INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

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.
{pkl27,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 $\text{LM2}_{\text{prune}}$ [8], was trained using all 650 million words of text data with a 160k vocabulary. A selec-

This work is in part supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). Chao Zhang is also supported by Cambridge International Scholarship from the Cambridge Commonwealth, European & International Trust. Supporting data for this paper will be available at the http://www.repository.cam.ac.uk data repository.

http://dx.doi.org/10.21437/Interspeech.2016-462

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 700h 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 700h 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.

	advice	childr.	comedy	compet.	docum.	drama	events	news
v1	15.5	9.3	3.7	13.9	16.3	5.6	7.3	28.3
v2	15.7	10.1	4.2	14.2	16.7	6.6	7.0	25.4
v3	15.7	10.7	4.4	14.4	17.3	6.9	6.6	24.0

Table 1: genre repartition for the 700h selections.

	advice	childr.	comedy	compet.	docum.	drama	events	news
v2%v1	21.3	26.6	32.6	22.0	16.7	29.5	41.5	24.2
v3%v2	13.3	16.0	18.5	13.8	9.7	14.4	28.3	17.0

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 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).

[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.

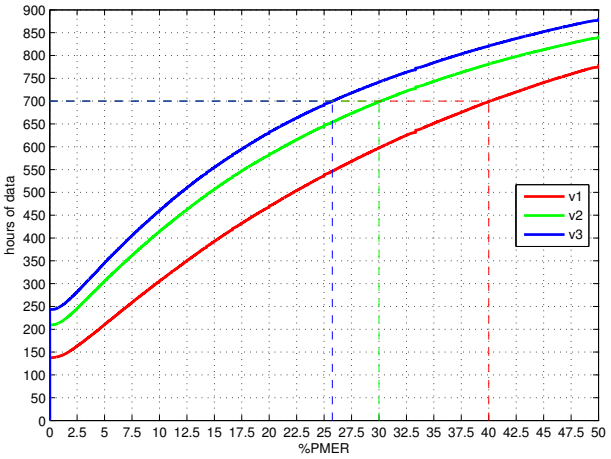

Figure 1: Cumulative duration of the selected training data according to a maximum threshold on PMER for both the v1 and v2 and v3 refined alignments/segmentations for 0.165s≤AWD≤0.66s. v1 refers to the alignments provided to MGB participants and v2 and v3 refer to the refined alignments.

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 $720 \times 1000^5 \times 9500$ 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 small reduction of 0.1% absolute is obtained for the DNN-HMM sys-

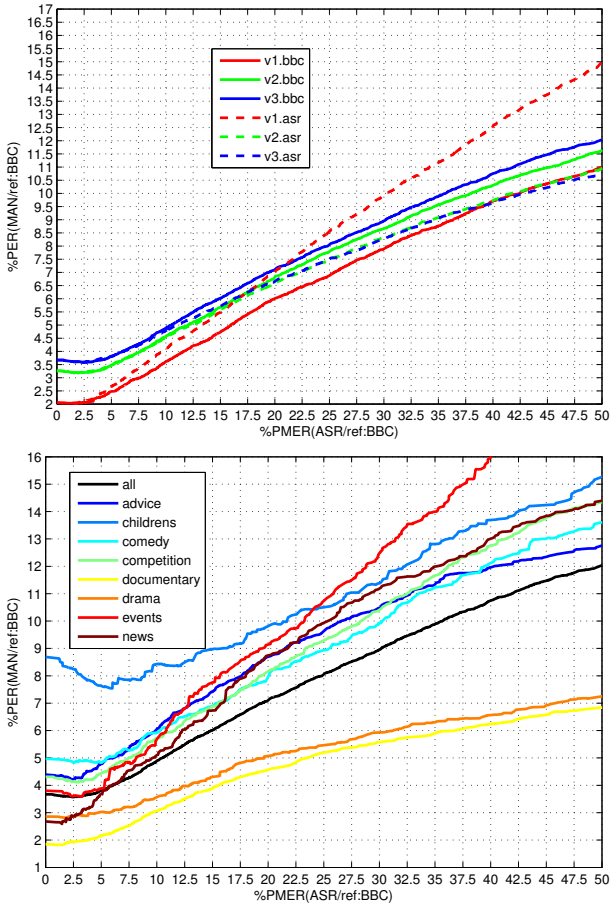

Figure 2: **top**: PER of the original (BBC) and lightly supervised (ASR) transcripts selection for a maximum threshold on PMER on the dev set. **bottom**: genre breakdown for BBC transcripts considering v3.

tem when considering the 700h-v3 selection instead of 700h-v2.

