EPSRC
Engineering and Physical Sciences
Research Council

MGB
CHALLENGE

Natural
Speech
Technology
Edinburgh – Cambridge – Sheffield

The MGB Challenge

Evaluating Multi-Genre Broadcast Media Recognition

Peter Bell, Jonathan Kilgour, Steve Renals, Mirjam Wester *University of Edinburgh*
Mark Gales, Pierre Lanchantin, Xunying Liu, **Phil Woodland** *University of Cambridge*
Thomas Hain, Oscar Saz *University of Sheffield*
Andrew McParland *BBC R&D*

mgb-challenge.org

Overview

Establish an open challenge in core ASR research with common data and evaluation benchmarks on broadcast data

Controlled evaluation of speech recognition, speaker diarization, and alignment

Used a broad, multi-genre dataset of BBC TV output

Challenge Task at ASRU 2015

Subtitles & light supervision

- Training data transcribed by subtitles (closed captions) – can differ from verbatim transcripts
  - edits to enhance clarity
  - paraphrasing
  - deletions where the speech is too fast
- There may be
  - words in the subtitles that were not spoken
  - words missing in the subtitles that were spoken
- Additional metadata includes speaker change information, timestamps, genre tags, …

MGB Resources

Fixed acoustic and language model training data

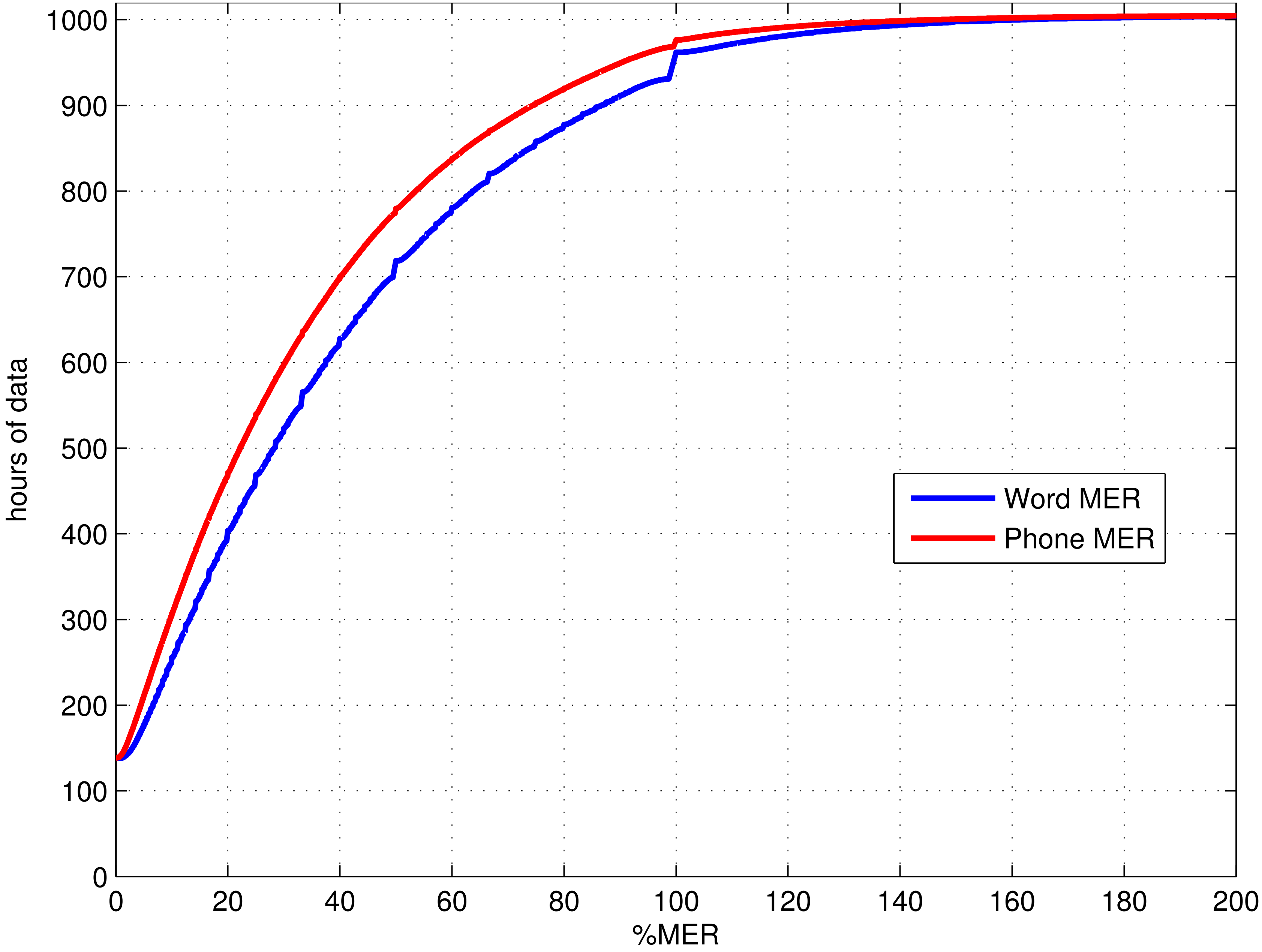
- precise comparison of models and algorithms
- data made available by BBC R&D Labs

- **Acoustic model training**
  1600h broadcast audio across 4 BBC channels (1 April – 20 May 2008), with as-broadcast subtitles – ~33% WER (26% deletions)
- **Language model training**
  640 million words BBC subtitles (1979–2013)
- **Lexicon**
  ASR version of Combilex

