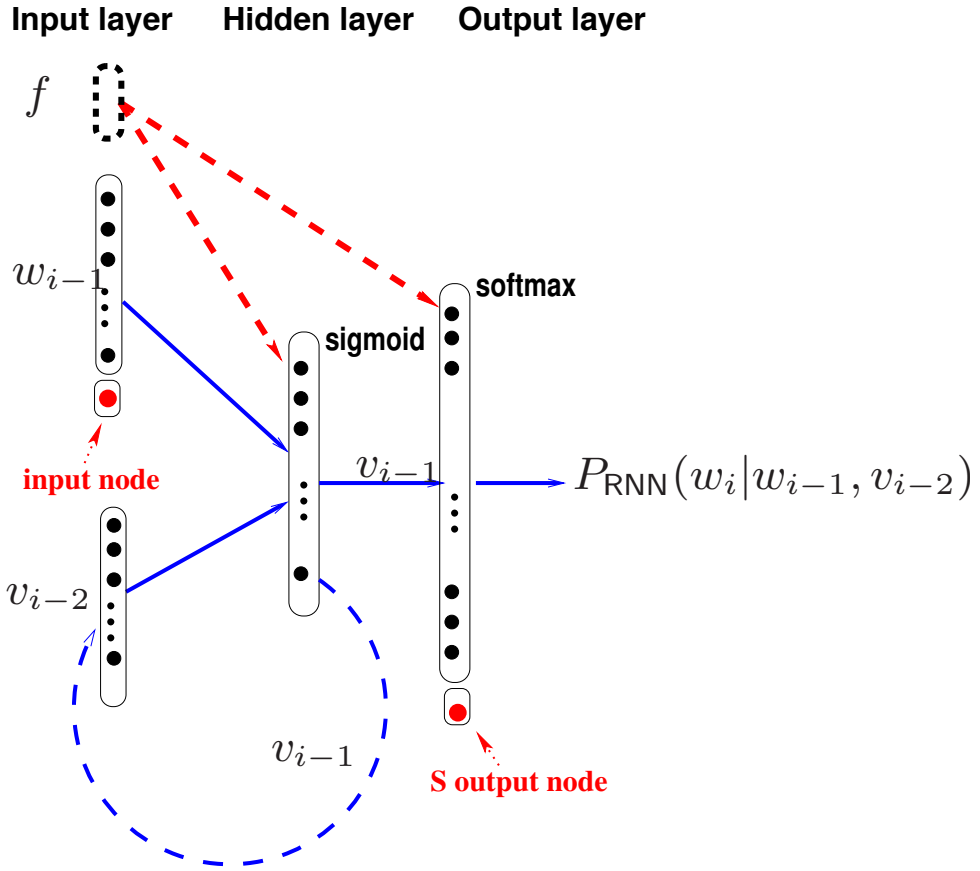
Input Transform	$p-$ReLU Adaptation	%WER
CMLLR	None	25.9
CMLLR	Bottom Layer	25.5
CMLLR	Bottom 3 Layers	25.0
CMLLR	Bottom 5 Layers	24.8

Table 5: %WER of 700h-v2 SA stacked hybrid system on dev.full. Automatic seg, 160k $LM2_{prune}$

Cambridge RNN Language Models

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- RNNLM trained on complete MGB training corpus
  - uses CUED-RNNLM GPU training
  - lattice rescoring from n-gram
  - Topic adaptation via latent Dirichlet allocation at input



AM	LM	dev.full	
		PPlex	%WER
700hr-v1 MPE hybrid	LM1	103.1	25.6
	LM1+RNN512	93.0	25.0
	LM1+RNN512.lda	85.1	24.7
	LM1+RNN1024.lda	81.0	24.4
700hr-v1 MPE hybrid	LM2	108.6	24.9
	LM2+RNN1024.lda	85.7	23.7

