# THE MGB CHALLENGE: EVALUATING MULTI-GENRE BROADCAST MEDIA RECOGNITION

*P Bell* [1], *MJF Gales* [2], *T Hain* [3], *J Kilgour* [1], *P Lanchantin* [2], *X Liu* [2],
*A McParland* [4], *S Renals* [1], *O Saz* [3], *M Wester* [1], *PC Woodland* [2]

(1) Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK
(2) Cambridge University Engineering Department, Cambridge CB2 1PZ, UK
(3) Speech and Hearing Group, Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
(4) BBC R&D, South Lab, London, W12 7SB, UK
mgb-admin@inf.ed.ac.uk www.mgb-challenge.org

## ABSTRACT

This paper describes the Multi-Genre Broadcast (MGB) Challenge at ASRU–2015, an evaluation focused on speech recognition, speaker diarization, and "lightly supervised" alignment of BBC TV recordings. The challenge training data covered the whole range of seven weeks BBC TV output across four channels, resulting in about 1,600 hours of broadcast audio. In addition several hundred million words of BBC subtitle text was provided for language modelling. A novel aspect of the evaluation was the exploration of speech recognition and speaker diarization in a longitudinal setting – i.e. recognition of several episodes of the same show, and speaker diarization across these episodes, linking speakers. The longitudinal tasks also offered the opportunity for systems to make use of supplied metadata including show title, genre tag, and date/time of transmission. This paper describes the task data and evaluation process used in the MGB challenge, and summarises the results obtained.

***Index Terms***— Speech recognition, broadcast speech, transcription, multi-genre, longitudinal, diarization, alignment

## 1. INTRODUCTION

The Multi-Genre Broadcast (MGB) Challenge at ASRU–2015 is a controlled evaluation of speech recognition, speaker diarization, and "lightly supervised" alignment using BBC TV recordings. The evaluation builds on a broad, multi-genre dataset, spanning the whole range of BBC TV output. The challenge used a training set of about 1,600 hours of broadcast audio, together with several hundred million words of subtitle text for language modelling, provided by the BBC. Transcriptions for the acoustic training data took the form of the broadcast subtitles which have a word error rate of about 32.9% (25.8% due to deletions) compared with verbatim transcripts, computed on `dev.full` (see Table 1), with a high variance across shows.

The MGB challenge had four main evaluation conditions:

1. **Speech-to-text transcription** of broadcast audio;
2. **Alignment** of broadcast audio to subtitles;
3. **Longitudinal speech-to-text transcription** of a sequence of episodes from the same series;
4. **Longitudinal speaker diarization** and linking, requiring the identification of speakers across multiple recordings.

Authors listed alphabetically. Supported by EPSRC Programme Grant grant EP/I031022/1, Natural Speech Technology (NST). The NST research data collection may be accessed at `http://datashare.is.ed.ac.uk/handle/10283/786`

The acoustic model and language model training data was fixed, with no additional data or annotations allowed to be used. This enabled the different models and algorithms used by MGB challenge participants to be directly compared, without needing to take account of different training data. It also opened the challenge to any research team, since the data was made available free of charge by the BBC (for the purpose of the MGB Challenge). A novel aspect of the MGB challenge was the exploration of speech recognition and speaker diarization in a longitudinal setting – i.e. recognition of several episodes of the same show, and speaker diarization across these episodes, linking speakers. The longitudinal tasks also offered the opportunity for systems to make use of supplied metadata including show title, genre tag, and date/time of transmission.

In this paper we describe the challenge data and metadata, with a focus on metadata refinement and data selection. We discuss the four evaluation conditions in greater detail, and also outline the baseline systems provided for each task. We then outline the different systems that participants developed for the challenge, and give an overview of challenge results.

## 2. MGB CHALLENGE DATA

The MGB Challenge used acoustic and language model training data provided to participants by the BBC under a non-commercial use license, which is summarised in Table 1. It included:

- Approximately 1,600 hours of broadcast audio taken from seven weeks of BBC output across four TV channels (BBC1, BBC2, BBC3, BBC4) over the period 1 April 2008 – 19 May 2008 (around 70–80% of the total broadcast time is speech.[1]);
- Subtitles (closed captions) for the provided audio as originally broadcast on TV, accompanied by baseline lightly-supervised alignments using an ASR system, with confidence measures;
- About 640 million words of BBC subtitles collected from 1979 – 2008. These are pre-recorded subtitles, with near duplicates (by ID) removed and no live subtitles included.

Both the training acoustic data and the subtitles were filtered to avoid partial overlap with development and evaluation data.

In addition participants were offered a non-commercial license to the and-compiled Combilex British English lexicon[2] [1, 2].

[1] The original set of data for this period contained 1,974 hours of audio, obtained from 2,759 shows; we have removed repeated shows and shows with damaged subtitle files.

