Selection of Multi-Genre Broadcast Data for the Training of Automatic Speech Recognition Systems

P. Lanchantin, M.J.F. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P.C. Woodland, C. Zhang

Cambridge University Engineering Department, Cambridge CB2 1PZ, U.K.
{pk127,mjfg,pk407,xl207,yq236,lw519,pcw,cz277}@eng.cam.ac.uk

Abstract

This paper compares schemes for the selection of multi-genre broadcast data and corresponding transcriptions for speech recognition model training. Selections of the same amount of data (700 hours) from lightly supervised alignments based on the same original subtitle transcripts are compared. Data segments were selected according to a maximum phone matched error rate between the lightly supervised decoding and the original transcript. The data selected with an improved lightly supervised system yields lower word error rates (WERs). Detailed comparisons of the data selected on carefully transcribed development data show how the selected portions match the true phone error rate for each genre. From a broader perspective, it is shown that for different genres, either the original subtitles or the lightly supervised output should be used for model training and a suitable combination yields further reductions in final WER.

1. Introduction

Recently there has been substantial interest in automatic transcription of general broadcast data and audio from web-based multimedia sources. This enables applications including content-based search but requires training suitable acoustic models. General broadcast data is recorded in diverse environments, includes dramas with highly-emotional speech, and often has overlaid background music or sound effects; word error rates (WERs) on such data are several times higher than for broadcast news and very variable across different genres. Work in this area has included automatic transcription of podcasts and other web audio [1], automatic transcription of Youtube [2, 3], the MediaEval speech retrieval evaluation which used blip.tv semi-professional user created content [4], the automatic tagging of a large radio archive [5], and automatic transcription of multi-genre media archive data [6]. Recently, systems were developed for the 2015 Multi-Genre Broadcast (MGB) challenge [7–10].

The MGB challenge [7] was an evaluation of speech recognition, alignment and speaker diarisation using audio from television programmes supplied by the British Broadcasting Corporation (BBC). It used audio from 7 weeks of programmes for acoustic model training data which were supplied with corresponding subtitles (closed captions). A key difficulty in training acoustic models on broadcast data is due to the variability in the quality of the available subtitles in terms of both segment-level timings and the transcription accuracy. In the case of closed captions, they may be approximate for various reasons, including the production process. A re-alignment of the captions is necessary to correct time stamps. Furthermore, confidence scores can be computed to find regions where the transcriptions are likely to be accurate in order to select data for subsequent training.

A lightly supervised approach [10–20] was used for the preparation of the data provided to MGB participants in which the output from a speech recogniser, using a language model biased towards the original transcripts, was compared to the original transcripts and a phone matched error rate (PMER) computed between the two for each recognised segment. The maximum PMER, along with an average word duration (AWD) threshold [10], allows segments to be selected for training while ensuring that the word/phone supervision information is reasonably accurate. We first re-processed the entire set of audio using lightly supervised decoding with an improved procedure that included acoustic models trained on 700 hours of audio taken from the information supplied for the MGB challenge. This led to revised alignments and segmentations of the BBC captions along with new confidence values on which improved acoustic models were trained on a revised 700 hour selection. The same procedure was repeated twice leading to a third 700 hour selection.

In this paper we compare these 700 hour selections to evaluate how they differ in terms of genre balance, transcript quality and WER of the trained acoustic models. We only used the aligned subtitle transcripts for acoustic training during our participation in the MGB challenge [8]. Hence, from a broader perspective, we also investigate if a genre-dependent combination of provided transcripts and the outputs of the revised lightly supervised hypotheses can lead to additional reductions in WER for the final transcription system.

