Automatic Transcription of Multi-genre Media Archives

P. Lanchantin[1], P.J. Bell[2], M.J.F. Gales[1], T. Hain[3], X. Liu[1], Y. Long[1], J. Quinnell[1]
S. Renals[2], O. Saz[3], M. S. Seigel[1], P. Swietojanski[2], P. C. Woodland[1]

[1]Cambridge University Engineering Department, Cambridge CB2 1PZ, UK
{pkl27,mjfg,xl207,yl467,jq228,mss46,pcw}@eng.cam.ac.uk
[2]Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK
{peter.bell,s.renals}@ed.ac.uk,p.swietojanski@sms.ed.ac.uk
[3]Speech and Hearing Research Group, University of Sheffield, Sheffield S1 4DP, UK
{t.hain,o.saztorralba}@dcs.shef.ac.uk

Abstract

This paper describes some recent results of our collaborative work on developing a speech recognition system for the automatic transcription or media archives from the British Broadcasting Corporation (BBC). The material includes a wide diversity of shows with their associated metadata. The latter are highly diverse in terms of completeness, reliability and accuracy. First, we investigate how to improve lightly supervised acoustic training, when timestamp information is inaccurate and when speech deviates significantly from the transcription, and how to perform evaluations when no reference transcripts are available. An automatic timestamp correction method as well as a word and segment level combination approaches between the lightly supervised transcripts and the original programme scripts are presented which yield improved metadata. Experimental results show that systems trained using the improved metadata consistently outperform those trained with only the original lightly supervised decoding hypotheses. Secondly, we show that the recognition task may benefit from systems trained on a combination of in-domain and out-of-domain data. Working with tandem HMMs, we describe Multi-level Adaptive Networks, a novel technique for incorporating information from out-of domain posterior features using deep neural network. We show that it provides a substantial reduction in WER over other systems including a PLP-based baseline, in-domain tandem features, and the best out-of-domain tandem features.

Index Terms: lightly supervised training, cross-domain adaptation, tandem, speech recognition, confidence scores, media archives

1. Introduction

The British Broadcasting Corporation (BBC) has a stated aim to open its broadcast archive to the public by 2022. Automatic transcription, metadata extraction and indexing of such material would give access to a large amount of content, indexing historic content, and enabling search based on transcriptions, speaker identity and other extracted metadata. However, technologies for this particular task are still underdeveloped. In the scope of the Natural Speech Technology EPSRC project and in collaboration with BBC Research and Development, we have begun to investigate the automatic transcription of broadcast material across different genres, using sparse or non-existent associated metadata and text resources.

This research was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). Thanks to Andrew McParland, Yves Raimond and Sam Davies of BBC R&D

Automatic transcription of arbitrary, multi-genre media content is a challenging task since the material to recognise may include broadcasts in diverse environments and drama with highly-emotional speech, overlaid background music or sound effects. Recent work on this task has for instance included automatic transcription of podcasts and other web audio [1] automatic transcription of Youtube [2, 3], the MediaEval rich speech retrieval evaluation which used blip.tv semi-professional user created content [4], and the automatic tagging of a large radio archive [5]. On the other hand, in order to train models for such large vocabulary continuous speech recognition systems, text resources and other metadata are highly desirable to provide in-domain training data. The problem is that the nature of these metadata may vary considerably over archive material in terms of completeness, reliability and precision. This partly reflects the large epoch (decades) that the data covers. A range of techniques have been proposed for this purpose such as the lightly supervised training approach [6], based on a biased language model (LM) decoding, and several methods have since been proposed along this line to improve upon this approach [7, 8, 9, 10].

In recent work described in [11, 12] which will be reviewed in this paper, we focused on two aspects related to the building of systems for automatic transcription of multi-genre media archives: lightly supervised training and evaluation using out-of-domain data. We recently proposed in [12] an approach in which phone level mismatch information is used to identify reliable regions where segment-level transcription combination can be used. Schemes for combining the imperfect original transcriptions with the confusion networks (CN) generated during the biased LM decoding can then be applied to leverage differences in the characteristics of the two forms of transcriptions. An evaluation technique based on ranking systems using imperfect reference transcripts was used to evaluate system performance. Secondly, in [11], we focused on the development of methods which can effectively combine in-domain and out-of-domain training data, using neural networks in the tandem framework [13] whereby context-dependent hidden Markov models (HMMs) with Gaussian mixture model (GMM) output distributions are trained on standard acoustic features concatenated with features derived from neural networks. A novel technique for posterior feature combination in a cross-domain setting and referred to as Multi-Level Adaptive Networks (MLAN) was then proposed. This technique has been investigated using a multi-genre broadcast corpus built from the data provided by the BBC, in terms of cross-domain speech recognition using different acoustic training data sources across different target genres.