[2] `http://www.cstr.ed.ac.uk/research/projects/combilex/`

| 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 | 3583 | 51466 |
| dev.long | 19 | 12 | 9 | 5 962 | 72 884 |
| eval.std | 16 | 11 | | | |
| eval.long | 19 | 14 | | | |

**Table 1**: Training, development and evaluation data for the MGB Challenge. train.short and dev.short are subsets of train.full and dev.full respectively. Evaluation tasks 1 (transcription) and 2 (alignment) use the same evaluation data (eval.std); evaluation tasks 3 (longitudinal transcription) and 4 (longitudinal diarization) use eval.long. dev.long contained 2 series each of 6 episodes, and 7 additional shows taken from 3 series. eval.long contained 2 series of 11 and 8 episodes.

### 2.1. Metadata

The subtitles provided with the training data included transcripts, speaker changes indicated by different text colours (used for subtitle display), time stamps, and other metadata such as an indication of music and sound effects, or indications of the way the text has been pronounced. The title, date, and time of transmission, TV channel, and genre of each show is also provided. The quality of the metadata varied considerably across genres and shows in terms of precision of the alignment, owing to varying subtitle time-lags, andd transcript reliability, owing to differences in the subtitle creation process (pre-recorded (offline) or live (re-speaking)). This metadata was refined in order to be used for the selection of training data.

To facilitate the task of the MGB challenge participants, we provide some output of the pre-processing we applied to the raw data. However participants were free to apply their own pre-processing to the raw subtitle data.

The metadata – including speaker changes, time stamps and transcripts – were first extracted from the BBC subtitle files and the transcripts were normalised. A two-step refinement procedure was carried out: alignment of the whole transcription for each show and computation of different measures for training data selection.

The alignment was based on a lightly supervised approach [3, 4]. Each audio file was segmented and the segments were clustered for speaker adaptation using the segmenter and clusterer part of the Cambridge University RT-04 diarisation system [5]. Each speech segment was decoded using a two-pass recognition framework [6, 7] including speaker adaptation, with the decoding employing a biased language model (LM) and tandem-SAT acoustic models trained on a subset of the training dataset. The biased LM was initially trained on the subtitle transcripts and interpolated with the overall language model, with a 0.9/0.1 interpolation weight ratio, resulting in an interpolated LM biased to the transcripts. The vocabulary was chosen to ensure coverage of words from the original transcripts. The decoder output was then compared with the original transcripts to identify matching sequences. Non-matching word sequences from the transcripts were force-aligned to the remaining speech segments. Once the whole transcript was aligned, each show was segmented according to silence duration and speaker change. The obtained segments were finally re-clustered.

A number of different measures were computed to facilitate the selection of training data. First, confusion networks were used for minimum word error rate decoding of the aligned segments considering the biased LM. The estimates of the word posterior probabilities encoded in the confusion networks could be used directly as confidence scores (which are essentially word-level posteriors), but they tended to over-estimate the true posteriors. To compensate for this, a decision tree was trained on a reference dataset to map the estimates to confidence scores.

Two other measures were computed by scoring the decoding against the aligned transcripts used as reference. *Phone Matched Error Rate* (PMER) and *Word Matched Error Rate* (WMER) were calculated as traditional error rates but are described as matched error rates since there are not accurate transcripts to be used as reference.

Finally an *Average Word Duration* (AWD, in seconds) was computed for each aligned segment in order to reject those having too large a portion of non-speech audio. Those segments were mainly due to unreliable transcripts which failed to be matched and aligned during the refinement procedure. It was found preferable to keep them in the transcripts for possible future processing and to reject them during the selection process. These measures were made available to participants.

Data from "week 6" was initially aligned using a GMM-based system discriminatively-trained on about 18 hours of hand-transcribed BBC Radio 4 data, plus 11 hours of subtitled TV data [8]. The resultant alignments were used to train a more elaborate GMM-based system with tandem features and speaker adaptive training, using $\texttt{WMER} \leq 40\%$ and $\texttt{AWD} \leq 1s$, which was then used to align all the MGB challenge data.

### 2.2. Data selection

In this subsection we show some examples of data selection according to the different measures provided in the refined transcripts. Without selection, the training set has a duration of 1197 hours.

**Selection according to average word duration:** A few segments contain a large portion of non-speech events mainly due to unreliable transcription which is not matched and aligned during the refinement procedure. Those segments can be detected and rejected according to the average word duration measure. At the top of Figure 1 we present the segment distribution according to the AWD value. At the bottom of the same figure, we present cumulative distribution of the selected training set according to a threshold on the average word duration. According to those plots, $0.2 < \texttt{AWD} < 0.7$ is a reasonable range for data selection and is used in the following examples. By doing so, we reject 16% of the 1197 hours leading to 1005 hours of training data.