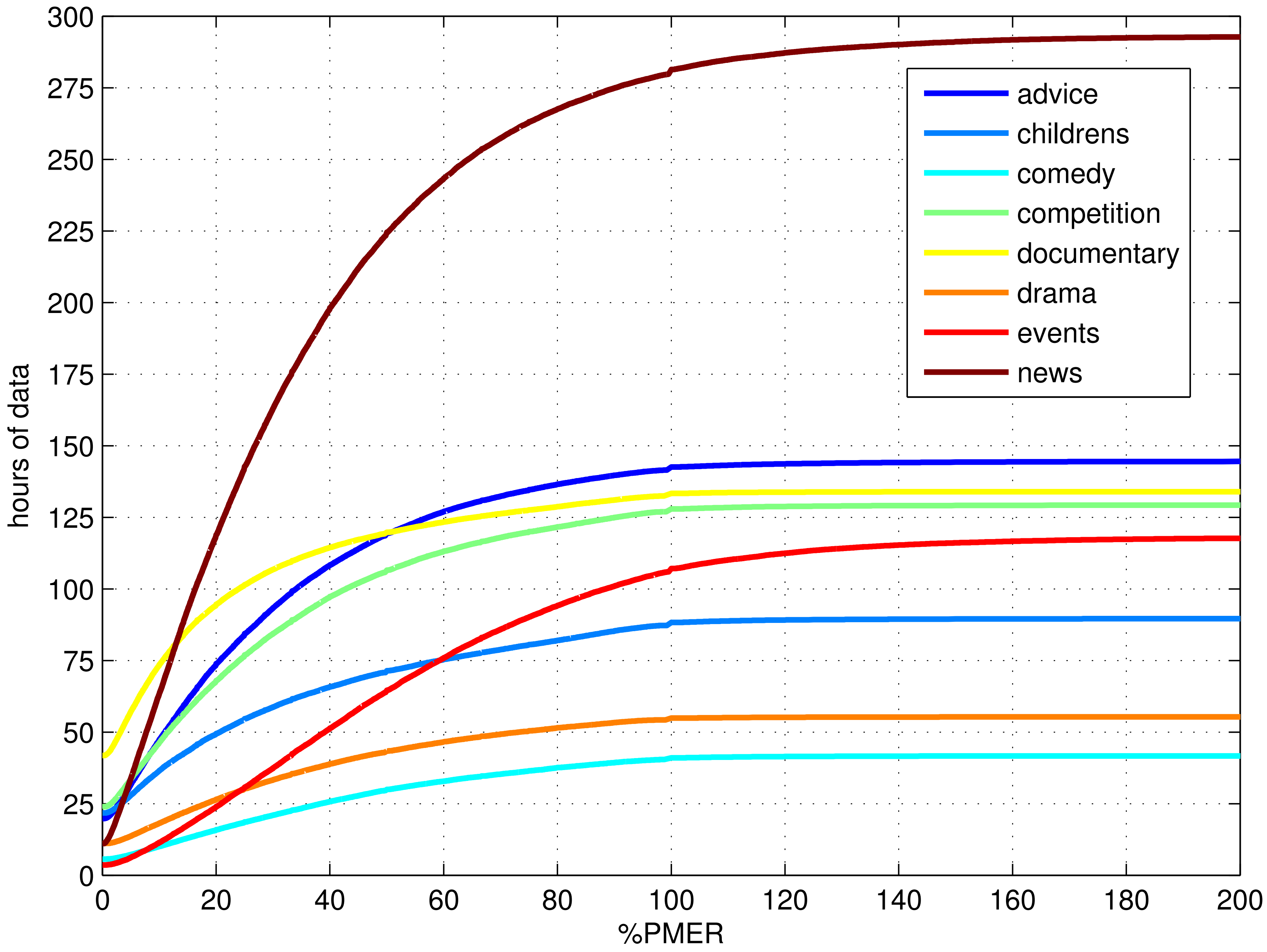
Pre-processing & data selection

- **Pre-processing**
  - transcript normalisation
  - acoustic segmentation
  - subtitle alignment
  - confusion scores computed for aligned segments using confusion networks and biased LM
- **Data Selection**
  - Average word duration – reject non-speech
  - Phone/word matched error rate (PMER/WMER) – decoding scored against aligned subtitles