Re-processing all the available audio with improved acoustic segmentations, acoustic models and strong episode-based biased language models led to a better alignment of the transcripts. It also modified the distribution of data across genres as well as the selected content, especially for harder genres such as comedy, drama and events. This led to an improvement in the WER performance of the trained acoustic models, at the computational cost of re-processing 1600h of audio. However, the last iteration of the whole process (leading to v3) did not lead to significant extra improvement in our experiments indicating a convergence of the proposed approach.

3. Transcript selection

Only the original BBC transcripts were used for acoustic training. 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 software to generate the corresponding text. Thus, they might reformulate 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

genre	advice	childr.	comedy	compet.	docum.	drama	events	news
%live	14.7	5.5	6.3	7.2	0.5	0.8	78.5	82.2

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 documentaries and drama have been transcribed live. Note that it doesn't necessarily mean that these "offline" transcripts are perfect 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 supervised decoding (denoted ASR) might be a better alternative as training material for some genres. In the following we explore the possibility of using or combining both types of transcripts in order to improve the WER of the trained models.

3.1. Comparison between BBC and ASR transcripts

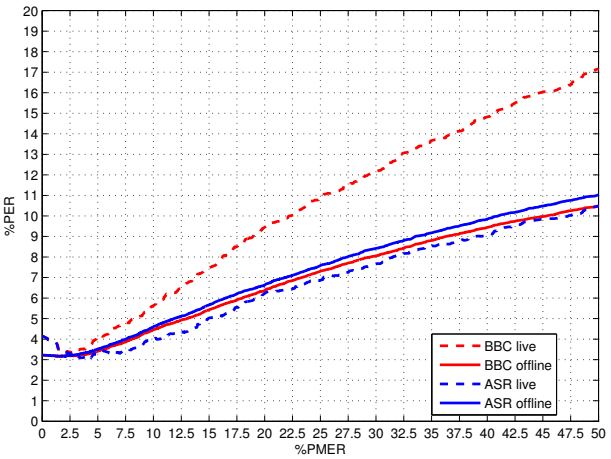

Figure 3: Cumulative %PER vs %PMER plot for live and offline transcribed shows on the development set considering BBC transcript (- - -) and ASR (—) lightly supervised output considering the v2 alignments/segmentations.

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$\leq$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 subtitled 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 quality of BBC transcripts is better than ASR. For the procedure used for the 700h-v1 training set for which PMER$\leq$40%, the PER is 12.7% for ASR compared to 9.7% for BBC transcripts. However, the order changes for the 700h-v2 training set, since at PMER$\leq$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 confirms that the quality of transcripts is genre-dependent. Ta-

	advice	childr.	comedy	compet.	docum.	drama	events	news
v1	-1.7	-2.4	-8.5	-2.7	-3.8	-10.9	-1.0	5.2
v2	1.5	0.4	-2.6	0.4	-0.5	-2.5	1.4	4.9
v3	1.7	0.6	-1.2	0.6	-0.3	-1.9	1.8	4.7

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.

ble 4 lists the dev set PER differences between the BBC and ASR transcripts for the 3 training sets. For the v1 alignments/segmentations, the quality of ASR transcripts is poorer than the BBC for all genres except for news. However, for the v2 alignments/segmentations, the quality of the ASR transcripts is better than BBC for most genres, e.g. events and the same tendency is observed for v3. as shown in Fig. 4. The exceptions are documentary, comedy and drama (see Fig. 4), for which the audio appears to be more difficult for automatic transcription as it might include highly expressive speaking styles for which recogniser performance yields higher WERs. Thus, while it was preferable to use the BBC transcripts for the MGB-provided v1 alignment for all genres, for the v2 and v3 alignments, it appears better to use a genre dependent combination of the ASR and BBC transcripts or solely the ASR transcripts for all genres as the performance of the acoustic models improve.