**Selection according to phone and word MER:** The selection can also be done according to the value of WMER or PMER [9]. In Figure 2, we present the cumulative duration of the selected training data according to a threshold on PMER and WMER. For instance, selecting segments having $\texttt{PMER} \leq 40\%$ leads to 700 hours of train-

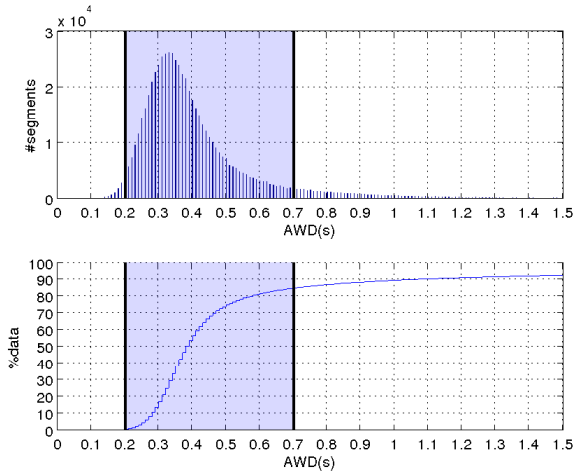


**Fig. 1**: Average word duration: *top*: segment distribution according to AWD value; *bottom*: cumulative duration in percentage of data of the selected training set according to threshold on AWD. The shaded portion corresponds to the region ($0.2 < \texttt{AWD} < 0.7$).

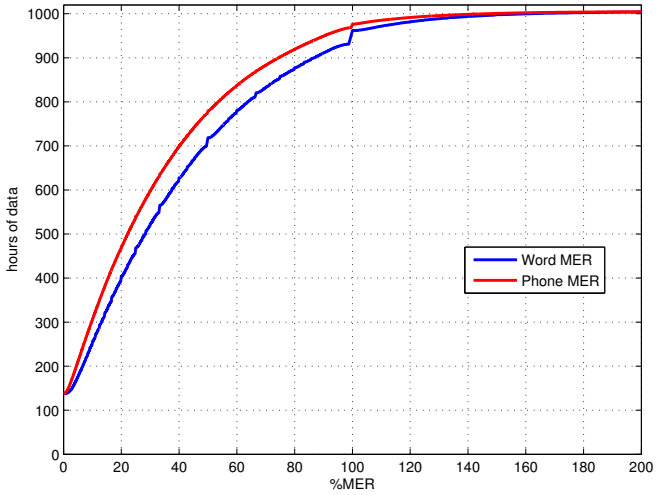


**Fig. 2**: Cumulative duration of the selected training data according to a threshold on PMER (red line) and WMER (blue line).

ing data. Note that considering phone instead of word level allows a significant increase of quantity of data for a given threshold.

MER-based selection can be refined by considering the genre of the shows. Figure 3 shows the cumulative duration of the selected training data by genre which varies significantly given the PMER threshold. The output of the lightly supervised decoding may be used as training material [10]. The selection of recognition hypothesis can be done by comparison with the original transcripts or using a confidence score.

### 2.3. Development and evaluation data

Two hand-transcribed development sets were also provided – one for the standard transcription and the alignment tasks (`dev.full`), and a second development set for the longitudinal transcription and diarization tasks (`dev.long`). Two evaluation data sets were released during the evaluation period (`eval.std` and `eval.long`).

The development and evaluation data sets were manually transcribed by two people. The data was supplied with time-aligned subtitles which were corrected to be verbatim transcriptions, using the AMI transcription guidelines [11]. It took an average 8 hours to transcribe 1 hour of broadcast data. To ascertain the quality of the transcriptions three 1-hour programs were cross-coded by both transcribers and 96% agreement was achieved.

## 3. EVALUATION TASKS

The MGB Challenge featured four evaluation tasks: for each the only allowable acoustic and language model training data was that specified above. To enable comparability, there was no option for participants to bring additional training data to the evaluation. Use of the provided other resources (e.g. dictionary) was optional.

### 3.1. Speech-to-text transcription

This is a standard speech transcription task operating on a collection of whole TV shows drawn from diverse genres. Scoring required ASR output with word-level timings. Segments with overlap were

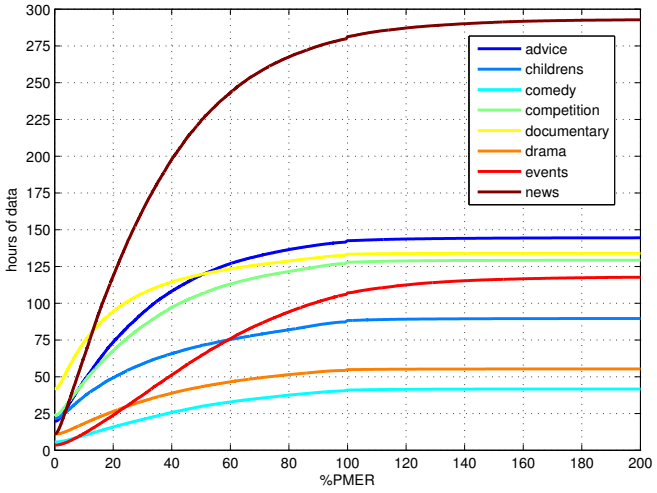


**Fig. 3**: Cumulative duration of the selected training data (1005 hours) according to a threshold on PMER depending on genre.

ignored for scoring purposes (where overlap is defined to minimise the regions removed – at word level where possible). Speaker labels were not required in the hypothesis for scoring.