The new technique was evaluated in terms of a discriminatively-trained speaker-adaptive speech recognition system, comparing in-domain and out-of-domain posterior features with the features obtained using MLAN.

The rest of the paper is organised as follows. In Section 2 the available BBC datasets are presented. Section 3 presents lightly supervised approaches for the correction of timestamp positions and the proposed transcription combination schemes. Finally, Section 4 presents the multi-level adaptive network scheme for the transcription of multi-genre data followed by conclusions in Section 5.

2. Description of the BBC datasets

The stated aim of the BBC to open its broadcast archive to the public by 2022 will give access to a very large amount of data: potentially 400,000 television programmes, over 700,000 hours of video and 300,000 hours of audio. A large amount of metadata associated to these data will be available from the *Infax* cataloguing system which allows to access tags manually attributed to programmes in varying levels of detail (more than 600,000 items) some of which are already publicly available. In the scope of our collaboration with BBC research and development started in 2011, six different sets of shows with their associated metadata have been provided for the investigation and the development of methods and systems for automatic transcription of broadcast material across the full range of genres.

2.1. Diverse shows/genres

The six sets contain speech that is mostly British English with a range of regional accents and audio contents covering a broad range of genres, environments and speaking styles that we describe below.

Radio4-1day: contains 36 talk-radio programmes broadcast on the same radio channel (BBC Radio 4) over 24 hours in February 2009. The duration of programmes range from 2 minutes for weather report to 3 hours for morning news/current affair programmes to give a total duration of 18 hours. The audio material covers different genres: news, weather reports, book readings, documentaries, panel games and debates.

Archives: contains 136 radio and TV programmes some of which are publicly available on the BBC archives website (http://www.bbc.co.uk/archive). It includes 399 episodes representing 271 hours of raw audio data with 146 hours of active speech. Episodes were recorded from 1970 to 2003. As for the Radio4-1day dataset, audio material covers a broad range of genres, environments and speaking styles.

Desert Island Discs: is a radio programme broadcast on BBC Radio 4. Each week, a guest is asked to choose eight pieces of music, a book and a luxury item that they would take if they were to be castaway on a desert island, whilst discussing their lives. It includes only two speakers in each show, the presenter and the guest, and small portions of music. This set includes 180 episodes representing 108 hours of raw data with 88 hours of active speech.

Reith Lectures: are a series of annual radio lectures on significant contemporary issues, delivered by leading figures from their relevant fields. The set includes 155 episodes, covering the years from 1976 to 2010. Each lecturer had 3-6 episodes presented at different times. Each episode is composed of several regions: the lecture region given by the lecturer, a non-lecture region which contains the introduction to the lecture by a presenter and since 1988, a question and answer session after the main lecture. The duration of each episode ranges from 18-35 minutes, to give a total audio duration of 72 hours from which 71.3 hours of lecture region data were extracted.

TV-drama: includes 14 episodes of a science fiction TV-drama series broadcast in 2010. Episode durations range from 40-75 minutes, to give a total duration of 11 hours.

TV-1week: includes 169 unique shows and 333 episodes broadcast on 4 BBC TV channels during the week of May 5th, 2008 through May 11th, 2008 representing 236 hours of raw audio data. The duration of the programmes ranges from 3 minutes to 4 hours. A list of genres covered by the programmes was provided with up to 85 different categories, although programmes typically get assigned to more than one genre. This categorisation includes drama series, soap operas, different types of documentaries, live sports, broadcast news, quiz shows or animation programmes.

The available audio material contained in these sets covers different genres and a broad range of environment and speaking style. For purposes of analysis, we divided the data into three categories by broad genre:

`studio`: in which speech is controlled, recorded in studio conditions or news reports, sometimes including telephone speech from reporters or contributors;

`location`: which includes material produced on "location" including for instance parliamentary proceedings;

`drama`: TV drama series, containing dramatic, fast emotional speech, and high background noise levels, making ASR particularly challenging.