2. Data Selection

2.1. Data used

The MGB challenge made available a total of 1600h of raw audio taken from 7 weeks of BBC programmes for acoustic model training. A 28 hour development test set (47 different programme episodes) was also provided. The data covers a wide variety of broadcast audio covering a full range of genres (e.g. documentaries, news, comedy, drama, sport events, etc). For language model training, a large corpus of additional text data of BBC closed captions was also provided for the MGB challenge yielding a total of 650 million words for language model training; 10M words of data from the 7 week acoustic transcripts; 640 million words from the additional subtitle data. A baseline 4-gram, denoted LM2 mono, [8], was trained using all 650 million words of text data with a 160k vocabulary. A selec-
tion of 700 hours of data, yielding the training set 700h-v1, used the MGB-provided alignments/segmentations (v1) of the BBC captions according to a maximum PMER [10]. Then sequence-trained hybrid acoustic models were trained on this selection and used to re-process the entire 1600 hours of audio with an improved acoustic segmentation and strong episode-based biased language models as described in [10]. This led to revised alignments/segmentations (v2) of the BBC captions along with new MER/PMER values. A second 700th selection of data according to maximum PMER then yielded a second training set: 700h-v2 on which new improved acoustic models were trained and used with an improved acoustic segmentation to led to revised alignments/segmentations (v3) of the BBC captions. A third 700th selection of data according to maximum PMER finally yielded a third training set: 700h-v3. In all cases only BBC subtitle word sequences were used for training.

2.2. Data selection analysis

The plots in Fig. 1 show the cumulative quantity of data according to a maximum PMER value\(^1\) as mentioned in the introduction. An AWD threshold\(^2\) was applied. The v2 and v3 alignment / lightly supervised output greatly increases the quantity of data having a zero PMER from 140h to 209h and 243h respectively. This in turn enabled the training of better DNN-based segmenters as described in [8]. To achieve the chosen operating point of 700h of selected training data, the maximum PMER decreases from 40% for v1 to 30% for v2 and 25.75% for v3 as indicated in Fig. 1. Note that PMER depends on several factors: the quality of the original transcripts, the acoustic models used for the lightly supervised decoding, and the alignment of the original transcripts. Hence a high PMER does not necessarily mean that the transcripts are incorrect given that the difference could be due, for instance, to the poor performance of the speech recogniser system in noisy acoustic conditions. The decrease in the maximum PMER for a 700h selection is due to better acoustic model performance, to the improved lightly supervised alignment procedure of the original transcripts (including an improved segmenter and strong episode-based biased language model) [10] or to a combination of both. Looking at Table 1 the new alignments significantly change the distribution of data across genres, reducing news and events data but increasing all others. They also increase the proportion of the harder genres such as drama and for those genres, the refined systems were better at identifying good transcript regions.

Table 1: genre repartition for the 700h selections.

<table>
<thead>
<tr>
<th></th>
<th>advice</th>
<th>childr.</th>
<th>comedy</th>
<th>compet.</th>
<th>docum.</th>
<th>drama</th>
<th>events</th>
<th>news</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>15.5</td>
<td>9.3</td>
<td>3.7</td>
<td>13.9</td>
<td>16.3</td>
<td>5.6</td>
<td>7.3</td>
<td>28.3</td>
</tr>
<tr>
<td>v2</td>
<td>15.7</td>
<td>10.1</td>
<td>4.2</td>
<td>14.4</td>
<td>17.3</td>
<td>6.9</td>
<td>6.6</td>
<td>24.0</td>
</tr>
<tr>
<td>v3</td>
<td>15.7</td>
<td>10.7</td>
<td>4.4</td>
<td>14.4</td>
<td>17.3</td>
<td>6.9</td>
<td>6.6</td>
<td>24.0</td>
</tr>
</tbody>
</table>

Table 2: Differences between the 700h selections in %frames.

The cumulative duration plots don’t show the difference in content between the training sets. Table 2 illustrates the difference in terms of the number of frames. Globally, 24.6% of

---

\(^1\)PMER/WMER is computed in the same way than the traditional segment-level phone and word error rate but with the original transcript as reference, which is not necessarily accurate.

\(^2\)AWD is computed by dividing the sentence duration in seconds by the number of words in the sentence.

Figure 1: Cumulative duration of the selected training data according to a maximum threshold on PMER for both the v1 and v2 and v3 refined alignments/segmentations for 0.165%≤AWD≤0.66%. v1 refers to the alignments provided to MGB participants and v2 and v3 refer to the refined alignments.

