Details of the MGB Challenge data preparation

P. Lanchantin, P. Bell,
J. Kilgour, X.Liu, O. Saz

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

27 May 2015

Dataset description & prepararation

Natural Speech Technology

- supply pre-processed data enabling participants to focus their time on novel research
- Task 1: **Transcription**
  - train (2193 episodes/ 1580h audio /1197h active speech) (subtitles)
  - dev (47e/28h/20h) (subtitles+manual)
  - eval (15e/11h) (manual)
- Task 2: **Alignment**
  - same as transcription dev+eval
- Task 3: **Longitudinal transcription**
  - dev: (19e/12h/8.5h) (subtitles+manual)
  - eval (19e/14h) (manual)
- Task 4: **Speaker diarisation and linking**
  - same as longitudinal dev+eval

Transcription refinement

Natural Speech Technology

- extracted from *TV-1week+TV-6week+Mediaeval* corpora
- provided transcriptions are generally **imperfect** and need to be refined to be used for the training of acoustic models (AM)
  - different source, constant lag, re-speaking ...
- forced-alignement of the transcript
  1. **automatic segmentation** into speech/nonspeech segments (rt-04)
  2. decoding of speech segments using **LM biased towards the transcripts**
  3. detection of **matching word sequences** between the decoded hypotheses and the transcript
  4. force-alignement of the remaining text on the remaining speech segments
  5. computation of confidence measure (**matching error rate** (Phone and Word MER) for each segment

Transcription refinement

Natural Speech Technology

- Iterative process
  - The same process is done iteratively using new acoustic models (AM) trained on the aligned transcrips

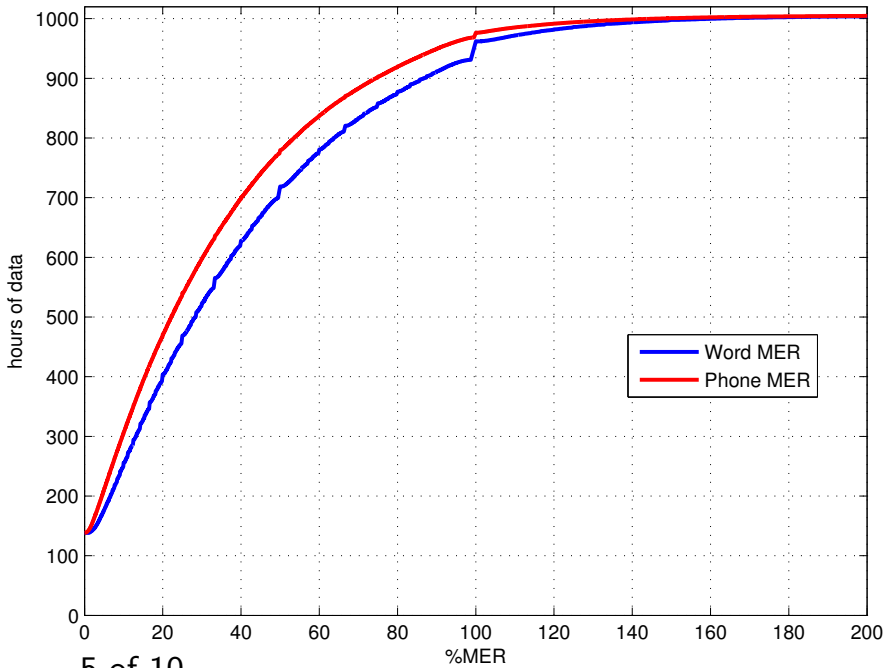
dataset	sel(h)	AM	LM	N-gram	Sub	Del	Ins	Err
cr4+who	21.63	mpe.0	LM8	tg(12)cn	29.3	28.5	2.1	**59.9**
cr4+who	21.63	mpe.1	LM8	tg(12)cn	31.7	20.8	3.2	**55.6**
1we+cr4+who	147.66	mpe.1	LM8	tg(12)cn	25.2	19.2	2.3	**46.7**
1we+cr4+who	147.66	tandem-sat.1	LM8	tg(21)cn	21.7	15.6	2.4	**39.7**
1we+cr4+who	147.66	tandem-sat.1	LM17	tg(21)cn	20.5	15.3	2.3	**38.1**
1we+cr4+who	147.66	tandem-sat.1	LM17	fg(27)cn	17.8	16.3	1.9	**36.0**