Training data selection



Training data by genre

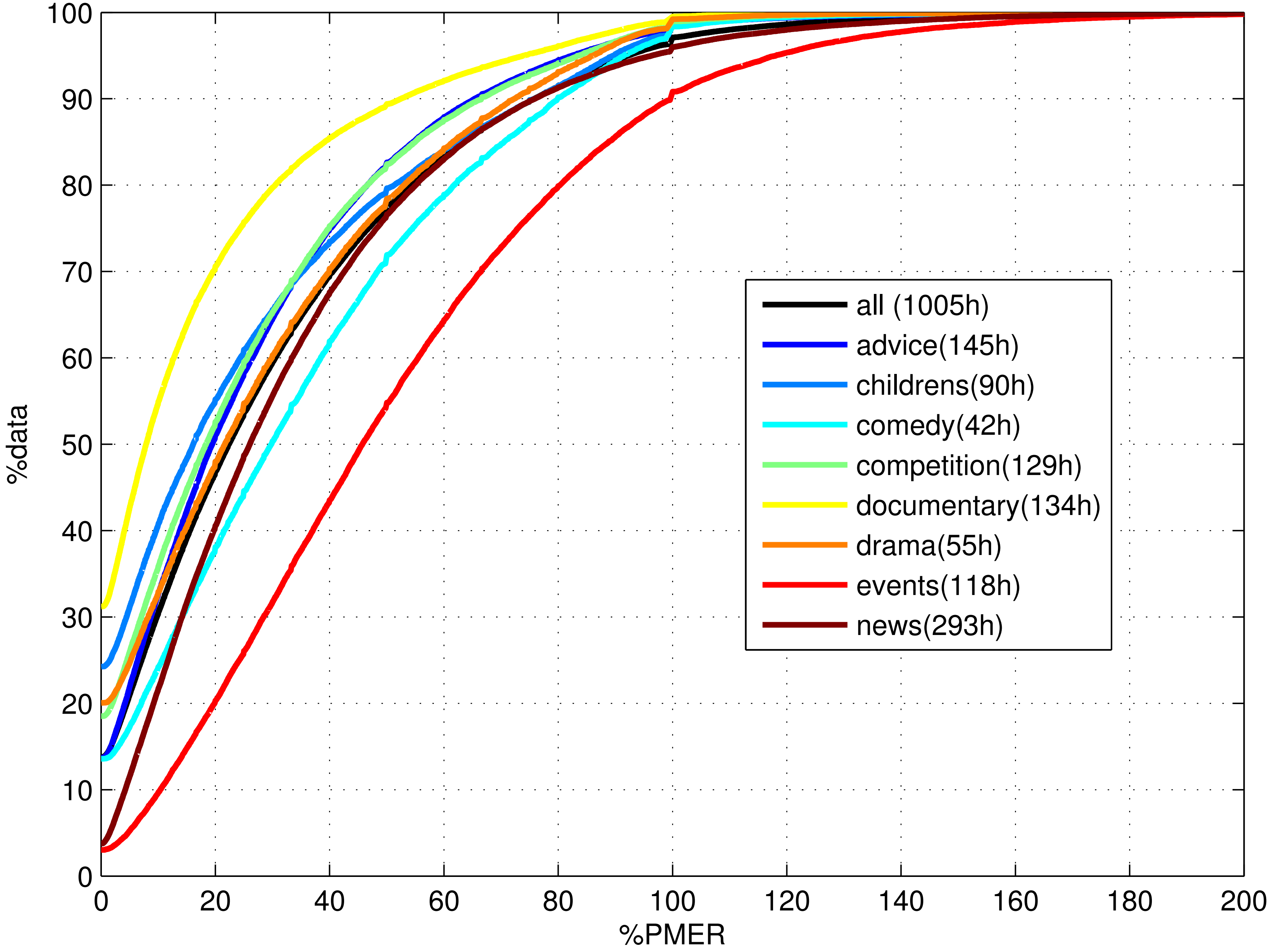


MGB CHALLENGE

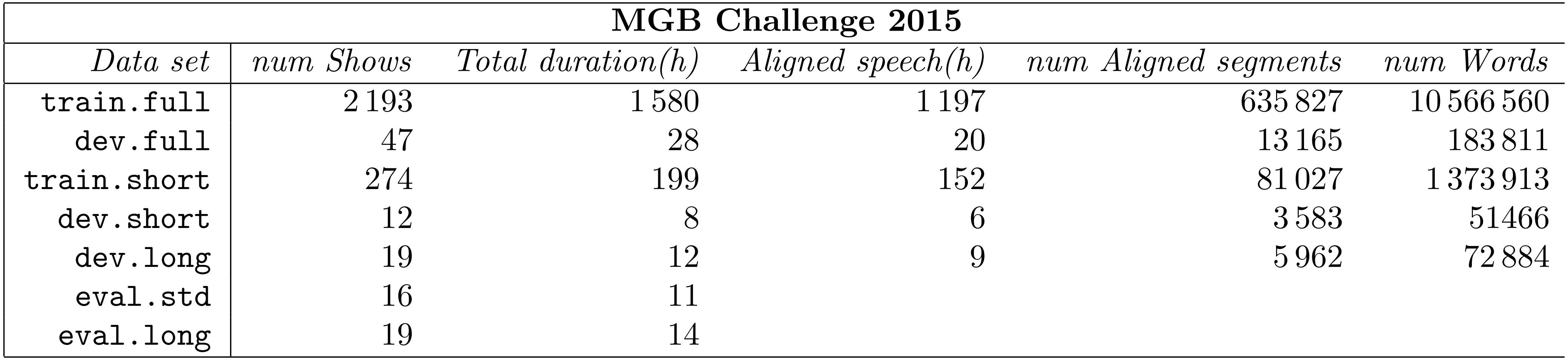
Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Training data by genre



MGB Data



MGB Challenge 2015					
Data set	*num Shows*	*Total duration(h)*	*Aligned speech(h)*	*num Aligned segments*	*num Words*
train.full	2 193	1 580	1 197	635 827	10 566 560
dev.full	47	28	20	13 165	183 811
train.short	274	199	152	81 027	1 373 913
dev.short	12	8	6	3 583	51466
dev.long	19	12	9	5 962	72 884
eval.std	16	11			
eval.long	19	14			

- Dev and eval data manually transcribed (by correcting subtitles)
    - 2 transcribers
    - 8x broadcast time
    - 96% agreement

Baseline Systems

- Use of Kaldi, XMLStarlet, SRILM, IRSTLM
- ASR – Speaker-adaptive GMM, DNN acoustic models
  - 11,500 tied triphone states
  - ML training using PLP, +LDA +MLLT +fMLLR
  - 3/4-gram LMs
  - 150k word lexicon (Combilex + g2p)
  - Training data selection based on WMER
  - DNN – 2 iters of CE training followed by sMBR sequence training (released post-evaluation)
- Segmenter
  - speech/non-speech DNN classifier (smoothed using HMM)
  - BIC-based speaker clustering
  - ~5% higher WER compared with gold-standard segmentation

MGB Tasks

1. *Speech-to-text transcription*
2. *Alignment*
3. *Longitudinal speech-to-text transcription*
4. *Longitudinal speaker diarization and linking*