The frames are different between the v1 and v2 selections and 15.6% between v2 and v3. Differences range from 16.7% for documentary to 41.5% for events between v1 and v2, and from 9.6% to 28.3% between v2 and v3 for the same genres. Audio content is then significantly different between the training sets especially for harder genres (comedy, drama and events).

The quality of the aligned BBC transcripts of both 700h training sets can be estimated by using the same lightly supervised procedures on the carefully transcribed development set and computing a phone error rate (PER). In the top plot of Fig. 2, the dev set PER is computed for selections with a maximum PMER value. For a maximum PMER varying from 0% to 50%, the global PER of the aligned BBC transcripts varies from 2.0% to 11.0% for v1, from 3.3% to 11.7% for v2 whereas it varies from 3.6% to 12.0% for v3. Considering the maximum PMER values in the 700h training sets, the quality of the transcripts increase for the v2 and v3 procedure since there is a 1.6% absolute reduction in maximum PER between v1 and v3 on the development set: PER=8.1% at PMER≤25.75% for v3, PER=8.7% at PMER≤30% for v2 whereas PER=9.7% at PMER≤40% for the v1 procedure. The PER depends on the quality of the transcripts and of their alignment to the audio. Hence for v2 and v3 it can be seen that improved acoustic models, segmentation and language models led to better alignments resulting in better quality aligned transcripts.

700h training sets were finally compared in terms of the WER of the trained acoustic models. For each training set, speaker independent (SI) hidden Markov models with Gaussian mixture model state output distributions (GMM-HMMs) were estimated using minimum phone error (MPE) training. These were used for recognition and for state-level alignment for training a hybrid SI system with a deep neural network (DNN) acoustic model in a DNN-HMM framework. The DNN architecture used is 720 × 1000\(^3\) × 9500 with rectified linear unit (ReLU) hidden activations and the cross entropy (CE) training criterion. All acoustic model training used a pre-release version of HTK 3.5 [21, 22]. Table 5 shows the WERs of the trained systems. A reduction of 0.4% and 0.7% absolute is obtained for the GMM-HMM and DNN-HMM systems respectively when considering the 700h-v2 selection instead of 700h-v1. A smaller reduction of 0.1% absolute is obtained for the DNN-HMM sys-
transcribed using real-time captioning: 37.4% of the 1600h data
be due to the speech recognition software.
formulate what they hear to enhance clarity and some errors can
silencer to repeat what they hear into speech recognition soft-
live sports: respeakers (or voice writers) uses a mask or speech
[23] is increasingly common since 2001 for news broadcast and
due to the subtitle production process. Real-time captioning
Only the original BBC transcripts were used for acoustic train-
a convergence of the proposed approach.
Table 3: %audio transcribed using real-time captioning.
has been transcribed “live” and most of the news (82.2%) and events (78.5%) programme episodes have been transcribed this way. On the other hand, only a very small portion of the doc-
umentaries and drama have been transcribed live. Note that it
doesn’t necessarily mean that these “offline” transcripts are per-
fect as they might be edited to enhance clarity, paraphrasing,
and generally don’t include hesitations or disfluencies. Given
these considerations, the transcripts generated by the lightly su-
pervised decoding (denoted ASR) might be a better alternative
as training material for some genres. In the following we ex-
plor the possibility of using or combining both types of tran-
scripts in order to improve the WER of the trained models.

3. Transcript selection

Only the original BBC transcripts were used for acoustic train-
ing. However, the quality of these transcripts can vary greatly
due to the subtitle production process. Real-time captioning
[23] is increasingly common since 2001 for news broadcast and
live sports: respeakers (or voice writers) uses a mask or speech
silencer to repeat what they hear into speech recognition soft-
ware to generate the corresponding text. Thus, they might re-
formulate what they hear to enhance clarity and some errors can
be due to the speech recognition software.

Table 3 gives the percentage of MGB data which has been
transcribed using real-time captioning: 37.4% of the 1600h data

![Graph 1](image1.png)

**Figure 2:** top: PER of the original (BBC) and lightly supervised (ASR) transcripts selection for a maximum threshold on PMER on the dev set. bottom: genre breakdown for BBC transcripts considering v3.

