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

MGB CHALLENGE

Overview

MGB CHALLENGE

Overview

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

MGB CHALLENGE

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

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

MGB CHALLENGE

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 CHALLENGE

MGB Resources

Fixed acoustic and language model training data

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

MGB CHALLENGE

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

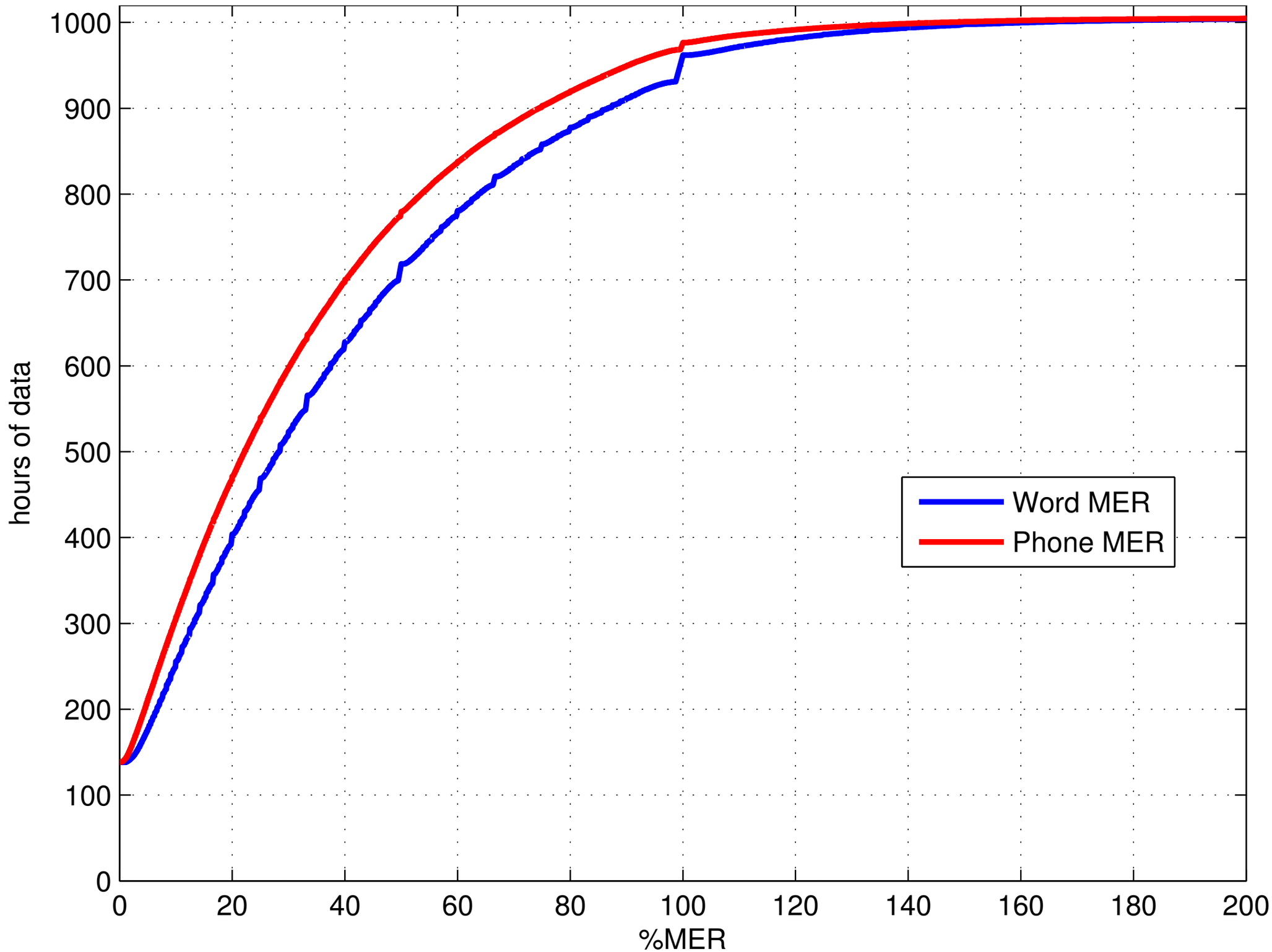
MGB CHALLENGE

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