For the evaluation data, show titles and genre labels were supplied. Some titles appeared in the training data, some were new. All genre labels were seen in the training data. Other metadata present in the development data was not supplied for the evaluation data. Speakers may be shared across training and evaluation data, and participants were free to automatically identify these themselves and make use of the information. Each show in the evaluation set was to be processed independently, so it was not possible to link speakers across shows. Systems for speech/silence segmentation were trained only on the official training set. A baseline speech/silence segmentation and speaker clustering for the evaluation data was supplied.

| System | 3gram | 4gram |
|---|---|---|
| GMM SI | 53.1 | - |
| + fMLLR | 51.3 | 48.5 |
| DNN CE 1 | 40.9 | 37.4 |
| DNN CE 2 | 40.5 | 37.1 |
| + sMBR | 37.1 | 33.7 |

**Table 2**: Baseline system results (WER%) on `dev.full`, decoding with a pruned 3gram LM; and rescoring with a 4gram LM

### 3.2. Alignment

In the alignment task participants were supplied with a tokenised version of the subtitles (the script) as they originally appeared on TV, without any timing information. The task was to align these subtitles to the spoken audio at word level, where possible. It should be noted that TV captioning often differs from the actual spoken words for a variety of reasons: there may be edits to enhance clarity, paraphrasing, and deletions where the speech is too fast. There may be words in the captions not appearing in the reference, and equally words missing in the subtitles that were spoken. As in the transcription task, it was possible to make use of the show title and genre labels, and any automatic speaker labelling across shows that participants choose to generate. Speaker change information was supplied as in the original captions.

Scoring was performed by a script that calculated a precision/recall measure, derived from automatic alignment of a careful manual transcription. A word is considered to be a match if both start and end times fall within a 100ms window of the associated reference word. Participants were only allowed to include words from the script in their output, however words could be removed. Their scoring therefore only matches the system output with the script prior to the matching process. As before, output was filtered to remove words falling in regions of overlapped speech.

### 3.3. Longitudinal speech-to-text transcription

This task aims to evaluate ASR in a realistic longitudinal setting – processing complete TV series, where the output from shows broadcast earlier may be used to adapt and enhance the performance of later shows. The evaluation data consisted of a collection of TV series with title and genre labels. Initial models were trained on the same data as for the standard transcription task. Systems then processed each series in strict broadcast order, producing output for each show using only the initial models, and optionally, adaptation data from shows that have gone before.

### 3.4. Longitudinal speaker diarization

This task evaluated speaker diarization in the longitudinal setting. Systems aimed to label speakers uniquely across the whole series. Speaker labels for each show were obtained using only material from the show in question, and those broadcast earlier in time. Participants were not able to use external sources of training data in their diarization systems (e.g. for building i-vector extractors).

## 4. BASELINE SYSTEM

We created an open recipe to enable participants to build a baseline ASR system using the Kaldi toolkit [12], as well as XMLStarlet[3], and the SRILM[4] and IRSTLM[5] toolkits. This baseline system simplified and automated the data pre-processing tasks, thus allowing participants to focus on more advanced aspects of ASR model-building. The version of the recipe distributed during the challenge constructed a speaker-adapted GMM system, which we have since extended to more state-of-the-art DNN acoustic models. The main features of the system were:

- Three-state cross-word triphone HMMs (11,500 tied states, 200,000 Gaussians in total);
- Maxiumum-likelihood training using PLP features with LDA and MLLT applied, with speaker-adaptive training using one FMLLR transform per speaker. For speaker labels, the automatic speaker-clustering supplied in the metadata was used;
- A language model trained on a normalised version of the supplied BBC subtitle text [13]. By default the recipe builds a 3-gram LM pruned with threshold of $10^{-7}$;
- A vocabulary of the 150k most frequently-occurring words in the text. An additional automatically-generated lexicon was supplied containing words is not present in Combilex. These pronunciations are generated automatically using Sequitur [14], or using a rule-based method for acronyms.
- An option for the user to automatically select training data according to a `WMER` threshold at the per-utterance level, based on the lightly-supervised transcriptions supplied. The default setting of 10% yields 260 hours of speech.

We also supplied a baseline speech/non-speech segmentation and speaker clustering for each show in the development and evaluation sets. The segmentation used a 2-layer DNN trained to detect 2 outputs (speech and non-speech). The posteriors were fed to a Viterbi decoder using a 2-state HMM to produce a smoothed segmentation. Clustering was based on an unsupervised iterative procedure where speakers were clustered using the Bayesian Information Criterion. When using the baseline segmentation, the performance of DNN-based systems with no speaker adaptation was worse by around 5% absolute WER compared to a gold-standard segmentation derived from the reference transcriptions.