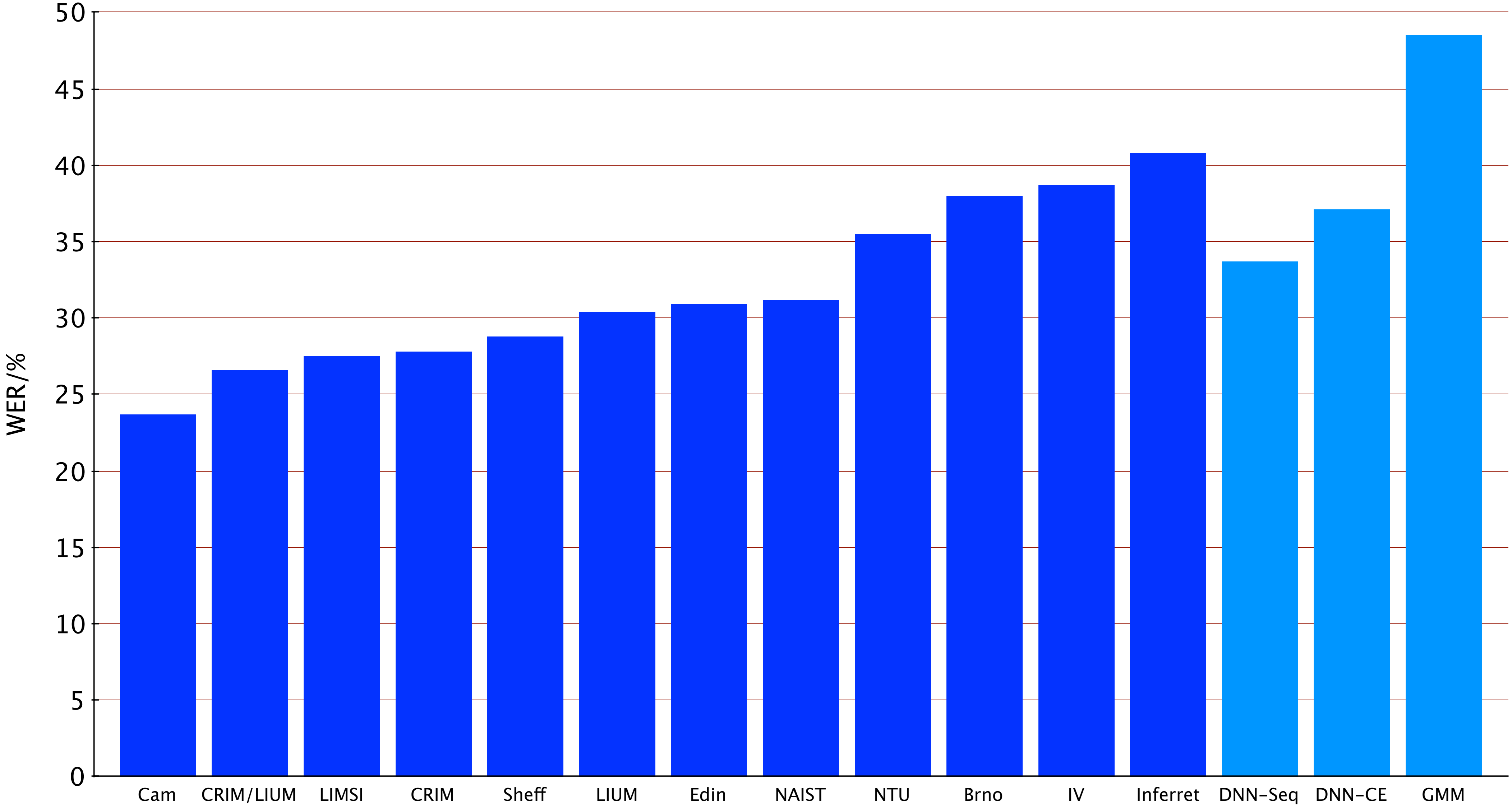
MGB participants

- Task 1 – transcription
  - BUT, Brno
  - CRIM
  - Inferret
  - Intelligent Voice
  - LIMSI
  - LIUM
  - NAIST
  - NTU, Singapore
  - Univ Cambridge
  - Univ Edinburgh
  - Univ Sheffield
- Task 3 – longitudinal trans.
  - Cambridge, Edinburgh, Sheffield
- Task 2 – alignment
  - CRIM
  - NHK
  - Quorate / Edinburgh
  - Cambridge
  - Sheffield
  - Vocapia / LIMSI
- Task 4 – diarization
  - IDIAP
  - Orange / LIUM
  - Cambridge
  - Edinburgh
  - Sheffield
  - Univ Zaragoza