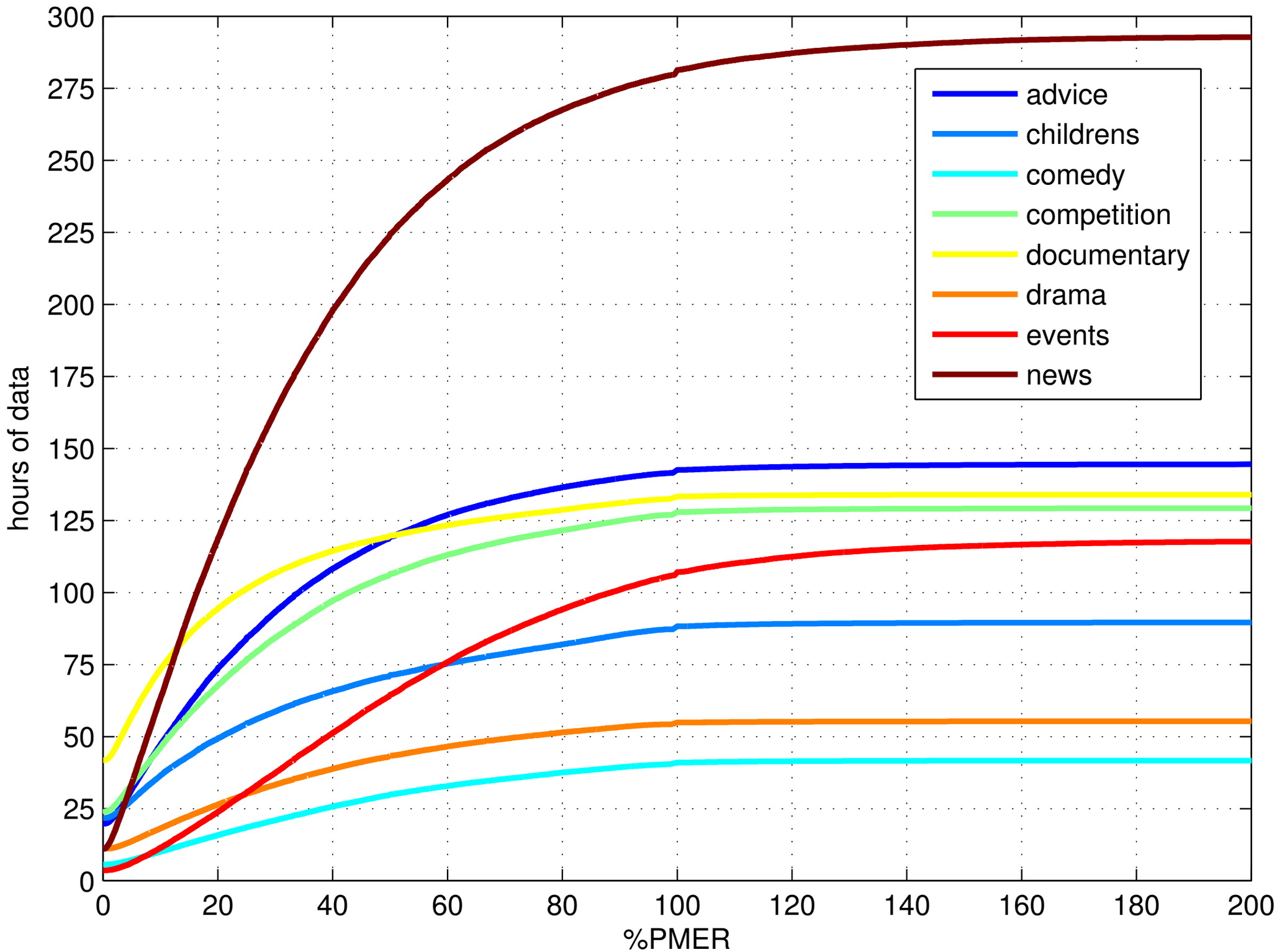
MGB CHALLENGE

Training data selection



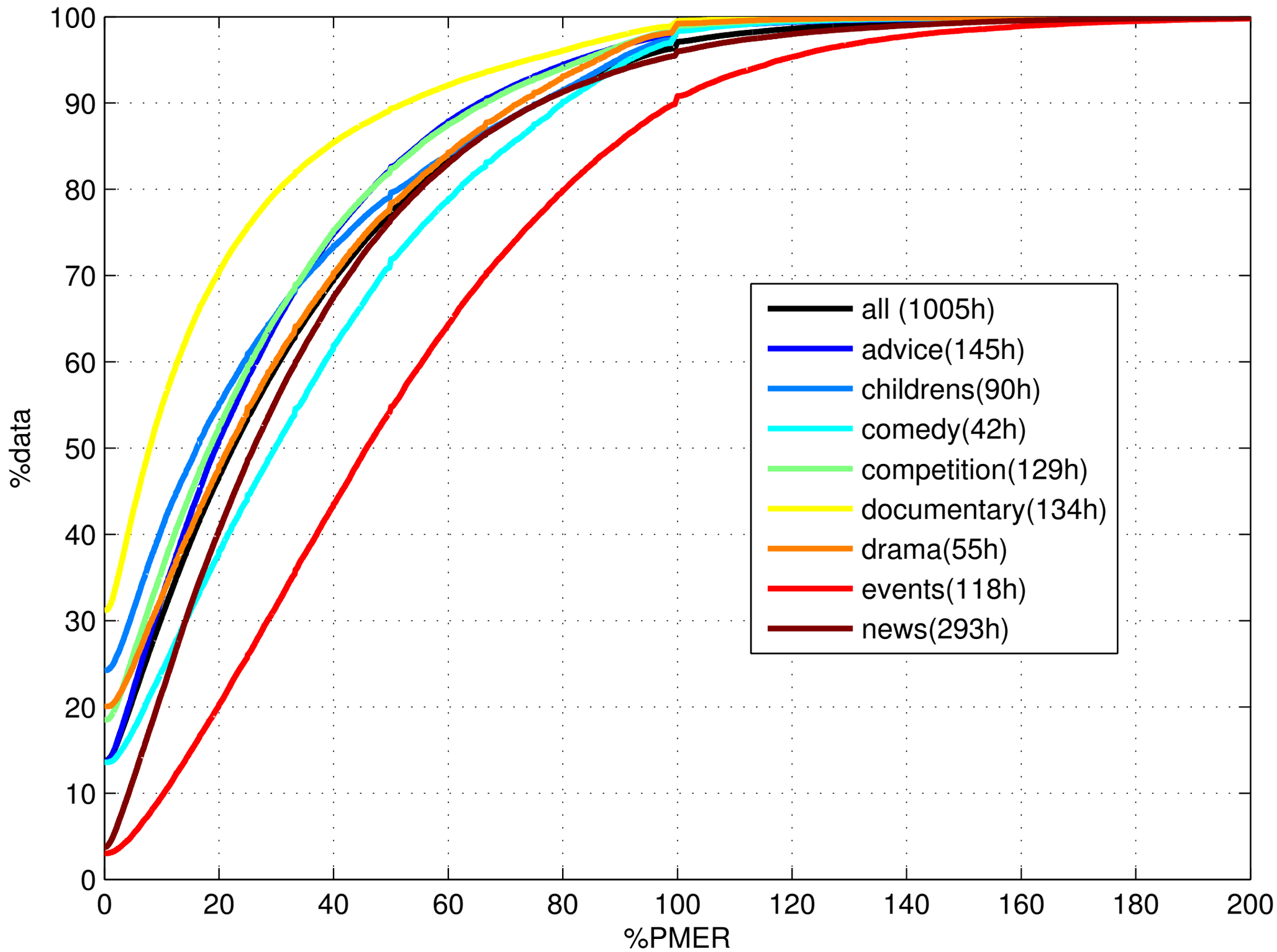
MGB CHALLENGE

Training data by genre



MGB CHALLENGE

Training data by genre



MGB CHALLENGE

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 CHALLENGE

MGB Tasks

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

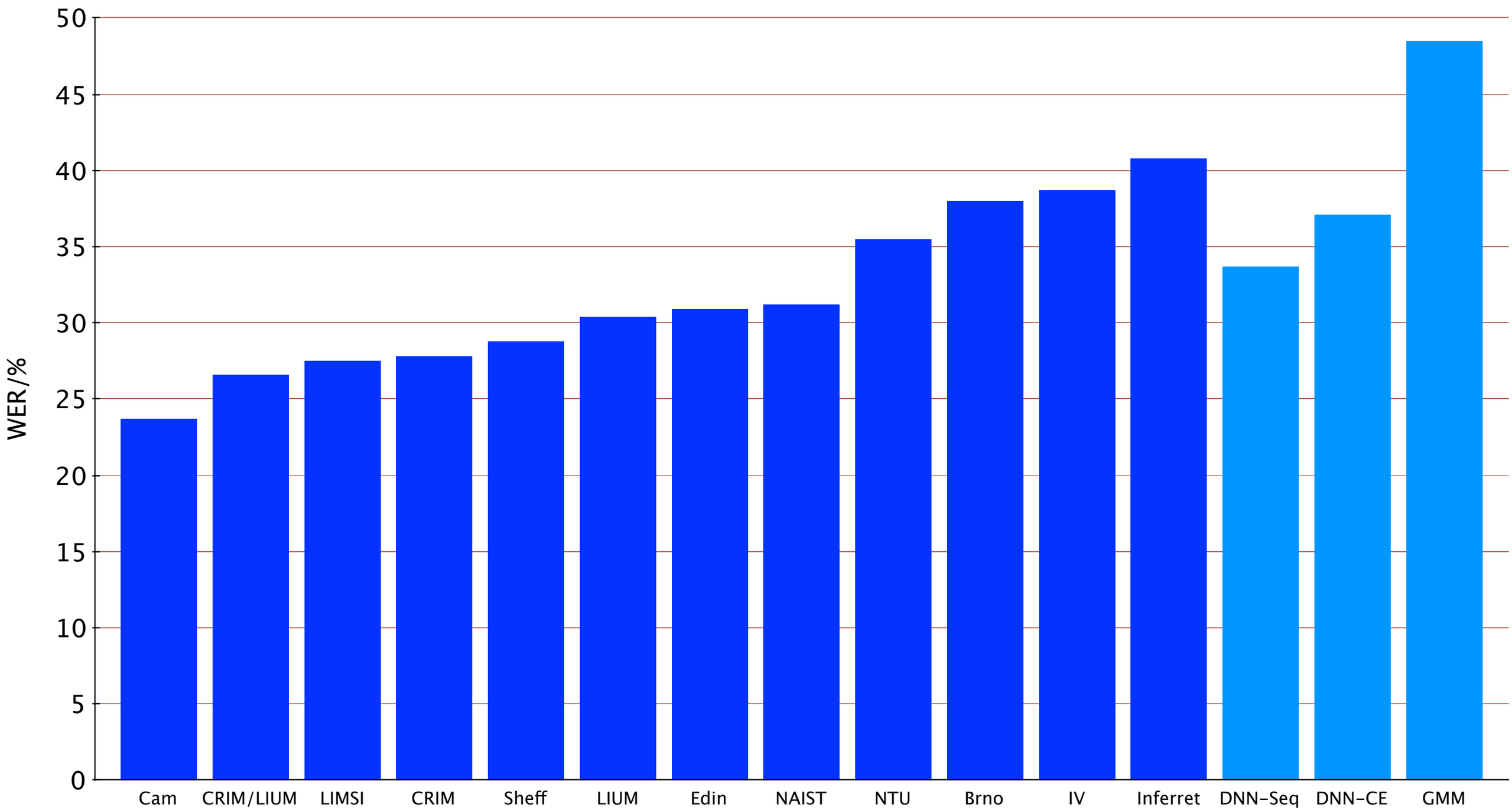
MGB CHALLENGE

MGB participants

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

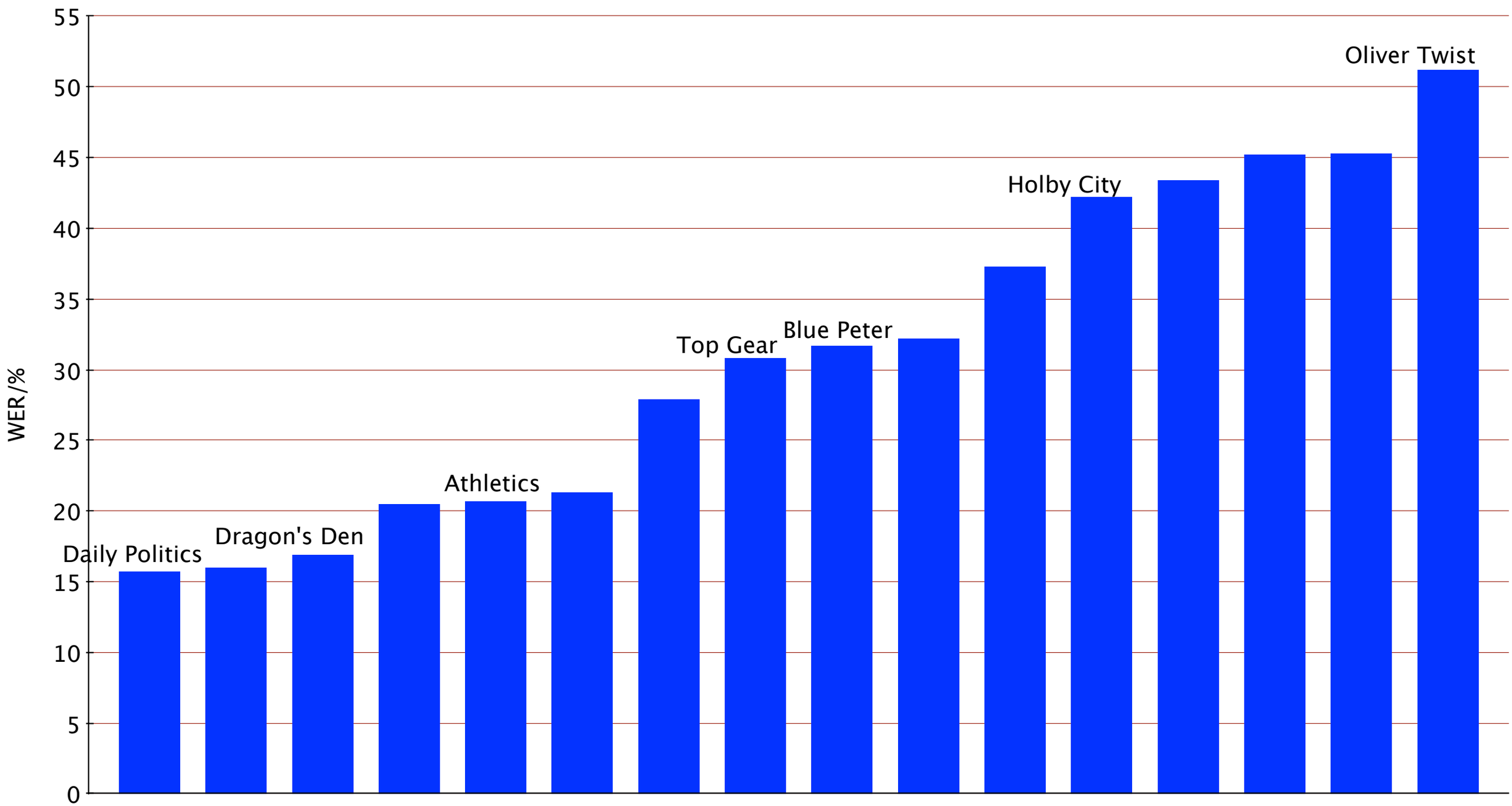
MGB CHALLENGE

Results – Transcription



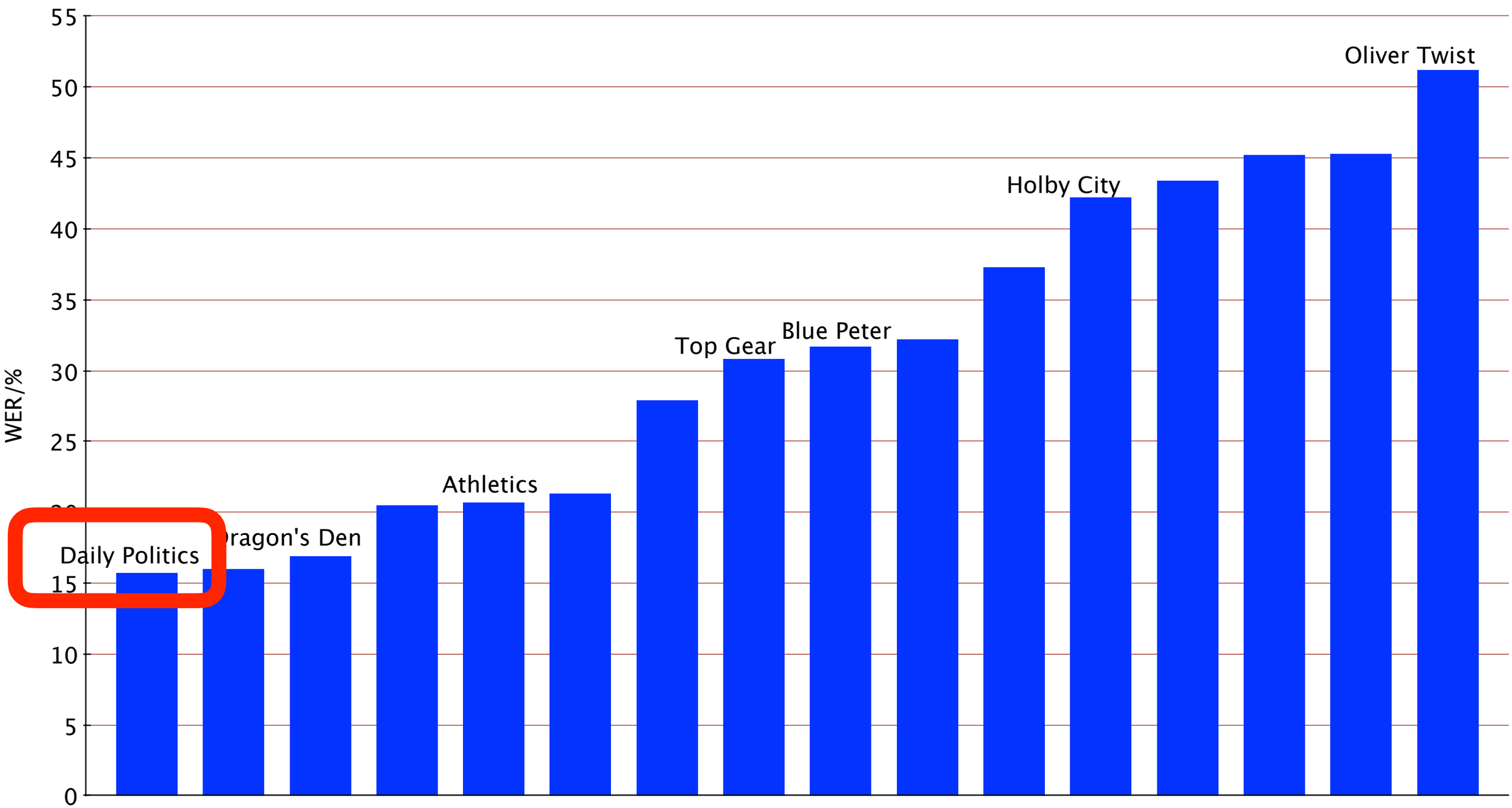
MGB CHALLENGE

Results by show – Transcription