Table 2 shows the performance on the Task 1 development set of the baseline recipe systems trained on 260hrs of data. The speaker-independent (SI) system uses the supplied baseline segmentation; adapted systems used the baseline speaker clustering. Our experience suggests that around 2% absolute WER is gained by expanding the training set to 700hrs by increasing the MER threshold to 40%. We also show the results from applying a standard DNN training recipe with CE training followed by sMBR sequence training [15]. Two iterations of CE training are used, with state alignments regenerated after the first iteration. The final system scores 34.9% WER on the evaluation set. Table 3 shows the performance of the baseline speaker segmentation and clustering systems for unlinked speaker

| | `dev.long` | `eval.long` |
|---|---|---|
| missed speech | 11.4 | 6.1 |
| false speech | 1.3 | 3.8 |
| speaker error | 34.1 | 37.2 |
| DER | 46.9 | 47.1 |

**Table 3**: Unlinked speaker diarization results (DER/%) for baseline systems on `dev.long` and `eval.long`

[3] `http://xmlstar.sourceforge.net/`

[4] `http://www.speech.sri.com/projects/srilm/`

[5] `https://hlt.fbk.eu/technologies/irstlm`

diarization on the development and evaluation data.

## 5. SUBMITTED SYSTEMS AND RESULTS

19 teams submitted systems to the MGB challenge, across the four tasks. We have attempted to highlight key features of the systems below. Detailed system descriptions will be publicly available at `http://mgb-challenge.org/` after ASRU.[6]

**Task 1: Speech-to-text transcription**

- *Brno University of Technology (BUT)*: train on matched portions of lightly supervised transcriptions (800 hrs). DNNs with sMBR sequence training, trained on fMLLR features
- *CRIM*: train on 1070hrs, utterances with higher WMER used only to update hidden layers; genre-dependent LMs used for rescoring. DNNs trained on TRAP+ivector features (various systems combined). Used LIUM's diarization in joint system.
- *Inferret*: 152 hours of data; DNNs trained on fMLLR features with 4 sMBR iterations
- *Intelligent Voice (IV)*: 170k utterances selected following data cleanup, with 1% WER improvement over WMER=50% selection. Kernel additive modelling investigated, but not effective. Final system used time-delay DNNs with i-vectors features included. Errors with acronym splitting.
- *LIMSI*: one-pass system, decoding entire shows, makes use of baseline segmentation and their own, three different NNLMs, additional data cleaning and normalisation. Incremental selection of training data using own alignment tool - 900 hrs finally. Features are BNs derived from PLP+TRAP features.
- *LIUM*: realignment of training data with own diarization; 5-gram feed-forward NNLM; AMs similar to CRIM.
- *NAIST*: p-norm DNNs with i-vector features; 338 hours of training data
- *NTU, Singapore*: 700 hrs of data; initial GMM trained on WMER=10% used to select remaining data with a confidence threshold. CE DNN training on full set; sequence training using only smaller most confident portion.
- *University of Cambridge (CU)*: Primarily HTK-based hybrid DNN and tandem systems via joint decoding. Trained on 700 hrs (PMER=30%). DNN-based segmenter. DNN adaptation by parametrised activation functions in Task 3 system. RNNLMs with adaptation. Also combination with Kaldi based CNN, DNN and LSTM systems.
- *University of Edinburgh (UE)*: data selected on WMER=40%. CNNs are trained on alignments generated from DNN. No sequence training.
- *University of Sheffield (SU)*: 1st pass - DNNs with PLP+bMMI features with global CMLLR xform. Speech segmented based on 1st-pass output. 2nd pass uses 3-complementary DNN systems (HTK and Kaldi), combined. Genre-dependent LMs used in some systems.

**Task 2: Alignment**

- *CRIM*: forced alignment with Task 1 AMs using a wide pruning beam – except used recognition output for two shows.
- *NHK*: DNN-based AM. Forced alignement with ASR putput used when subtitles not given.
- *Quorate / UEDIN*: data selected on WMER=40%; DNNs with SMBR sequence training; 2-pass alignment with factor-transducers
- *CU*: DNN-based segmentation. Lightly supervised decoding with SI DNN. Text-aligned to original script to get anchor points, followed by forced alignment. Removed words if suspect insertions or substitutions by comparison with confidence-marked lightly supervised output.
- *SU*: lightly supervised decoding. After alignment with the text, force-alignment is used to obtain final word timings.
- *Vocapia / LIMSI*: Rover combination of task 1 primary system with a system using LM biased towards show captions.

**Task 3: Longitudinal speech-to-text transcription**

- No participants attempted explicit longitudinal adaptation. There were some system improvements due to the additional week allowed for submitting results for this task.