MGB CHALLENGE

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

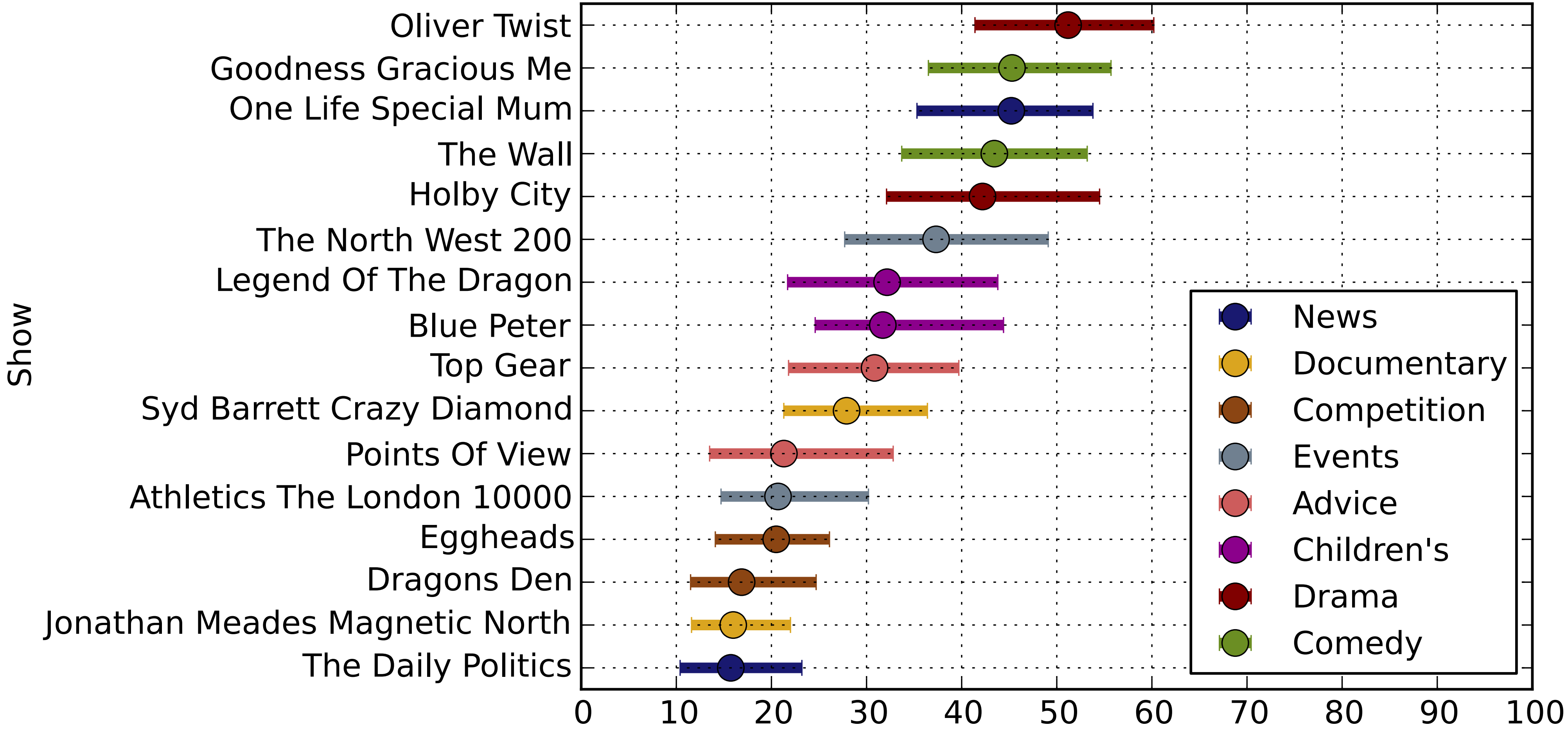
Results – Transcription



MGB CHALLENGE

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Results by Show - Transcription



Results by show – Transcription

Results by show – Transcription

Longitudinal Transcription

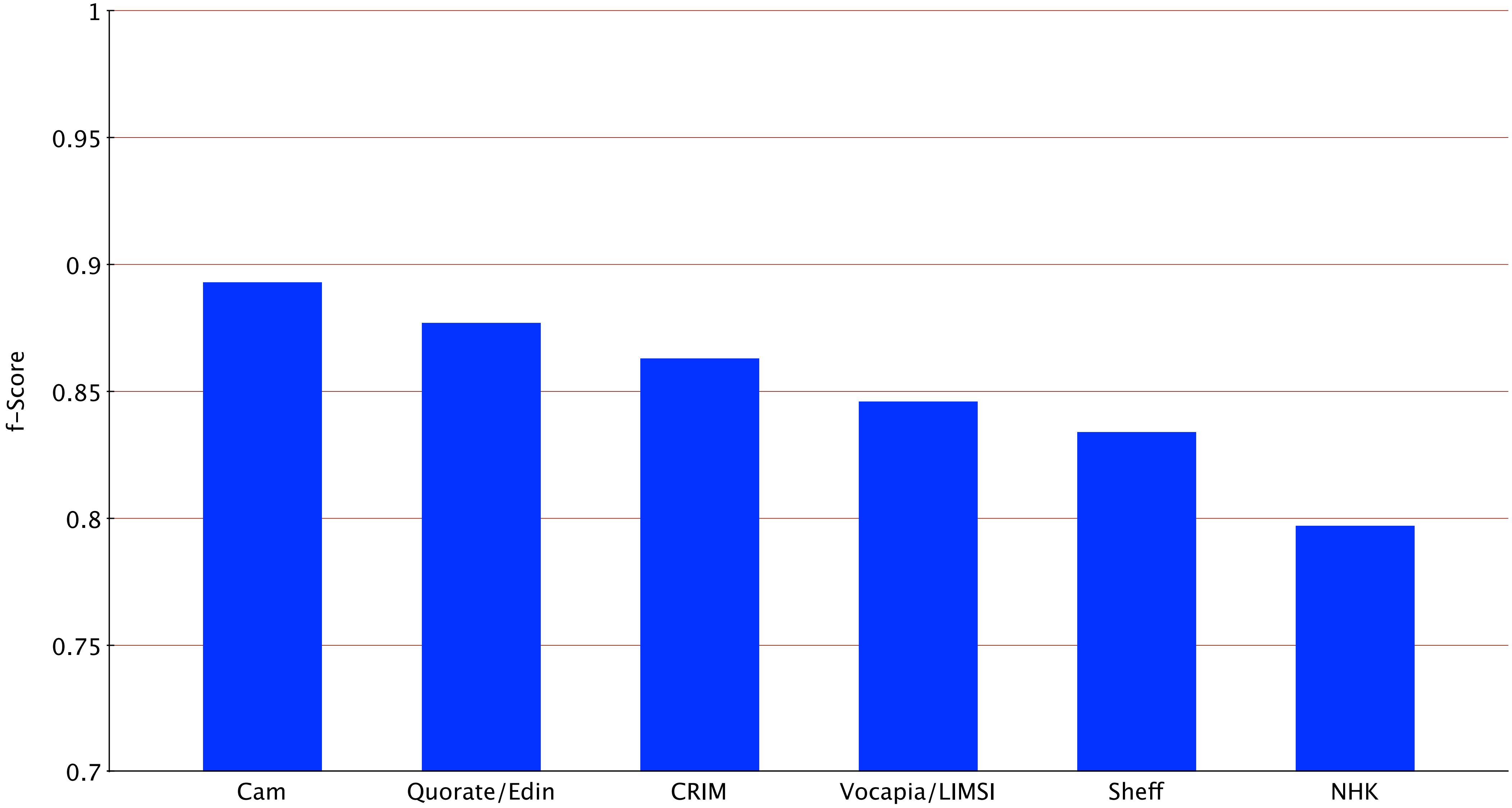
- Aimed at causal adaptation across episodes of same series (different test data to task 1).
  - No site did series based adaptation
  - Deadline one week later: NST sites updated systems! (perhaps 1.5-2% abs lower WER same data).