2.2. Available metadata

Metadata associated to the dataset presented in the last section varies over time, shows and media type. These can be more or less complete, accurate and reliable. In the following we first classify the metadata into three types. We then introduce the issues related to each type of metadata.

`type1`: transcriptions are produced manually and timestamps are provided (quantised to 1s) as well as speaker names and additional metadata such as indications of music or sound effects. This type of metadata is available for Radio4-1day, Desert Island Discs and the Archives dataset.

`type2`: transcriptions are not verbatim, timestamps are not provided and a number of errors which depend on the degree to which the speaker deviated from the original script. This type of metadata is typical of the Reith Lectures dataset in which scripts were used by lecturers from which they were free to deviate.

`type3`: transcriptions are derived from subtitles for hearing impaired, timestamps are provided as well as and other metadata such as an indication of music and sound effects, or indications of the way the text has been pronounced. Most of the shows include several speakers. Speaker identities are indicated by the use of several different text "colours" (which are used for subtitle display) . Timestamps were found to be unreliable due to time-lags that occur in subtitles, presumably arising from the re-speaking process for subtitle creation. This type of metadata is the one used for the TV-drama and TV-1week datasets.

These different types of metadata can be characterised in terms of completeness, accuracy and reliability. The metadata

can be more or less *complete*: the transcription can cover all the episode, or just a part of it, the timestamp information can also be available or not (e.g `type2`). The available metadata also varies over shows: some include speaker ID, sound event indications, title of music, programme genre. In terms of *accuracy*, transcriptions may include annotation of disfluencies and quantisation of the timestamps also may vary over shows (e.g 1ms for `type3` to 1s for `type1`). Finally, the *reliability* varies over the different types of metadata: `type1` include manual transcriptions and are considered to be more reliable even though they might include some variations depending on the transcriber and some episodes transcribed according to `type3` were found to have time-lag. Finally the reliability of `type2` metadata varies over episodes depending on speakers who can deviate differently from scripts.

3. Lightly Supervised Approaches

Most of the issues related to metadata described in the last section may be solved by lightly supervised approaches. In conventional lightly supervised training [6], a biased language model (LM) trained on the transcriptions (closed-captions) is used to recognise the training audio data. The recognition hypotheses are then compared to the close-captions and matching segments are filtered to be used in re-estimation of the acoustic model parameters. The entire process is carried out iteratively, until the amount of training data obtained converges. This kind of approach can first be used for the correction of timestamps when these are unreliable, imprecise or simply non-existent such as `type2` metadata. It then can be used when transcriptions are unreliable in order to select data for the training of acoustical models. We first describe our method for timestamp correction before presenting our approach for non-reliable transcription based on combined transcriptions. We finally investigate an evaluation technique based on ranking systems using imperfect reference transcription when no reference transcription is available.

3.1. Timestamp correction

Timestamps can be inaccurate due to quantisation effects (`type1`), unreliable due to time-lags that can occurs in subtitles (`type3`) or simply nonexistent (`type2`). They can however be corrected using a lightly supervised approach in the following way [14], which will also be used in section 3.2. Each show is first segmented and segments are clustered according to speakers using the CU RT-04 diarisation system [15]. Each speech segment is decoded using a two-pass[1] (P1-P2) recognition framework [16, 17] including speaker adaptation, with the decoding employing a biased language model (LM). This biased LM is initially trained on the original transcription (denoted as *origTrans* in the following) and then interpolated with a generic language model, with a 0.9/0.1 interpolation weight ratio. This results in an interpolated LM biased to the original in-domain transcripts. The vocabulary is chosen to ensure coverage of words from the original transcripts. The decoder output is then compared with the raw transcription to identify matching sequences. Non-matching word sequences from the raw transcription are force-aligned to the remaining speech segments. Finally, once realigned, the position of timestamps can be corrected.

[1] the output lattices generated in the second pass (P2 stage) when generating the 1-best hypotheses are used to generate confidence scores for both automatic transcriptions and the original transcriptions in section 3.2.