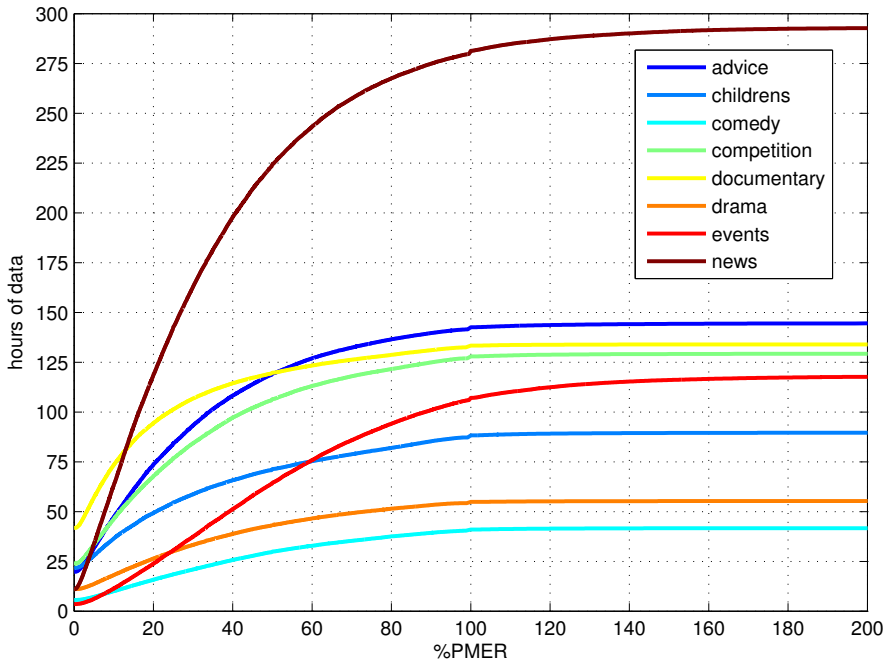
- decoding was run on the manual segmentation, scoring was done considering the **manual devset**, excluding overlapping segments
- **LM8**: 4-gram interpolated LM combining 4 LMs (BBC-1week, BBC-radio4/tv-drama, BNE data, UW web data) pruned at $1.0e^{-9}$
- **LM17**: 4-gram interpolated LM combining 5 LMs (BBC subtitles(680M) minus 2M words from the month of May 2008 to avoid overlap with the 1 week dev data, BBC-1week, BBC-radio4/tv-drama, BNE data, UW web data) pruned at $1.0e^{-9}$