MGB CHALLENGE

Results by show – Transcription



MGB CHALLENGE

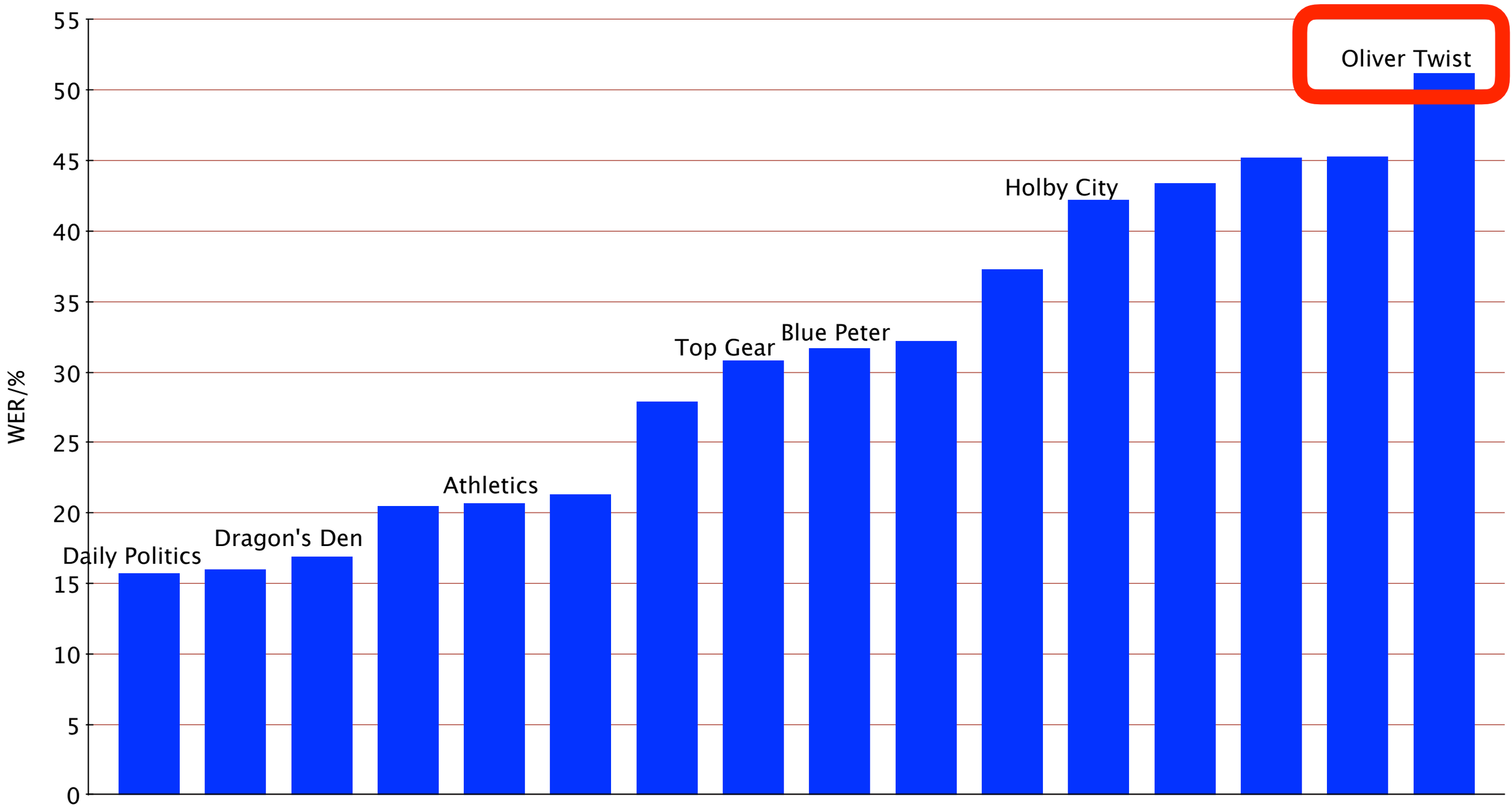
Results by show – Transcription

MGB CHALLENGE

Results by show – Transcription

MGB CHALLENGE

Results by show – Transcription



MGB CHALLENGE

Results by show – Transcription

MGB CHALLENGE

Results by show – Transcription

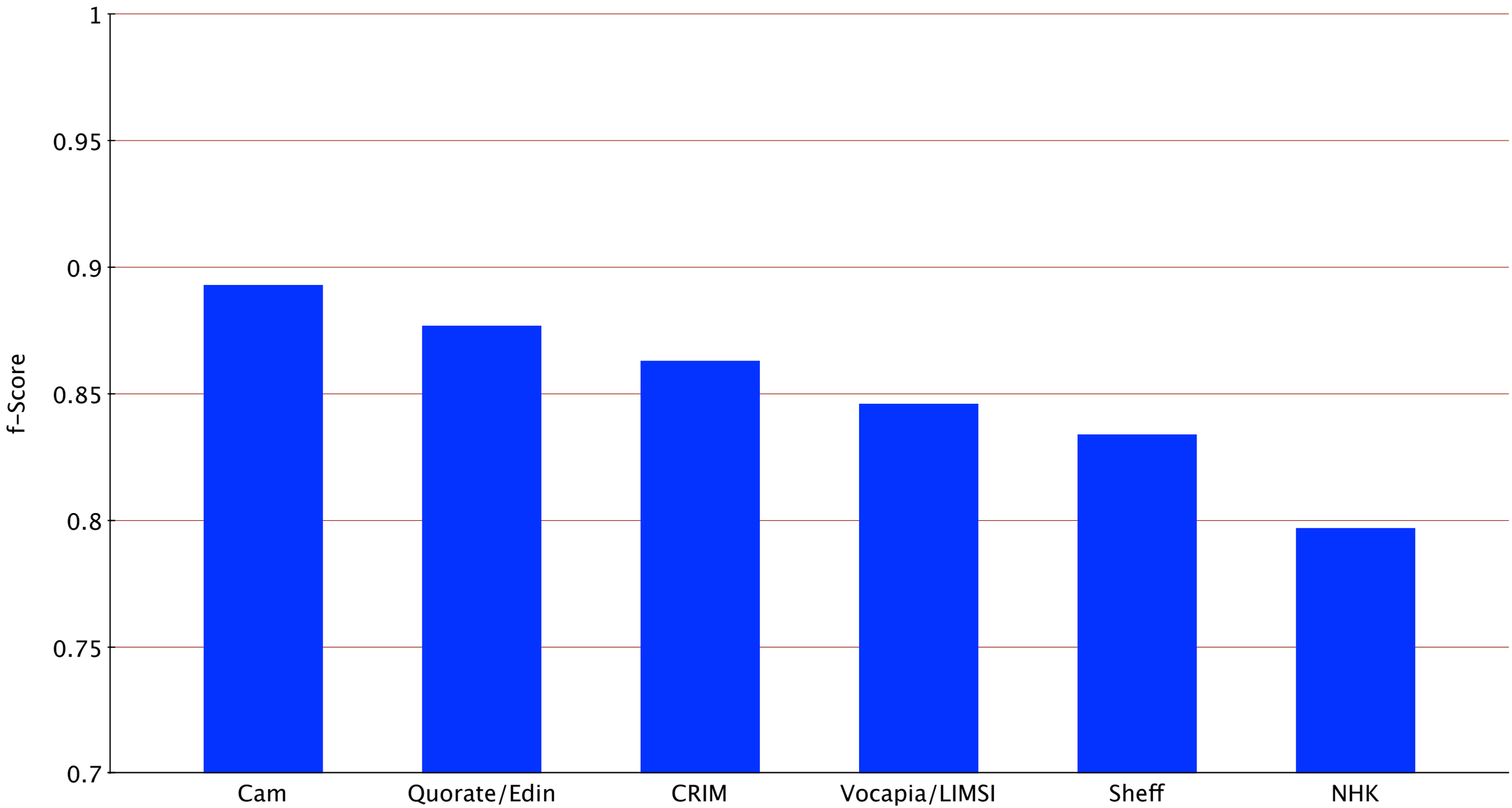
MGB CHALLENGE

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

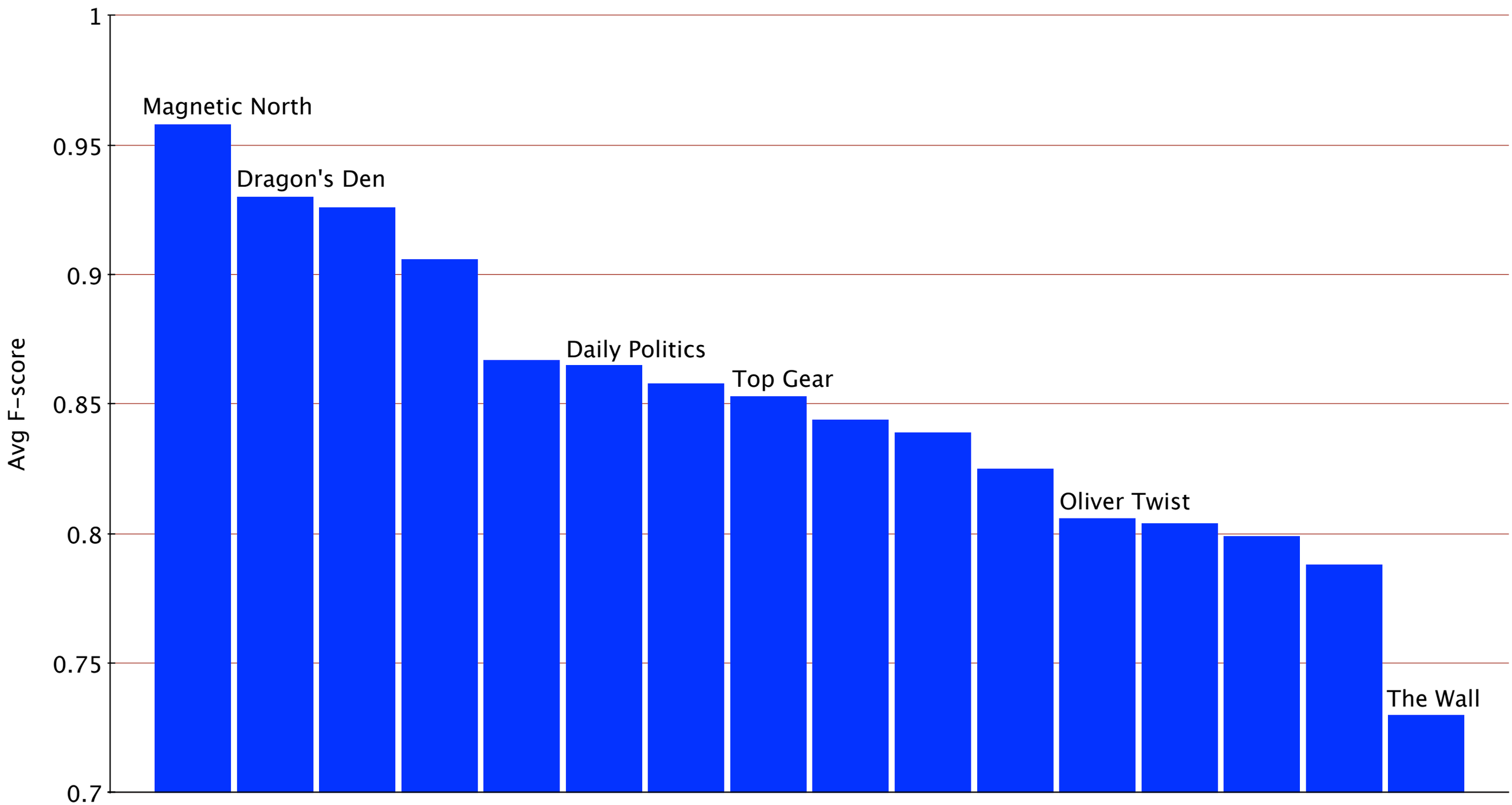
MGB CHALLENGE

Results – Alignment



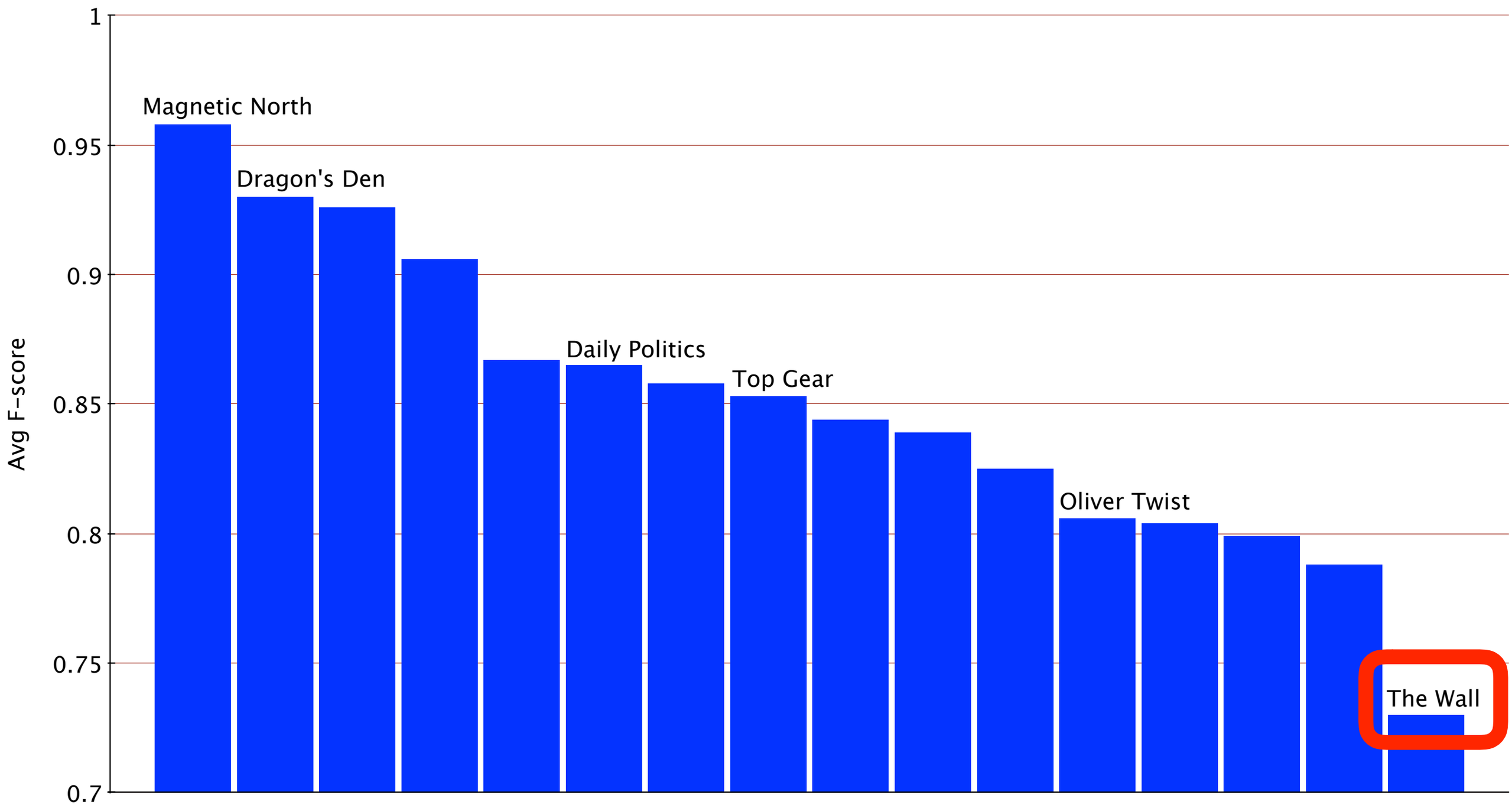
MGB CHALLENGE

Results by show – Alignment



MGB CHALLENGE

Results by show – Alignment