3.2. Combined transcriptions

There are two main issues with the conventional lightly supervised approaches related to `type2` metadata. As the original imperfect transcriptions deviate more from the correct ones, the constraints provided by the biased LM are increasingly less appropriate. This leads to a greater mismatch between the original transcriptions and the biased LM decoding hypotheses, which results in a reduction in the amount of usable training data after filtering is applied. Moreover, information pertaining to the mismatch between the original transcriptions and the automatic decoding outputs is normally measured at the sentence or word level. As acoustic models used in current systems are normally constructed at the phone level, the use of phone level mismatch information is preferable [9]. In [12], we proposed a method for the selection of training data using unreliable transcriptions. In this method, phone level mismatch information is used to identify reliable regions where segment-level transcription combination can be used. Schemes for combining the imperfect original transcriptions with the confusion networks (CN), generated during the biased LM decoding, can then be applied to leverage the different characteristics of the two forms of transcriptions.

3.2.1. Segment-level combination

Mismatch information at phone level is useful in order to derive combined transcriptions for the selection of training data. In order to exploit this information when the original and automatically decoded transcriptions disagree significantly, segment level phone difference rate[2] (PDR) is used to select the segments in the original transcriptions (*origTrans*) that can be combined with the automatically derived hypotheses (*aHyp*) outputs. To do so, (i) *origTrans* is first mapped into each of the *aHyp* segments using standard dynamic programming alignment, unmapped words being discarded. (ii) The mapped transcriptions are then force-aligned to obtain the phone sequences from which (iii) the PDR between the two force-aligned phone sequences can be calculated, if both exist. Finally, (iv) segment selection can be performed by selecting segments from *origTrans* which have a PDR values less than a threshold optimised on a held-out dataset. The remaining segments are then filled in to yield the transcriptions for the full training data set.

3.2.2. Word-level combination

When the mismatch between the original transcripts and the 1-best biased LM decoding hypotheses is large, the amount of training data is reduced dramatically. In this case, the hypotheses can be combined with the original transcripts by considering word level consensus networks [18], in order to limit this reduction. However, the assumption that the imperfect transcription is always present in the biased LM CN network can be too strong in cases like `type2` transcriptions in which lecturers may deviate significantly from their initial script. To handle this issue, a modified word level CN based transcription combination scheme can be used: if the word given by the original transcription is not found in the lattice, the word with the highest confidence score in the biased LM lattice is selected. To do so, (i) *origTrans* is first mapped into each of the *aHyp* segments as was carried out for the segment-level combination. (ii) Using the lattices generated in Section 3.1 to obtain the *aHyp* segments, the lattice arc posterior ratio (LAPR) presented in [19] is calculated as the confidence score (CS) for each word in *aHyp*. (iii) A "virtual" confidence score (because they are not confi-

[2] the traditional segment-level phone error rate is calculated but this is a PDR as there are no accurate transcriptions

dence scores in the usual sense) based on hard assignment is associated with each word in the mapped *origTrans*. If there are alternative word candidates in the lattices which agree with the word in *origTrans*, a score larger than the maximum value of LAPR is assigned as the confidence score (1.2), otherwise, the confidence score is set to 0.0. Finally, (iv) after confidence scores have been assigned to all words in both *aHyp* and in *origTrans*, ROVER [20] is used, taking the confidence scores into account, to do the transcript combination, yielding the final set of "best" word sequences for each segment.

3.3. Evaluation considering relative measures

Most lightly supervised training research has been focused on improving only the quality of the training transcriptions, assuming that the correct transcriptions are available for test data used in performance evaluation. However, for many practical applications accurate transcriptions that cover many diverse target domains can be impractical to manually derive for both the training and test data. Hence, alternative testing strategies that do not explicitly require correct test data transcriptions are preferred [21]. Here, we investigated the reliability of a performance rank ordering, given by the *origTrans* as an approximate reference transcription. Should such a rank ordering be consistent with that generated by the gold standard reference on the hand labelled data, it was then hoped that *origTrans* could be used for other larger sized test sets that don't have accurate transcripts