Participant	Substitutions	Deletions	Insertions	Word Error Rate
CU	8.6%	7.9%	2.8%	19.3%
SU	11.7%	9.8%	3.2%	24.8%
UE	10.9%	12.6%	2.8%	26.3%

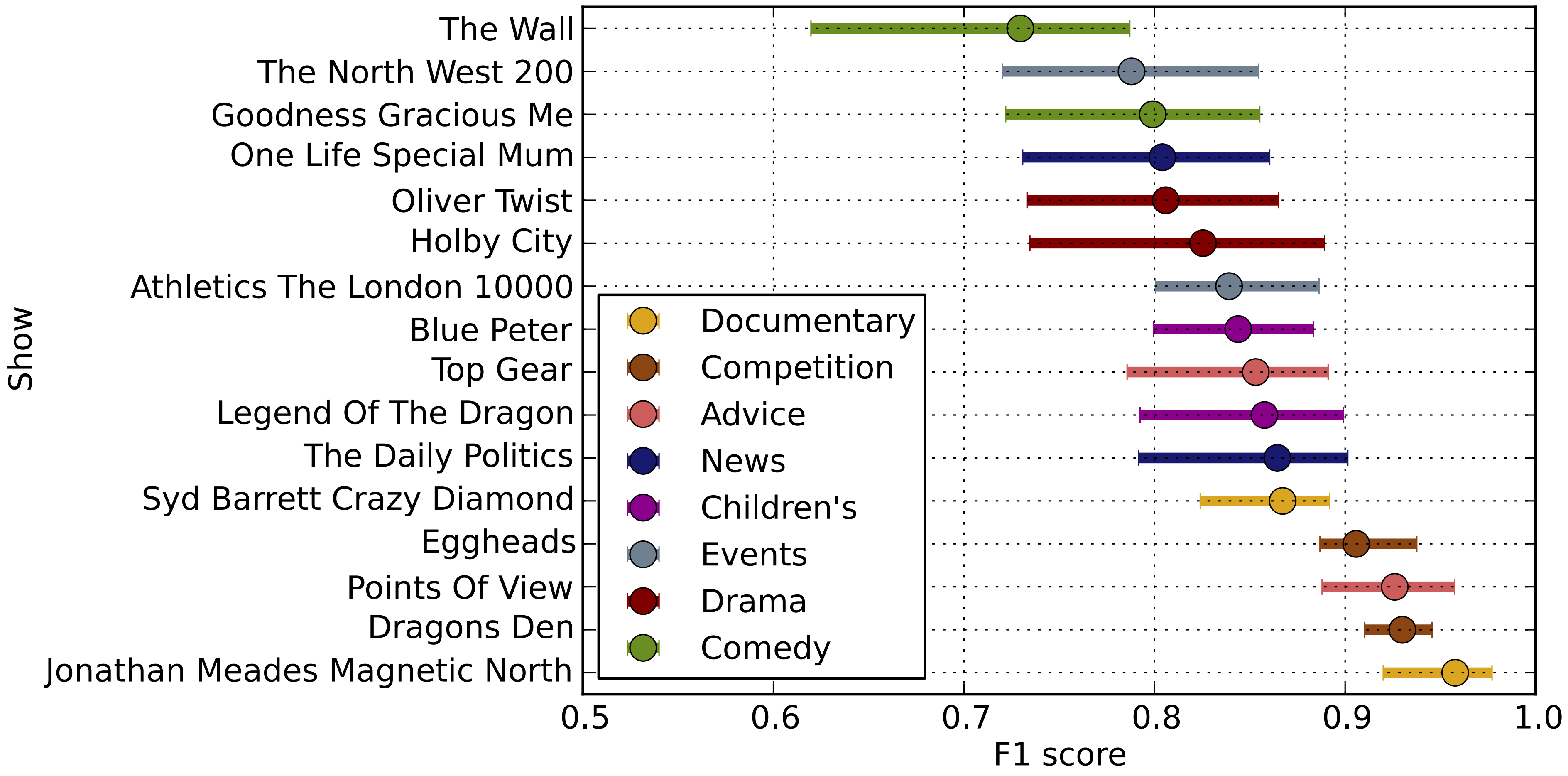
Alignment

- Task: align tokenised subtitles to spoken audio at word level (where possible)
- Scoring performed by calculating precision & recall (summarised as f-score), derived from automatic alignment of a careful manual transcription.
- A word matches if both start and end times fall within a 100ms window of the associated reference word.
- Only words from the script to be aligned
- Regions of overlapped speech not evaluated