Example: Refined transcription trainset

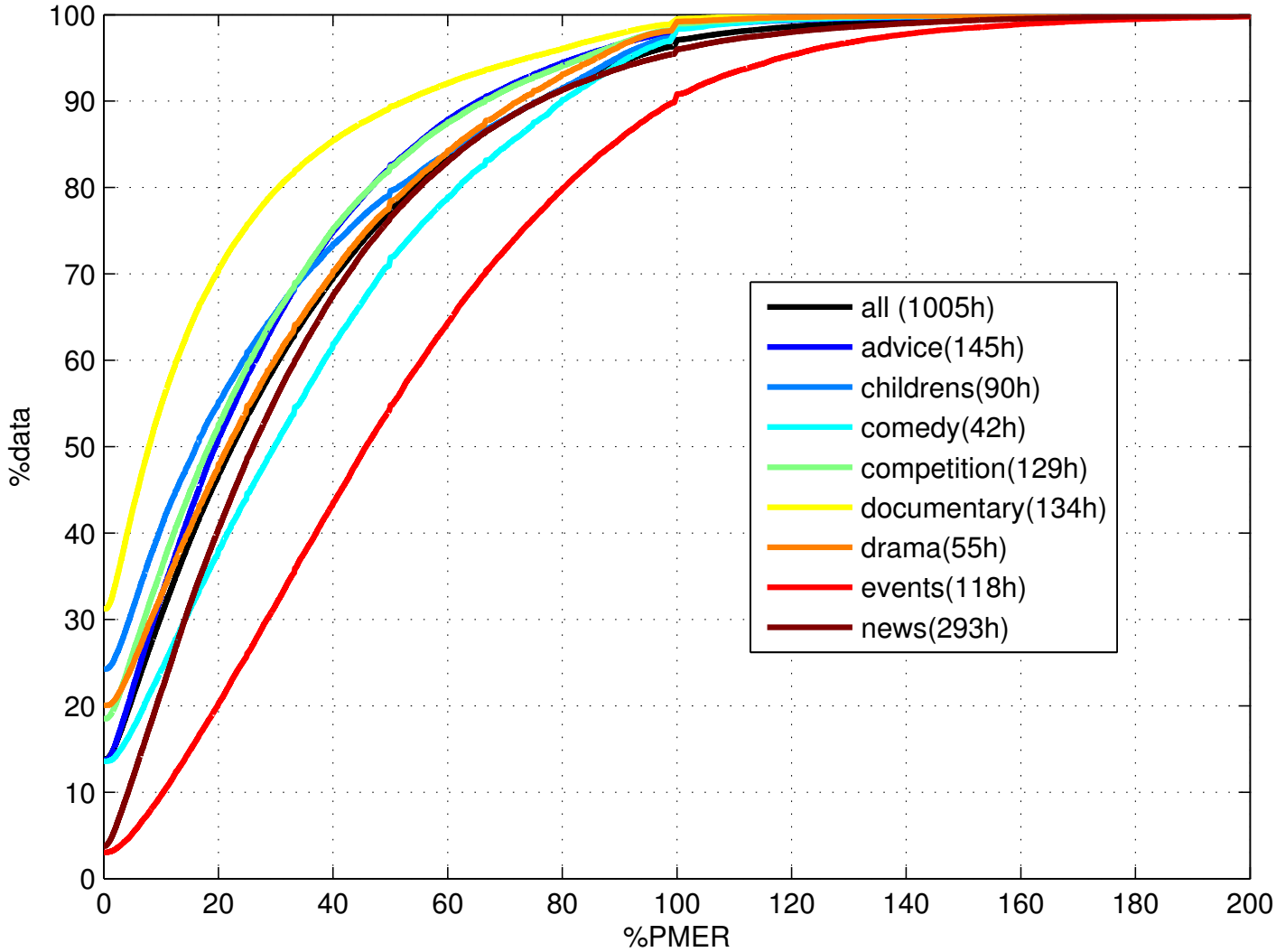
Natural Speech Technology

- 1197 hours of aligned transcripts/speech
  - **Matching Error Rate** (MER) used for **selection** of training data
  - *left*: cumulative duration(h) vs Word and phone MER
  - *right*: the cumulative duration(%) vs PMER for each genre: selection can be done by taking horizontal or vertical thresholds





Refined transcription trainset



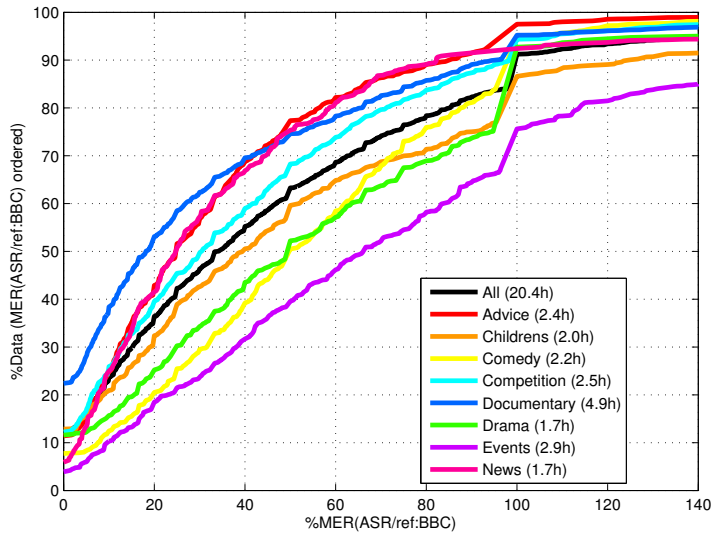
Data Selection analysis per genre

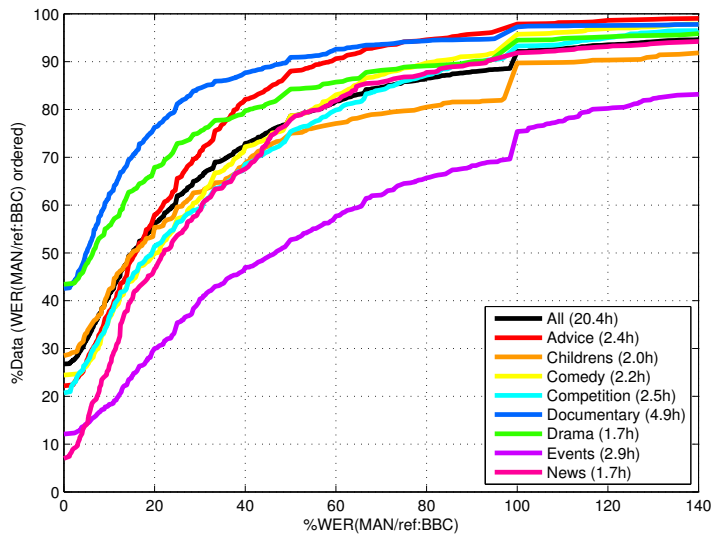
Natural Speech Technology

- We consider 8 primary genres extracted from json(infax)
  - advice, childrens, comedy, competition, documentary, drama, events, news
- in the previous experiment, we assumed that MER was representative of the quality of the transcript independently of the genre but is that really the case?
  - MER depends on the quality of the transcription but also on the performance of the AM
  - both can vary across genre
- we ran an data selection analysis per genre on the carefully transcribed dev data