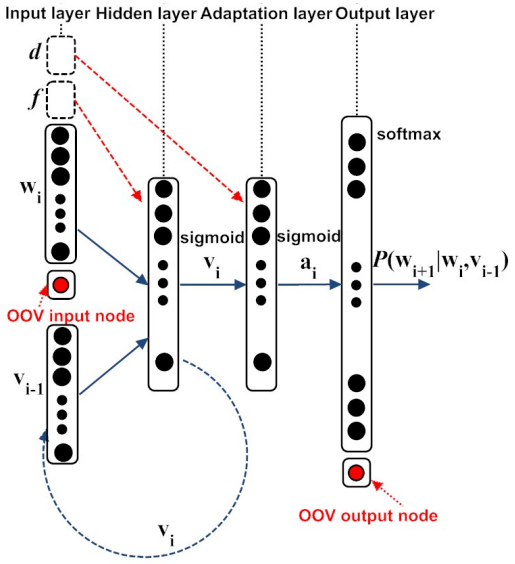
Table 3: 700h-v1 MPE hybrid acoustic models, manual segments

- **1.2 % abs reduction in WER** from topic-adapted RNNLM

RNNLM Adaptation (Sheffield)

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- RNNLM hybrid adaptation
  - Latent topics used as auxiliary features in the input layer
  - Linear Hidden Network (LHN) adaptation based on the genre labels



	WER
n–gram	30.1%
RNNLM	29.2%
LDA features	28.7%
Genre LHN	28.9%
Hybrid	28.6%

Poster session – S. Deena: Combining Feature and Model-Based Adaptation of RNNLMs for Multi-Genre Broadcast Speech Recognition

Sheffield Transcription

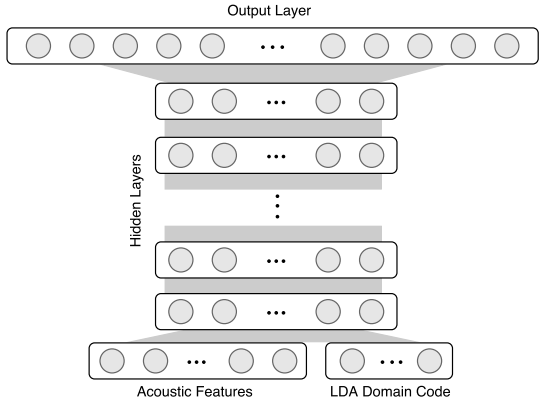
Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Automatic speech segmentation with a DNN–HMM system
- Decoding with a DNN–HMM system with PLP–bMMI features
- Re–segmentation using the output of decoding
- Automatic clustering using BIC and PLP features
- Decoding with 3 separate systems (4-gram LMs)
  - 2 DNN-HMM systems (one with adapted features) and one adapted Tandem GMM-HMM with DNN-derived features
- System combination with ROVER
- Poster session – O.Saz: The 2015 Sheffield System for Transcription of Multi–Genre Broadcast Media

Acoustic Domain Adaptation (Sheffield)

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- DNN acoustic domain adaptation
  - Acoustic domains are inferred from training data using Latent Dirichlet Allocation (LDA)
  - An auxiliary vector with domain weights is used for adaptation for each segment



	WER
Baseline	33.3%
SAT	31.4%
LDaT	30.6%
LDaT+SAT	28.9%

Oral session – M. Doulaty: Latent Dirichlet Allocation Based Organisation of Broadcast Media Archives for Deep Neural Network Adaptation

Cambridge Alignment

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Uses system to generate training data alignments
  - Lightly supervised decoding using hybrid models
  - Modified due to no initial time-stamp information
  - Finally time align words using Tandem model
  - Maximise precision using confidence-based filtering of output + comparing word times between aligned script & lightly supervised outputs (B1 primary, B4 modified version using complete conf-net)

system	F	Precision	Recall	N_{hyp}	N_{match}
B1	0.9120	0.9283	0.8936	141,404	131,260
B4	0.9160	0.9311	0.9013	141,754	131,991

Poster session – The Development of the Cambridge University Alignment Systems for the Multi-Genre Broadcast Challenge

Alignment System (Sheffield)