3.4. Experiments and results

To validate our proposed approach, experiments were run on the Reith Lectures dataset for which metadata are of `type2` as lecturers deviated more or less from their original prepared scripts during their speech. For the experiments, data were divided into a training set of 68 hours, a test set of 2.5 hours and two episodes of 0.8 hours of gold standard transcripts. A first comparison between *origTrans* and *aHyp* transcriptions carried out at the episode level, according to the word difference rate (WDR[3]) in the lecture regions, showed that difference rates vary strongly between speakers. The effectiveness of the segment and word level combination approaches was then validated on the gold standard transcrips, both word-level and best segment-level combined transcriptions achieving similar significant reductions in phone error rate (PER) and word error rate (WER) over the performance of the *origTrans* and *aHyp* transcriptions indicating that more accurate transcriptions could be obtained from the transcriptions combination. Given these preliminary results, we then investigated how real speech transcription systems are affected by training acoustic models using the combined training data transcriptions. Results obtained from the real transcription systems and detailed in [12] showed that both of the combination approaches investigated provide more accurate transcriptions than the original lightly supervised transcriptions, resulting in improved ML and MPE models. For MPE models, a reduction of 0.6% absolute and 1.1% absolute of WDR is obtained when using segment and word level combined transcriptions respectively, instead of *aHyp* (17.4% WDR), when added to a multi-genre broadcast dataset with accurate transcriptions. We also showed that rank ordering of the WER and WDR pairs derived from *origTrans* and from the gold standard transcript was consistent, allowing to use the *origTrans* as reference for other larger sized test sets that don't have accurate transcripts.

[3]The WDR is calculated in the same manner as the traditional word error rate, but this is a WDR as there are no accurate transcriptions

4. Multi-genre transcription using out-of-domain data

We now move our focus to a second aspect of the development of systems for the automatic transcription of media Archives which aim to effectively combine in-domain and out-of-domain training data. State-of-the-art transcription systems built for domains such as conversational telephone speech (CTS), and North American broadcast news (BN) perform with low accuracy on multi-genre data such as the BBC ones described in section 2. This is mostly due to the high mismatch in environment, speaking style, speaker and accent. Unsurprisingly, in-domain HMM-GMM systems trained on these data outperform these out-of-domain (OOD) systems, despite the fact that there is an order of magnitude less in-domain training data. For the purpose of the transcription of BBC archives, we then focused on the development of methods which can effectively combine in-domain and OOD training data using neural networks. Intensive research has been carried out recently on deep neural networks (DNNs) with promising results [22, 23]. We have used DNNs with generative pre-training to obtain posterior features used in the tandem framework [13] which is attractive for cross-domain modelling, since it allows independent adaptation of the GMM and DNN parameters. We recently proposed in [11] a novel technique called Multi-Level Adaptive Networks (MLAN) for posterior feature combination in a cross-domain setting. This technique, which will be presented below, has been investigated on a subset of the BBC dataset presented in section 2 in terms of cross-domain speech recognition using different acoustic training data sources across different target genres. It has then been evaluated in terms of a discriminatively-trained speaker adaptive speech recognition system, by comparing in-domain and out-of-domain (OOD) posterior features obtained using the proposed method.

4.1. Multi-Level Adaptive Networks

In our proposed method, DNNs are trained to model frame posterior probabilities over monophones. The structure of the DNNs is fixed following analysis of the frame error rate on held-out validation data and monophone log-posterior probabilities output from the nets are decorrelated using a single PCA transform with dimensionality reduced to 30 [13] to obtain the posterior features. These are then concatenated with the original acoustic features. Using initial OOD DNN adapted to a new domain, can be viewed as imposing a form of regularisation on the resulting net. However we observed small benefits when using deep architectures and fairly large quantities of in-domain data. We therefore proposed an alternative adaptation procedure called *Multi-level Adaptive Networks* (MLAN). In the first level of this MLAN scheme, networks trained on OOD acoustic data are used to process in-domain acoustic data to generate posterior features, which are concatenated with the original in-domain acoustic features as in the tandem framework. We would expect the OOD posterior features to enhance the discriminative abilities of the simple in-domain acoustic features. In the second level, additional DNNs are trained, using the first level tandem features as input, to minimise an in-domain objective function of log-posterior phone probabilities. The outputs from these DNNs are then used to generate the final tandem features for HMM training. Finally, by expanding the input tandem feature vector used at the second level, output from multiple networks, trained on different domains, may be included with no modification to the architecture. The main motivation for the MLAN scheme is that the new DNNs, trained discriminatively,

Feature set	1-pass (unadapted)				2-pass (adapted)			
	Studio	Location	Drama	All	Studio	Location	Drama	All
PLP	12.0	25.9	58.8	32.7	11.5	23.6	58.9	31.8
BBC tandem	11.7	23.3	54.9	30.4	11.3	22.3	54.4	29.8
AMI tandem	11.3	22.6	55.0	30.1	11.1	21.5	54.2	29.4
AMI+CTS MLAN	10.2	20.9	50.5	27.6	9.8	20.0	50.2	27.1

Table 1: Final MPE system results (WER%) on the 2.3h test set using PLP, tandem and MLAN features.

are able to learn which elements of the OOD posterior features are useful for discrimination in the new domain; whilst the direct inclusion of in-domain acoustic features in the input means that the resulting frame error rates ought never to be worse than DNNs trained purely in-domain. The additional generative pretraining carried out ensures that the new DNN does not over-fit to the in-domain data. More details (e.g DNN structure) and explanation of the method can be found in [11].

