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, …
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

Word MER
Phone MER
Training data by genre

%PMER

- advice
- childrens
- comedy
- competition
- documentary
- drama
- events
- news

hours of data

0 20 40 60 80 100 120 140 160 180 200

Training data by genre
Training data by genre

%PMER

%data

all (1005h)
advice (145h)
childrens (90h)
comedy (42h)
competition (129h)
documentary (134h)
drama (55h)
events (118h)
news (293h)
## MGB Data

<table>
<thead>
<tr>
<th>Data set</th>
<th>num Shows</th>
<th>Total duration(h)</th>
<th>Aligned speech(h)</th>
<th>num Aligned segments</th>
<th>num Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>train.full</td>
<td>2193</td>
<td>1580</td>
<td>1197</td>
<td>635 827</td>
<td>10 566 560</td>
</tr>
<tr>
<td>dev.full</td>
<td>47</td>
<td>28</td>
<td>20</td>
<td>13 165</td>
<td>183 811</td>
</tr>
<tr>
<td>train.short</td>
<td>274</td>
<td>199</td>
<td>152</td>
<td>81 027</td>
<td>1 373 913</td>
</tr>
<tr>
<td>dev.short</td>
<td>12</td>
<td>8</td>
<td>6</td>
<td>3 583</td>
<td>51 466</td>
</tr>
<tr>
<td>dev.long</td>
<td>19</td>
<td>12</td>
<td>9</td>
<td>5 962</td>
<td>72 884</td>
</tr>
<tr>
<td>eval.std</td>
<td>16</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eval.long</td>
<td>19</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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
Finally, two genres did not present a specific pattern. The two "Events" shows achieve different WERs and this seems to relate to the type of live event broadcast: quieter background in athletics competition ("Athletics: The London 10000") compared to noisier background in motorbike racing ("The North West 200"). The two "News" shows present the highest variability, as they range from a more conventional studio–based daily news show ("The Daily Politics") to a more factual documentary–like show where one of the main characters has a severe disability ("One Life Special: Mum and Me").

This analysis indicates that more advances in ASR are necessary to reduce the performance of shows with challenging acoustic and speaking conditions. Successful transcription of media data requires all genres to produce a more balanced performance. The achieved results indicate that shows from "Competition", "Documentary", "News" and even "Events" genres can produce less than 20% WER. But with genres such as "Comedy" and "Drama" obtaining more than 40% WER, a robust system for audio analysis of media data still needs more improvements.

### Task 2: Lightly supervised alignment

The results for the 6 submitted systems for Task 2 are presented in Table 22.2. The results are ranked in terms of F1 score, with precision and recall also being detailed. The best system was submitted by CU, with an F1 score of 0.8927, and all systems achieved an F1 score of 0.80 or above. The best precision was achieved by CU (0.92) and the best recall by Quorate (0.89).

Comparing the ranking of the results in Table 22.2 with the 3 strategies in the systems submitted by the participants shows that the best performing system (CU) employs a lightly supervised decoding approach, although the other 2 participants that used a similar approach (SU and Vocapia) rank 4th and 5th. This may be due to the better
Results by show – Transcription

Daily Politics
Dragon's Den
Athletics
Top Gear
Blue Peter
Oliver Twist
Holby City
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).

<table>
<thead>
<tr>
<th>Participant</th>
<th>Substitutions</th>
<th>Deletions</th>
<th>Insertions</th>
<th>Word Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>CU</td>
<td>8.6%</td>
<td>7.9%</td>
<td>2.8%</td>
<td>19.3%</td>
</tr>
<tr>
<td>SU</td>
<td>11.7%</td>
<td>9.8%</td>
<td>3.2%</td>
<td>24.8%</td>
</tr>
<tr>
<td>UE</td>
<td>10.9%</td>
<td>12.6%</td>
<td>2.8%</td>
<td>26.3%</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>Cam</th>
<th>Quorate/Edin</th>
<th>CRIM</th>
<th>Vocapia/LIMSI</th>
<th>Sheff</th>
<th>NHK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Table 22.2

<table>
<thead>
<tr>
<th>Participant</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CU</td>
<td>0.9165</td>
<td>0.8701</td>
<td>0.8927</td>
</tr>
<tr>
<td>Quorate</td>
<td>0.8651</td>
<td>0.8899</td>
<td>0.8773</td>
</tr>
<tr>
<td>CRIM</td>
<td>0.8451</td>
<td>0.8822</td>
<td>0.8632</td>
</tr>
<tr>
<td>Vocapia</td>
<td>0.8847</td>
<td>0.8109</td>
<td>0.8462</td>
</tr>
<tr>
<td>SU</td>
<td>0.8654</td>
<td>0.8052</td>
<td>0.8342</td>
</tr>
<tr>
<td>NHK</td>
<td>0.7690</td>
<td>0.8268</td>
<td>0.7969</td>
</tr>
</tbody>
</table>

The performance of the decoding systems developed by CU as seen in the results in Task 1. The system based on factor transducer (Quorate) ranks second, which indicates that this is also a good performing strategy. Finally, both systems that were based mostly on Viterbi forced alignment (CRIM and NHK) also perform differently, 3nd and 6th. This seems to indicate that more complex systems based on lightly supervised decoding or decoding with automata are the most successful strategies in the lightly supervised alignment tasks.

As for Task 1, the results broken down per show are given in Figure 22.2. These are average, lowest and highest F1 scores across all the 6 participants. The show with the highest score, "Jonathan Meades Magnetic North" achieves an average F1 score of 0.96, with a maximum individual score of 0.98. While the show with the lowest average score, "The Wall" achieves an average score of 0.73 and a lowest score of 0.62.

Regarding the influence of genre, the ranking of the genres is similar to that achieved in Task 1, with "Comedy" and "Drama" achieving the poorest scores and "Documentary" and "Competition" achieving the best ones.

Figure 22.2: Average, minimum and maximum F1 score per show in Task 2

Figures 22.1 and 22.2 hinted to a certain relationship between the results per show achieved in both Tasks 1 and 2. In general, genres and shows with low WERs in Task 1 also tend to have higher F1 scores in Task 2.
Results by show – Alignment

progress, it is Jimmy Carr. Keep
Transcription-Alignment Correlation

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

- Separate live subtitles and offline

- Increase WER by 1% gives 0.004 worse f-score

![Graph showing correlation between WER and F1 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
Results – Diarization

Unlinked
Linked

Cambridge
Orange/LIUM
Zaragoza
Sheffield
Edinburgh
Idiap

DER/%
MGB–2 (& beyond?)

- 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