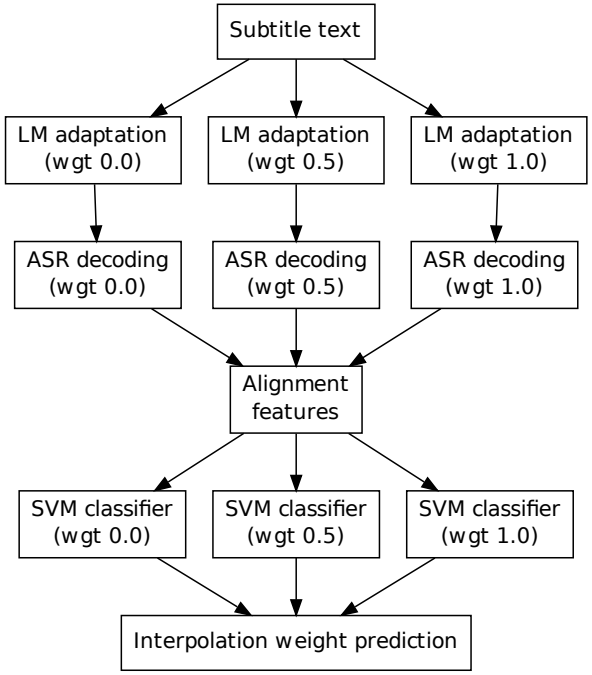
Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Automatic speech segmentation, clustering and decoding with the Task 1 system
  - Interpolating background n–gram with subtitles n–gram for each show
  - DTW alignment of decoding output to show subtitles
  - Word–level time stamps using Viterbi forced alignment

Alignment highlights (Sheffield)

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Selection of interpolation weights in lightly supervised decoding
  - Segments in a show are better covered by subtitles than others
  - SVM is used to estimate best interpolation weight per segment
  - Improves WER, but not F–measure



	WER	F
Baseline	23.9%	0.8593
Segmentwise	22.7%	0.8590
Oracle	21.1%	0.8611

Poster session – B. Khaliq et al.: Segmentwise language model interpolation for lightly supervised alignment of broadcast ubtitles

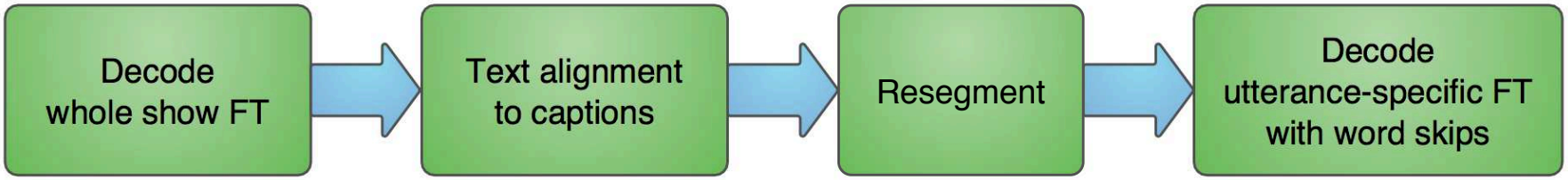
Edinburgh/Quorate Alignment System

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Uses a factor automaton (or transducer) to apply strong text constraints during decoding limiting each utterance to substrings of the reference text (one per show in first pass).
  - Much improved decoding accuracy in difficult acoustic conditions
  - search space is highly constrained → more efficient decoding
  - robust to insertions (words spoken but not in script) but not deletion (words in script but not spoken) – both are common in this data
- Second pass WFSTs are generated dynamically per utterance by selecting surrounding text, and word skips are allowed, giving robustness to deletions

Edinburgh Alignment (ctd)

Natural Speech Technology
Edinburgh – Cambridge – Sheffield



- Results show good performance
- Importance of adding second pass to boost recall
- Further details in poster.