MGB CHALLENGE

Results by show – Alignment

MGB CHALLENGE

Results by show – Alignment

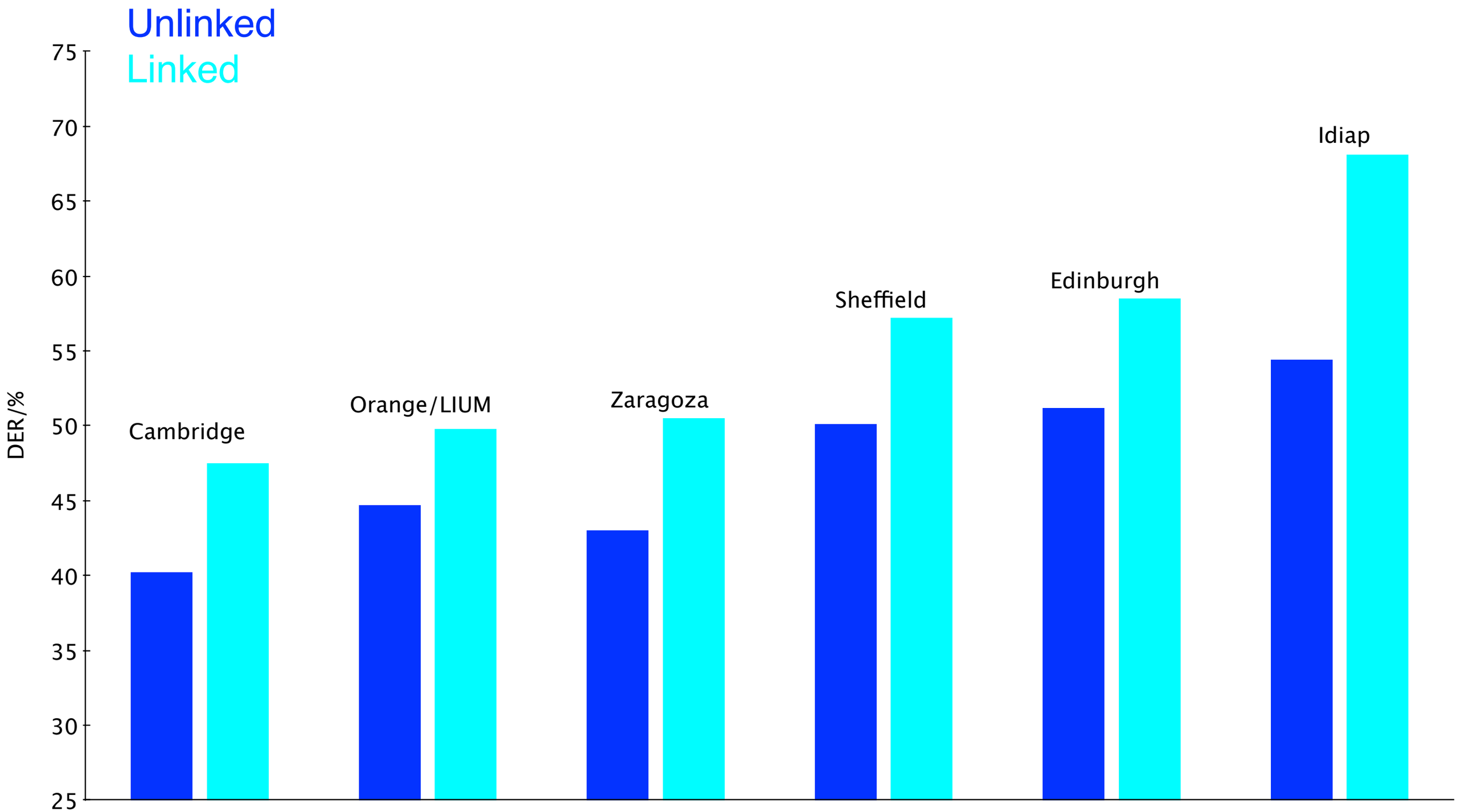
MGB CHALLENGE

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

Results – Diarization



MGB CHALLENGE

Conclusions

- **MGB was a real challenge!**
- Multi-genre broadcast speech presents a substantial challenge – highly variable across shows
- Speaker diarization of this data, in particular, is highly challenging

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

MGB CHALLENGE

MGB–2

- The MGB challenge will run again in 2016
- Same four tasks, same training data, new dev + eval, also some speaker annotations