**Task 4: Longitudinal speaker diarization**

- *IDIAP*: speaker clustering using a fusion of Information Bottleneck clustering and a traditional HMM-GMM method. Agglomerative clustering using i-vectors with the Hotteling t-square statistic distance metric.
- *Orange/LIUM*: segmentation based on ASR output; three clustering schemes were investigated, the best performing being i-vectors with PLDA distance metric
- *CU*: DNN-based segmentation including change point-detection and iterative agglomerative clustering to get homogeneous segments. Speaker clustering using feature warped data on a MAP-adapted UBM model for each cluster to optimise a cross-likelihood ratio (CLR). Linking uses complete-linking clustering with a distance measure based on CLR.
- *UE*: GMM-based agglomerative speaker clustering
- *SU*: clustering using the SHoUT toolkit; posteriors are generated from a Speaker Separation DNN, used as input to a speaker-state HMM
- *University of Zaragoza (UZ)*: i-vector approach using unsupervised version of PLDA with dirichelet prior, approximated with variational bayes

The results for all participants across all tasks are summarised in Table 4. For the recognition and alignment tasks the results are presented per show and as an overall average. The variance across shows is quite high: for example the most accurate system for task 1, has an average WER of 23.7%, with the WER per show varying from 10.4 – 41.4% across the 16 test shows.

## 6. CONCLUSIONS AND FUTURE CHALLENGES

The first MGB Challenge developed a process for evaluating systems for multi-genre broadcast speech recognition, diarization, and alignment, using fixed training sets for acoustic and language model training. We achieved wide participation from 19 teams. Recognition, alignment, and diarization of multi-genre broadcast speech is indeed a substantial challenge. Performance is highly variable across shows – for example, the WER by show varied from 10 – 40% for the most accurate system in the transcription task. Speaker diarization of this broadcast contentis considerably more difficult than what is typically addressed in the literature. Evaluation of alignment and longitudinal evaluation conditions were novel aspects. The alignment evaluation had a tight temporal constraint (looking for matches with 0.1s of the start/end times of each word) which was difficult. Finally, no team attempted direct longitudinal modelling for task 3, probably because of the tight deadlines of the challenge. We plan to continue the MGB Challenge, using the current tasks and training data. Possible extensions tasks include moving to significantly larger training set for speech-to-text transcription, and to extend the challenge to different languages.

[6]Citations to accepted system papers will be added to the final version of the paper.

| Show | CU | CRIM/LIUM | LIMSI | CRIM | SU | LIUM | UE | NAIST | NTU | BUT | IV | Inferret | *Average* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Daily Politics* | 10.4 | 13.0 | 11.8 | 13.5 | 13.6 | 14.5 | 15.4 | 17 | 16.8 | 19.4 | 20.1 | 23.2 | 15.7 |
| *Magnetic North* | 11.6 | 12.5 | 13.0 | 13.8 | 16.9 | 14.9 | 14.1 | 13.2 | 18.8 | 22.0 | 20.1 | 20.9 | 16.0 |
| *Dragons' Den* | 11.5 | 13.0 | 13.5 | 13.9 | 14.3 | 14.1 | 15.8 | 15.7 | 21.2 | 22.9 | 21.8 | 24.7 | 16.9 |
| *Eggheads* | 14.1 | 16.9 | 17.6 | 17.5 | 19.2 | 19.8 | 19.1 | 20.5 | 24.5 | 26.1 | 24.9 | 25.8 | 20.5 |
| *Athletics London* | 14.7 | 16.8 | 15.8 | 17.7 | 20.7 | 19.4 | 19.3 | 21.8 | 20.6 | 25.1 | 26.2 | 30.2 | 20.7 |
| *Point of View* | 13.5 | 17.1 | 14.2 | 17.4 | 18.9 | 23.4 | 21.7 | 21.8 | 22.2 | 25.6 | 27.1 | 32.8 | 21.3 |
| *Syd Barrett* | 21.3 | 22.7 | 23.2 | 23.9 | 24.0 | 25.7 | 28.4 | 29.4 | 30.8 | 32.8 | 36.2 | 36.4 | 27.9 |
| *Top Gear* | 21.8 | 25.7 | 26.3 | 27.6 | 27.2 | 29.3 | 31.4 | 28.9 | 36.1 | 37.8 | 38.3 | 39.7 | 30.8 |
| *Blue Peter* | 24.6 | 26.6 | 25.6 | 27.8 | 28.4 | 30.4 | 31.1 | 31.1 | 34.3 | 38.3 | 37.9 | 44.4 | 31.7 |
| *Legend of the Dragon* | 21.7 | 25.2 | 26.1 | 26.0 | 25.2 | 31.7 | 29.9 | 33.0 | 41.7 | 42.6 | 43.8 | 39.0 | 32.2 |
| *The North West 200* | 27.7 | 30.4 | 31.6 | 31.2 | 32.2 | 34.4 | 36.9 | 38.3 | 43.4 | 45.6 | 46.9 | 49.1 | 37.3 |
| *Holby City* | 32.1 | 36.6 | 40.9 | 38.0 | 39.3 | 41.7 | 39.1 | 38.4 | 47.3 | 49.8 | 48.6 | 54.5 | 42.2 |
| *The Wall* | 33.7 | 39.2 | 38.7 | 40.4 | 40.8 | 43.8 | 42.6 | 42.7 | 46.1 | 48.8 | 51.2 | 53.2 | 43.4 |
| *One Life Special Mum* | 35.3 | 40.2 | 40.5 | 42.2 | 42.2 | 43.8 | 45.1 | 45.7 | 49.6 | 51.4 | 53.8 | 52.8 | 45.2 |
| *Goodness Gracious Me* | 37.2 | 36.5 | 41.9 | 37.6 | 42.5 | 45.1 | 46.0 | 45.3 | 48.7 | 52.9 | 55.7 | 54.1 | 45.3 |
| *Oliver Twist* | 41.4 | 44.2 | 50.1 | 45.9 | 49.4 | 52.2 | 49.2 | 48.9 | 55.4 | 58.8 | 58.6 | 60.2 | 51.2 |
| ***Overall WER (%)*** | **23.7** | **26.6** | ***27.5*** | ***27.8*** | **28.8** | ***30.4*** | ***30.9*** | ***31.2*** | **35.5** | **38.0** | **38.7** | **40.8** | |