![Graph 2](image2.png)

**Figure 3:** Cumulative %PER vs %PMER plot for live and offline transcribed shows on the development set considering BBC transcript (---) and ASR (---) lightly supervised output considering the v2 alignments/segmentations.

Figure 3 shows the PER for selections based on both the original (BBC) and lightly supervised transcripts (ASR) for live and offline transcribed episodes with a maximum PMER value on the development set with the v2 alignments/segmentations. The PER of BBC transcripts for live programmes is 4.1% absolute higher than for offline programme at PMER ≤ 30. This large difference doesn’t occur for ASR transcripts, for which PER is comparable for both live and offline transcribed episodes, with a maximum PER difference of 0.6% absolute in favour of live transcribed episodes. However, for offline transcribed shows, the BBC transcripts are slightly better than ASR ones. Hence, this suggests that ASR transcripts should be used for live subti-
tled episodes and the BBC transcripts for the offline ones.

The above development set comparison can also include a
genre breakdown. In the top plot in Fig. 2 for v1, the qual-
ity of BBC transcripts is better than ASR. For the procedure
used for the 700h-v1 training set for which PMER ≤ 40%, the
PER is 12.7% for ASR compared to 9.7% for BBC transcripts.
However, the order changes for the 700h-v2 training set, since
at PMER ≤ 30% the ASR transcripts are better than the BBC
ones with PER=8.3% for the ASR transcript compared to 8.7%
for the BBC transcripts, the same effect being observed for
700h-v3. A genre breakdown shown in bottom of Fig. 2 con-
firms that the quality of transcripts is genre-dependent. Ta-

![Graph 3](image3.png)
Table 4: PER difference between the BBC and ASR transcripts for the 700h-v1 selection (PMER ≤ 40), 700-v2 selection (PMER ≤ 30) and 700-v3 selection (PMER ≤ 25.75). Colour indicates the transcript minimising the PER.

For the 700h-v2 training set segmentation, the transcripts were modified using the information in Table 4, using ASR tran-
scripts for all genres. However no extra reduction of 0.3% and 0.9% absolute is obtained for the GMM-
HMM and DNN-HMM based systems respectively when considering the ASR transcripts for all genres.

3.2. Combined transcription

For the 700h-v2 training set segmentation, the transcripts were modified using the information in Table 4, using ASR transcripts for all genres except comedy, documentary and drama for which ASR transcripts were retained to yield a new training set 700h-v2mix. GMM-HMM and DNN-HMM based systems were trained in the same way as described in section 2 and results are presented in Table 5: a reduction of 0.7% absolute in the DNN-based hybrid system WER is obtained when using the transcript combination instead of the BBC transcripts. Thus, despite the results on the dev set, it appears that using the ASR transcripts for all genres on the training set lead to a bigger reduction in WER, confirming the trend observed in Table 4.

<table>
<thead>
<tr>
<th>AM</th>
<th>Sel. Transcripts</th>
<th>fg_en</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-HMM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v1 BBC</td>
<td>40.7</td>
<td></td>
</tr>
<tr>
<td>v2 BBC</td>
<td>40.5</td>
<td></td>
</tr>
<tr>
<td>ASR</td>
<td>40.0</td>
<td></td>
</tr>
<tr>
<td>v3 BBC</td>
<td>40.3</td>
<td></td>
</tr>
<tr>
<td>ASR</td>
<td>40.0</td>
<td></td>
</tr>
<tr>
<td>DNN-HMM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v1 BBC</td>
<td>27.5</td>
<td></td>
</tr>
<tr>
<td>v2 BBC</td>
<td>26.8</td>
<td></td>
</tr>
<tr>
<td>ASR</td>
<td>25.9</td>
<td></td>
</tr>
<tr>
<td>v3 BBC</td>
<td>26.7</td>
<td></td>
</tr>
<tr>
<td>ASR</td>
<td>25.9</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: %WER of SLPME GMM-HMMs and CE DNN-HMM systems on dev.full using manual segmentation, confusion network decoding and 4-gram LM (fg_en). Original (BBC) and its genre-dependent combination with lightly supervised transcripts (ASR) included for 700h-v2.
5. References