- New tasks
  - Robust transcription (using BBC MGB data) – not all genres represented in the training data
  - Arabic transcription and dialect identification (using data from Al-Jazeera)
- **Interested in MGB–2?**
  meet at 13:00, Thursday – Fuego Room

mgb-challenge.org

MGB CHALLENGE

MGB Posters (15:30–17:00)

- *The 2015 Sheffield System for Transcription of Multi-Genre Broadcast Media*, Saz et al
- *The 2015 Sheffield System for Longitudinal Diarisation of Broadcast Media*, Milner et al
- *Cambridge University Transcription Systems for the Multi-Genre Broadcast Challenge*, Woodland et al
- *The Development of the Cambridge University Alignment Systems for the Multi-Genre Broadcast Challenge*, Lanchantin et al
- *The NAIST ASR System for the 2015 Multi-Genre Broadcast Challenge: On Combination of Deep Learning Systems Using a Rank-Score Function*, Do et al
- *Speaker Diarisation and Longitudinal Linking in Multi-Genre Broadcast Data*, Karanasou et al
- *Variational Bayesian PLDA for Speaker Diarization in the MGB Challenge*, Villalba et al
- *A System for Automatic Alignment of Broadcast Media Captions using Weighted Finite-State Transducers*, Bell et al
- *CRIM and LIUM Approaches for Multi-Genre Broadcast Media Transcription*, Gupta et al