Table 5 shows the WERs of the systems trained on the same segmentation, but using the ASR transcripts. For v2, A reduction of 0.3% and 0.9% absolute is obtained for the GMM-HMM and DNN-HMM systems respectively when considering the ASR transcripts for all genres. However no extra reduction is obtained when considering v3.

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 BBC 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.

AM	Sel.	Transcripts	fg_cn
GMM-HMM	v1	BBC	40.7
	v2	BBC	40.3
		ASR	40.0
		BBC+ASR	40.1
	v3	BBC	40.3
		ASR	40.0
DNN-HMM	v1	BBC	27.5
	v2	BBC	26.8
		ASR	25.9
		BBC+ASR	26.1
	v3	BBC	26.7
		ASR	25.9

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

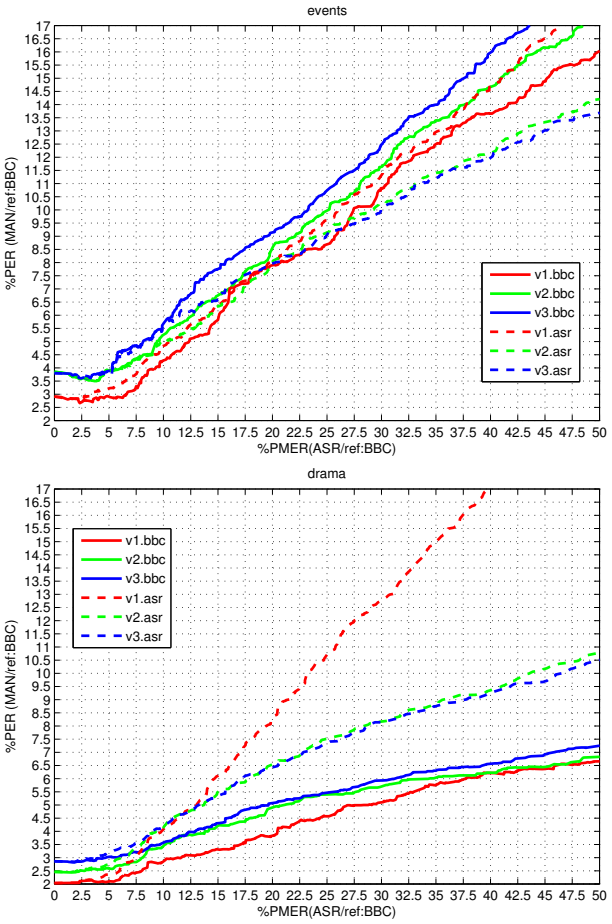

Figure 4: PER of the original (BBC) and lightly supervised (ASR) transcripts selection defined by a maximum threshold on PMER considering the v1, v2 and v3 alignments for drama and events. The manual transcription of the development set was used as reference.

4. Conclusion

This paper compared selecting 700h training sets from the multi-genre broadcast data based on lightly supervised training in terms of genre distribution, transcript quality and WER of the subsequently trained acoustic models. It was shows that reprocessing of the entire set of MGB audio with improved acoustic segmentation, acoustic models and strong episode-based biased language models led to a better alignment of the transcripts and modified the distribution of data across genres as well as the selected content. The major change was an increase in data for harder genres such as comedy, drama and events. This led to improvements in the WER given by both GMM-HMM and DNN-HMM acoustic models. It also allows the lightly supervised ASR transcripts to be used for most of the genres, and by combining the original subtitle-based transcripts and ASR transcripts in a genre-dependent fashion gave a new training set which yields a further reduction in WER. After few iterations of the procedure, the ASR transcripts can be used for all genres. In future, we plan to explore the combination of several confidence measures in addition to PMER for the selection of training data and the corresponding transcripts.

5. References

[1] J. Ogata & M. Goto, "Podcastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription", *Proc. Interspeech*, Brighton, 2009.

[2] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina & O. Siohan, "An audio indexing system for election video material", *Proc ICASSP*, Taipei, 2009.