**Task 1: Speech-to-text transcription (WER/%)**. Italics in adjacent systems indicates no significant difference at the 1% level; all differences between systems two or more places apart in ranking were found to be statistically significant at 1% (Matched Pairs Sentence Segment Word Error Test).

| Show | CU* | Quorate/UE | CRIM | Vocapia/LIMSI | SU | NHK | *Average* |
|---|---|---|---|---|---|---|---|
| *Magnetic North* | 0.977 | 0.962 | 0.971 | 0.920 | 0.973 | 0.944 | 0.958 |
| *Dragons' Den* | 0.946 | 0.944 | 0.912 | 0.935 | 0.934 | 0.910 | 0.930 |
| *Points of View* | 0.957 | 0.941 | 0.934 | 0.888 | 0.929 | 0.907 | 0.926 |
| *Eggheads* | 0.938 | 0.904 | 0.920 | 0.894 | 0.892 | 0.887 | 0.906 |
| *Syd Barrett* | 0.892 | 0.887 | 0.850 | 0.877 | 0.874 | 0.824 | 0.867 |
| *Daily Politics* | 0.901 | 0.887 | 0.870 | 0.888 | 0.792 | 0.849 | 0.865 |
| *Legend of the Dragon* | 0.899 | 0.867 | 0.879 | 0.833 | 0.876 | 0.792 | 0.858 |
| *Top Gear* | 0.891 | 0.886 | 0.826 | 0.876 | 0.855 | 0.786 | 0.853 |
| *Blue Peter* | 0.883 | 0.860 | 0.863 | 0.854 | 0.799 | 0.803 | 0.844 |
| *Athletics London* | 0.886 | 0.848 | 0.822 | 0.875 | 0.803 | 0.801 | 0.839 |
| *Holby City* | 0.883 | 0.880 | 0.889 | 0.785 | 0.780 | 0.735 | 0.825 |
| *Oliver Twist* | 0.863 | 0.865 | 0.849 | 0.738 | 0.787 | 0.733 | 0.806 |
| *One Life Special Mum* | 0.856 | 0.844 | 0.860 | 0.766 | 0.767 | 0.731 | 0.804 |
| *Goodness Gracious Me* | 0.855 | 0.832 | 0.835 | 0.761 | 0.790 | 0.722 | 0.799 |
| *The North West 200* | 0.855 | 0.801 | 0.822 | 0.794 | 0.737 | 0.720 | 0.788 |
| *The Wall* | 0.787 | 0.773 | 0.760 | 0.742 | 0.696 | 0.620 | 0.730 |
| ***Overall f-score*** | **0.893** | **0.877** | **0.863** | **0.846** | **0.834** | **0.797** | |

**Task 2: Alignment (f-score)**. (*: After a bugfix, the CU system has an overall f-score of 0.900).

| Show/Episode | CU | SU | UE | *Average* |
|---|---|---|---|---|
| *Celebrity Masterchef* Ep 1 | 18.9 | 23.3 | 25.1 | 22.4 |
| Ep 2 | 15.4 | 20.5 | 21.2 | 19.0 |
| Ep 3 | 15.7 | 20.6 | 22.1 | 19.5 |
| Ep 4 | 17.4 | 21.4 | 24.1 | 21.0 |
| Ep 5 | 13.8 | 19.2 | 19.3 | 17.4 |
| Ep 6 | 19.5 | 24.1 | 26.9 | 23.5 |
| Ep 7 | 20.2 | 26.6 | 27.8 | 24.9 |
| Ep 8 | 15.5 | 21.8 | 24.0 | 20.4 |
| Ep 9 | 23.5 | 30.7 | 34.9 | 29.7 |
| Ep 10 | 23.6 | 30.3 | 33.1 | 29.0 |
| Ep 11 | 13.7 | 17.5 | 18.5 | 16.6 |
| *The Culture Show Uncut* Ep 1 | 20.3 | 27.5 | 26.0 | 24.6 |
| Ep 2 | 22.5 | 28.4 | 29.0 | 26.6 |
| Ep 3 | 14.9 | 20.7 | 21.9 | 19.2 |
| Ep 4 | 22.2 | 27.7 | 27.9 | 25.9 |
| Ep 5 | 23.3 | 30.7 | 29.0 | 27.7 |
| Ep 6 | 19.6 | 23.3 | 25.0 | 22.6 |
| Ep 7 | 21.0 | 25.4 | 26.5 | 24.3 |
| Ep 8 | 21.5 | 27.5 | 29.6 | 26.2 |
| ***Overall WER (%)*** | **19.3** | **24.8** | **26.3** | |
| *Task 1 WER (%)* | 22.1 | 28.8 | 29.7 | |