- *Lightly Supervised Training based on Genre Dependent Caption Accuracy*, Oku et al
- *Speaker Diarization in an Unsupervised i-Vector/PLDA Training Framework*, Le Lan et al

MGB CHALLENGE

MGB Posters (15:30–17:00)

- *The 2015 Sheffield System for Transcription of Multi-Genre Broadcast Media*, Saz et al
- *The 2015 Sheffield System for Longitudinal Diarisation of Broadcast Media*, Milner et al
- *Cambridge University Transcription Systems for the Multi-Genre Broadcast Challenge*, Woodland et al
- *The Development of the Cambridge University Alignment Systems for the Multi-Genre Broadcast Challenge*, Lanchantin et al
- *The NAIST ASR System for the 2015 Multi-Genre Broadcast Challenge: On Combination of Deep Learning Systems Using a Rank-Score Function*, Do et al
- *Speaker Diarisation and Longitudinal Linking in Multi-Genre Broadcast Data*, Karanasou et al
- *Variational Bayesian PLDA for Speaker Diarization in the MGB Challenge*, Villalba et al
- *A System for Automatic Alignment of Broadcast Media Captions using Weighted Finite-State Transducers*, Bell et al
- *CRIM and LIUM Approaches for Multi-Genre Broadcast Media Transcription*, Gupta et al
- *Lightly Supervised Training based on Genre Dependent Caption Accuracy*, Oku et al
- *Speaker Diarization in an Unsupervised i-Vector/PLDA Training Framework*, Le Lan et al