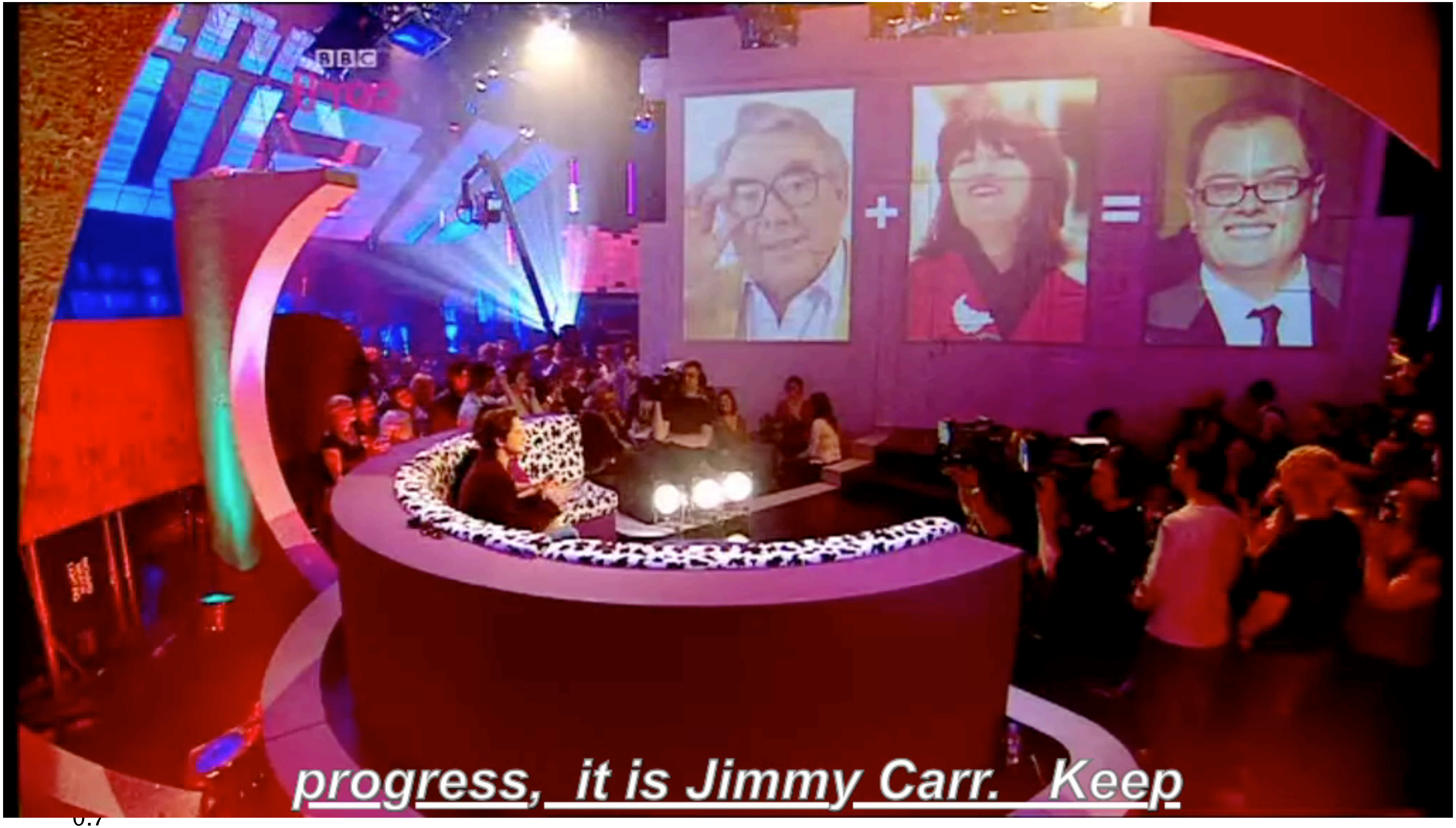
Results – Alignment



Results by show – Alignment



Results by show – Alignment



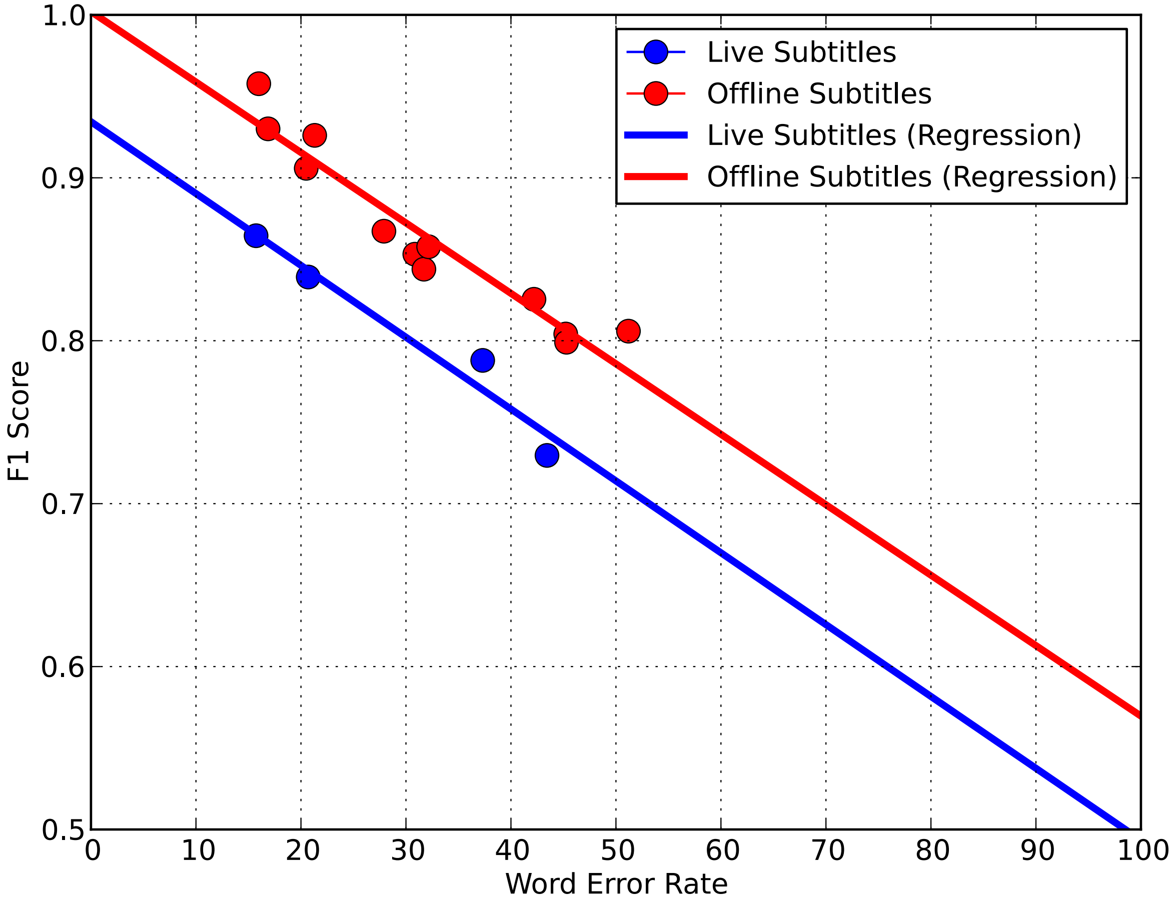
progress, it is Jimmy Carr. Keep

0.7

Transcription-Alignment Correlation

- Plot the correlation between WER and alignment f-score measure across shows

- Separate live subtitles and off-line
- Increase WER by 1% gives 0.004 worse f-score