**Task 3: Longitudinal speech-to-text transcription (WER/%)**.

| | CU | Orange/ LIUM | UZ | SU | UE | IDIAP |
|---|---|---|---|---|---|---|
| Linked Speaker Diarization | | | | | | |
| missed speech | 4.5 | 10.0 | 6.1 | 2.1 | 6.0 | 6.0 |
| false speech | 2.8 | 1.3 | 4.0 | 4.8 | 4.1 | 4.1 |
| speaker error | 40.2 | 38.5 | 40.4 | 50.3 | 48.4 | 57.9 |
| diarization error | **47.5** | **49.8** | **50.5** | **57.2** | **58.5** | **68.1** |
| masterchef error | 51.9 | 51.4 | 55.3 | 64.6 | 61.6 | 67.9 |
| culture show error | 40.2 | 48.1 | 42.7 | 45.3 | 53.4 | 68.3 |
| Unlinked Speaker Diarization | | | | | | |
| missed speech | 4.5 | 10.0 | 6.1 | 2.1 | 6.0 | 6.0 |
| false speech | 2.6 | 1.2 | 3.9 | 4.7 | 3.9 | 4.0 |
| speaker error | 33.1 | 33.5 | 33.0 | 43.2 | 41.3 | 44.4 |
| diarization error | **40.2** | **44.7** | **43.0** | **50.1** | **51.2** | **54.4** |
| masterchef error | 44.6 | 47.9 | 47.6 | 55.8 | 52.8 | 58.2 |
| culture show error | 33.1 | 39.5 | 35.2 | 40.8 | 48.6 | 48.1 |

**Task 4: Longitudinal speaker diarization (DER/%)**.

**Table 4**: Evaluation results for the four MGB Challenge tasks. The error rate of applying the Task 3 ASR systems to Task 1 is also shown.

## 7. REFERENCES


[1] K Richmond, R Clark, and S Fitt, "Robust LTS rules with the Combilex speech technology lexicon," in *Proc Interspeech*, 2009, pp. 1295–1298.

[2] K Richmond, R Clark, and S Fitt, "On generating Combilex pronunciations via morphological analysis," in *Proc Interspeech*, 2010, pp. 1974–1977.

[3] N. Braunschweiler, M.J.F. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. Interspeech*, 2010, pp. 2222–2225.

[4] P. Lanchantin, P. J. Bell, M. J. F. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, M. S. Seigel, P. Swietojanski, and P. C. Woodland, "Automatic transcription of multi-genre media archives," in *Proc. of SLAM workshop*, Marseille, France, 2013.

[5] S.E. Tranter, M.J.F. Gales, R. Sinha, S. Umesh, and P.C. Woodland, "The development of the Cambridge University RT-04 diarisation system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[6] G. Evermann and P.C. Woodland, "Design of fast LVCSR systems," in *Proc. ASRU Workshop*, 2003.

[7] M.J.F. Gales, D.Y. Kim, P.C. Woodland, H.Y. Chan, D. Mrva, R. Sinha, and S.E. Tranter, "Progress in the CU-HTK Broadcast News Transcription System," in *IEEE Trans. on Audio, Speech, and Language Processing*, 2006, vol. 14, pp. 1513–1525.

[8] P.J. Bell, M.J.F. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P.C. Woodland, "Transcription of multi-genre media archives using out-of-domain data," in *Proc. SLT*, 2012.

[9] Y. Long, M. J. F. Gales, P. Lanchantin, X. Liu, M. S. Seigel, and P. C. Woodland, "Improving lightly supervised training for broadcast transcriptions," in *Proc. Interspeech*, 2013.

[10] L. Lamel, J.L. Gauvain, and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," in *Computer Speech and Language*, 2002, vol. 16, pp. 115–129.

[11] J Moore, M Kronenthal, and S Ashby, "Guidelines for AMI speech transcriptions," 2005, `http://groups.inf.ed.ac.uk/ami/corpus/Guidelines/speech-transcription-manual.v1.2.pdf`.

[12] D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlıcek, Y Qian, P Schwarz, J Silovský, G Semmer, and K Veselý, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[13] P.J. Bell, F. McInnes, S. Gangireddy, M. Sinclair, A. Birch, and S. Renals, "The UEDIN English ASR system for the IWSLT 2013 evaluation," in *Proc. International Workshop on Spoken Language Translation*, 2013.

[14] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, 2008.

[15] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech*, 2013.