[3] R. C. van Dalen, J. Yang & M.J.F. Gales, "Generative kernels and score-spaces for classification of speech: Progress report", *Tech. Rep. , Cambridge University Engineering Department*, 2012.

[4] M. Larson, M. Eskevitch, R. Orderlman, C. Kofler, S. Schmiedeke & G.J.F. Jones, "Overview of Mediaeval 2011 Rich speech retrieval task and genre tagging task", *Working Notes Proceedings of the MediaEval 2011 Workshop*, 2011.

[5] Y. Raimond, C. Lowis, R. Hodgson & J. Tweed, "Automatic semantic tagging of speech audio", *Proc. WWW 2012*, 2012.

[6] 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 & P.C. Woodland, "Automatic transcription of multi-genre media archives", *Proc. first Workshop on Speech, Language and Audio in Multimedia*, 2013.

[7] P. Bell, M.J.F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester & P.C. Woodland, "The MGB Challenge: evaluating multi-genre broadcast media transcription", To appear *Proc IEEE ASRU*, Scottsdale, 2015.

[8] P.C. Woodland, X. Liu, Y. Qian, C. Zhang, M.J.F. Gales, P. Karanasou, P. Lanchantin & L. Wang, "Cambridge University transcription systems for the Multi-Genre Broadcast Challenge", To appear *Proc IEEE ASRU*, Scottsdale, 2015.

[9] P. Karanasou, M.J.F. Gales, P. Lanchantin, X. Liu, Y. Qian, L. Wang, P.C. Woodland & C. Zhang, "Speaker diarisation and longitudinal linking on multi-genre broadcast data", To appear *Proc IEEE ASRU*, Scottsdale, 2015.

[10] P. Lanchantin, M.J.F. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P.C. Woodland & C. Zhang, "The Development of the Cambridge University Alignment Systems for the Multi-Genre Broadcast Challenge", To appear *Proc IEEE ASRU*, Scottsdale, 2015.

[11] J. Robert-Ribes & R. Mukhtar, "Automatic generation of hyperlinks between audio and transcript", *Proc. Eurospeech*, Rhodes,1997.

[12] P.J. Moreno, C. Joerg, J.-M.V. Thong & O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments", *Proc. ICSLP*, Sydney, 1998.

[13] L. Lamel, J.L. Gauvain & G. Adda, "Lightly supervised and unsupervised acoustic model training", *Computer Speech and Language*, vol. 16, pp. 115–129, 2002,

[14] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, VRR. Gadde & J. Zheng, "An efficient repair procedure for quick transcriptions", *Proc. ICSLP*, Jeju Island, 2004.

[15] H.Y. Chan & P.C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training", *Proc. ICASSP*, Montreal, 2004.

[16] L. Mathias, G. Yegnanarayanan & J. Fritsch, "Discriminative training of acoustic models applied to domains with unreliable transcripts", *Proc. ICASSP*, Philadelphia, 2005.

[17] B. Lecouteux, G. Linares, P. Nocera & J.F. Bonastre, "Imperfect transcript driven speech recognition", *Proc. Interspeech*, Pittsburgh, 2006.

[18] A. Haubold & J. Kender, "Alignment of speech to highly imperfect text transcriptions", *Proc. IEEE International Conference on Multimedia and Expo*, 2007.

[19] N. Braunschweiler, M.J.F Gales & S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," *Proc. Interspeech*, Makuhari, 2010.

[20] S. Hoffman & B. Pfister, "Text-to-speech alignment of long recordings using universal phone models," *Proc Interspeech*, Lyon, 2013.

[21] S. Young, G. Evermann, M. Gales, T. Hain., D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book (for HTK version 3.4)*, Cambridge University Engineering Department, 2006.

[22] C. Zhang & P.C. Woodland, "A general artificial neural network extension for HTK", *Proc. Interspeech*, Dresden, 2015.

[23] C. Eugeni, "Respeaking the BBC news: a strategic analysis of respeaking on the BBC," *The Sign Language Translator and Interpreter*, vol. 3, no. 1, pp. 29–68, 2009.