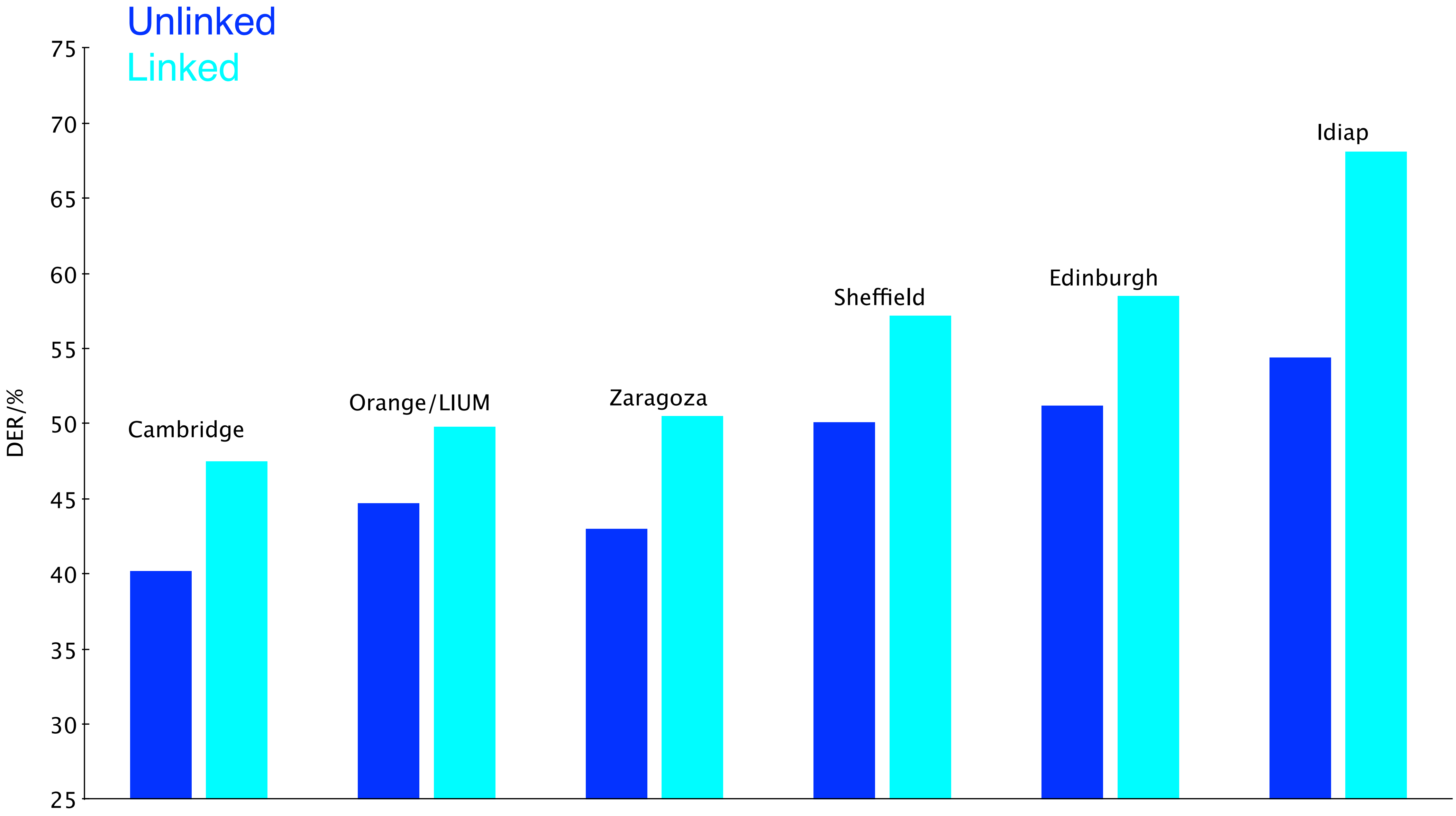
Diarization

- Evaluation of speaker diarization in a *longitudinal* setting
- Systems aimed to label speakers uniquely across a whole series (linked diarization)
- Speaker labels for each show were obtained using only material from the show in question, and those broadcast earlier in time
- No external sources of training data permitted (e.g. for building i-vector extractors)
- As a contrast also evaluated single-show unlinked diarisation

MGB CHALLENGE

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Results – Diarization



MGB CHALLENGE

MGB–2 (& beyond?)

Natural Speech Technology
Edinburgh – Cambridge – Sheffield

- BBC based challenge data not possible to use in 2016
  - problem due to resolving permissions issues in time: hope to use this data again in future
- New Arabic task arranged for 2016 (QCRI / Edinburgh)
  - Evaluated ASR on multi-genre TV data from Aljazeera
  - 1,200h of TV programmes released as training data, along with lightly-supervised alignment of captions from QCRI system.
  - 110M words from Aljazeera.com website (2004-2011) for LMs
  - Verbatim transcripts of 20 hours of programmes from 2015 manually created for use as development and evaluation data
  - 10 (non NST) labs submitted systems. Entries from the US, Japan, China, Europe and several from Arabic-speaking world

mgb-challenge.org

Conclusions

- **MGB was a real challenge!**
- Multi-genre broadcast speech presents a substantial challenge – highly variable across shows
- All tasks tackled showed interesting range of performance (across systems and shows)
- Speaker diarization of this data, in particular, is highly challenging

Supported by **EPSRC** and **NVIDIA**