System	Precision	Recall	F-score
Preliminary DNN AMs			
Pass 1 FT	0.8816	0.7629	0.8180
+ force align	0.8290	0.7855	0.8066
Pass 2 FT+del	0.8679	0.8563	0.8620
Final DNN AMs			
Pass 1	0.9009	0.8128	0.8546
Pass 2 FT+del	*0.8856*	*0.9013*	*0.8934*
Pass 2(b) FT	0.8896	0.8946	0.8921
Pass 2(b) FT+del	0.8928	0.9002	**0.8965**

Sheffield Diarization

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Automatic speech segmentation with a DNN–HMM system
- Decoding with a DNN–HMM system with PLP–bMMI features
- Re–segmentation using the output of decoding
- Fine–tuning of the speech segmentation DNN using the show data
- Re–segmentation with the fine–tuned DNN–HMM system
- Automatic clustering using BIC and MFCC features
- Fine–tuning of a speaker separation DNN using the show data
- Re–clustering using DNN–HMM speaker separation system with the fine–tuned DNN
- Speaker linking using BIC and PLP features
- Poster session – R. Milner et al. The 2015 Sheffield system for longitudinal diarisation of broadcast media

Diarization Highlights (Sheffield)

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Three–step speech segmentation leads to low segmentation error
  - 2–output DNN trained to separate speech from non–speech using 700 hours of acoustic training dataInitial
  - Decoding output is used to filter out areas of non–speech
  - The DNN is fine–tuned on the test data and speech is re–segmented

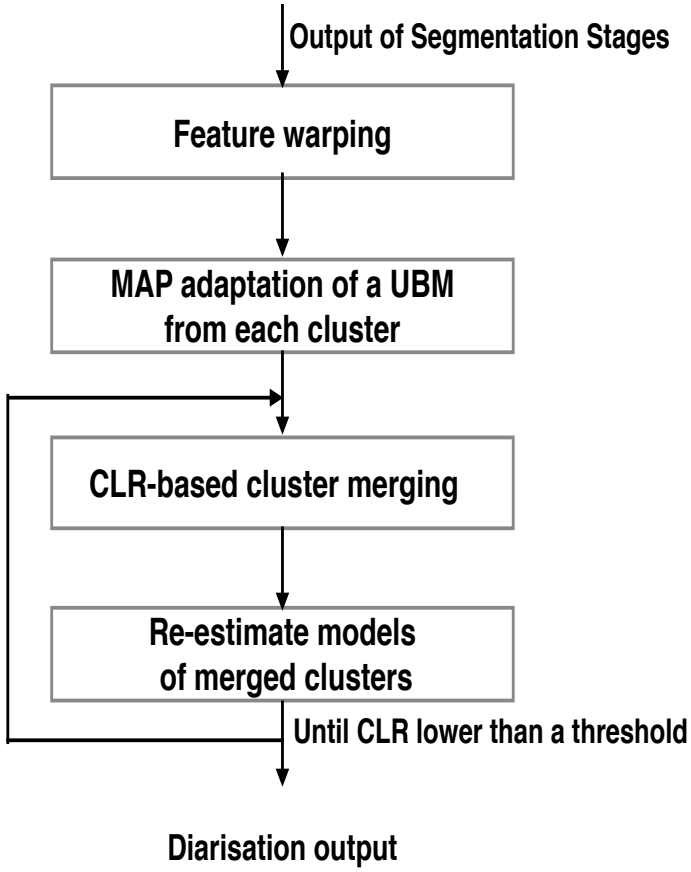
	Miss	False	SER
DNN–HMM	4.1%	8.5%	12.6%
+Decoding	6.7%	2.7%	9.4%
+DNN fine–tuning	4.4%	3.8%	8.2%

Poster session – R. Milner et al.: The 2015 Sheffield system for longitudinal diarisation of broadcast media

Cambridge Diarisation Approach

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- Use segmenter developed for transcription
- Uses classic UBM-based representation of each cluster based on warped features
- Cross-likelihood ratio between clusters used as a distance measure for clustering
- Cross-episode linking applied after basic diarisation, and uses a complete linkage clustering between clusters



Poster session – Speaker diarisation and longitudinal linking in multi-genre broadcast data

Conclusions

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

- Many systems developed for MGB as part of NST across range of NST tasks
- Improved segmentation algorithms are important
- Use of subtitle data and refined alignment for training
- Various types of acoustic model, adaptation and combination
- RNNLM applied to a large scale task with adaptation
- Different approaches to the alignment task
- Diarisation and linking used different techniques between sites
- **Many posters** that give more details!