Data Selection analysis per genre

- taking the BBC transcripts as reference, we plot the cumulative duration (%) given MER(ASRvsBBC) and WER(MANvsBBC)
  - left graph:
    - by comparison with the MER graph for the train dataset, we show that the devset is representative of the whole dataset in term of MER
  - right graph:
    - quality of the transcript depends on the genre (assuming MAN indep.)
    - MER & WER graphs have roughly the same shape: MER seems to be a good indicator of the quality of the transcript

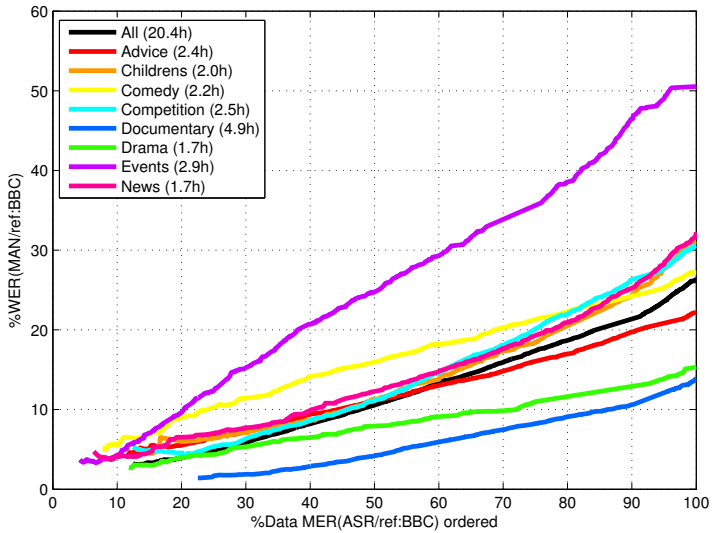


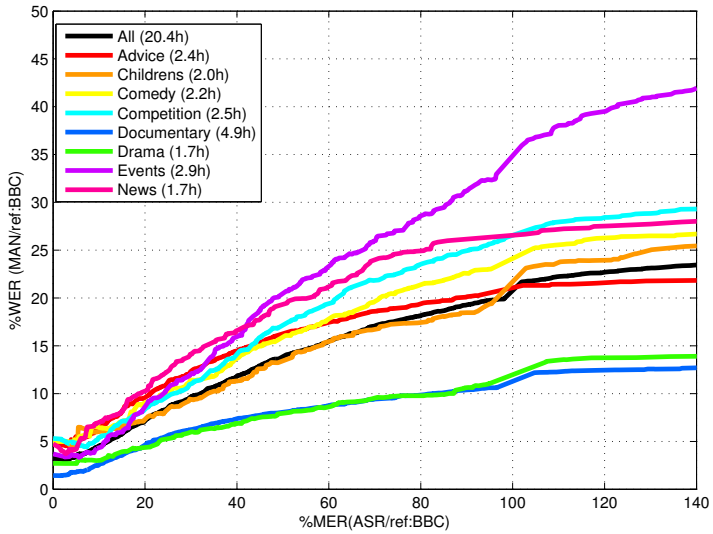


Data Selection analysis per genre

Natural Speech Technology

- Which transcript quality should we expect by doing a selection according to the %data (left graph) or to the %MER(right graph)?
- Selection threshold should depend on genre given that the dev set is representative of the whole dataset in term of transcription quality per genre (future experiments)





Sync Demo

Natural Speech Technology

- Original subtitles might be imperfect
  - imprecise time stamps position, episode-specific lag
  - approximate transcription (re-speaking)
- We present the result of our alignment method based on a lightly supervised (LS) approach on 3 shows
  - `v20080609_223000_bbcfour_the_alan_clark_diaries` (drama) [-20:16]
  - `v20080628_191000_bbcone_doctor_who` (drama) [-25:10]
  - `v20080618_224500_bbcone_uefa_euro_2008_match` (sport event) [-35:38]

- in green: original subtitle
- in yellow: re-aligned subtitle (LS approach)
- in white: output of the LS decoding