4.2. Experiments

Experiments were conducted on the Radio4-1day and the TV-drama dataset divided into the three categories by broad genres defined in Section 2.1 (`studio`, `location`, `drama`). Transcriptions were found to be reliable but timestamps were corrected according to the procedure detailed in Section 3.1, giving a total of 23 hours of transcribed and aligned speech in total. The data were divided at the show level into a training set of 20.7 hours and a test set of 2.3 hours, each containing roughly the same balance across genres. For the out-of-domain data, two diverse sets were used. The first one included 277 hours of US-English conversational telephone speech (CTS) taken from the switchboard I, switchboard II and CallHome corpora. The second set consisted of Recordings from the Augmented Multi-Party Interaction (AMI) corpus. Concerning the system architectures, development experiments were performed using a simple one-pass system and the final evaluation system was trained using MPE discriminative training [24] and had a two-pass decoding architecture.

4.2.1. Development experiments

Recognition of the test set was first performed using two OOD acoustic models trained on PLP features from the AMI and CTS training set. The results demonstrate the large acoustic mismatch between these domains and the BBC domain. The performance of tandem features was then investigated by comparing models trained purely on the BBC training set with models trained on tandem features obtained using OOD nets. It was found that OOD tandem features from AMI and CTS improved performance for all genres (with the overall WER initial value equal to 39.4% reduced by 5.6% absolute and 3.9% absolute using AMI and CTS features respectively) compared to simple PLP features supporting earlier work suggesting that posterior features are portable across domains. With respect to the broad genres, it was found that CTS and AMI OOD posteriors are both better for Studio speech by comparison with the BBC tandem results, AMI is best for Location speech and equally matched with in(domain features for Drama speech, which is the genre most mismatched to the OOD acoustic models. Performance of the MLAN was then investigated and showed substantial additional gains over standard tandem features, for both domains. The CTS posteriors which were worst-matched to the BBC domain, gain the most benefit from MLAN with a 3.6% absolute WER reduction overall (initial value 35.5%). The combination of both OOD posterior features with MLAN reduces WER still further, suggesting the second-level DNN is successfully able to exploit complementary information between AMI and CTS.

4.2.2. Final system evaluation

For the final system evaluation, the best-performing in-domain and out-of-domain tandem features, and the best MLAN features, were selected for use in training a more competitive final system. Table 1 shows the final system results on the test set with and without speaker adaptation. The HMMs were trained with MPE only on the BBC training set using STC-projected PLP features and the relevant posterior features. All the new features outperformed the baseline PLP features in both the unadapted and speaker adapted MPE systems. This supports the preliminary results from the development system and indicates that the posterior features can bring complementary information to the PLP features even when the HMMs are trained using MPE. Moreover, the overall improvement over the baseline PLP features, in both the unadapted speaker-adapted systems was dramatic, with absolute WER reductions of 5.1% and 4.7% respectively. Table 1 shows that speaker adaptation is effective in reducing the WER for all three posterior feature sets, compared with the baseline PLP features which only offers gains for the Location and Studio subsets, although for these two subsets, the gains from adaptation are larger than for the posterior features. It was then hypothesised that the posterior features are better able to capture speaker-invariant information in these subsets, whilst in the noisy drama subset, are able to model speaker-dependent structures more effectively than PLPs.

5. Conclusions and Future work

We presented our joint work on the development of a speech recognition system for multi-genre media archives from the BBC using limited text resources. We first described the different BBC datasets which were provided with their diverse audio content and metadata.We then focused on improving the transcription quality of acoustic model training data for the BBC archive task. Combination at both the word and segment level-level of the original transcriptions, with the lightly supervised transcription generated by recognising the audio using a biased language model has been presented. This provides more accurate transcriptions than the original lightly supervised transcriptions, resulting in improved models. We then presented the MLAN method for recognition of multi-genre media archives with neural network posterior features, successfully using out-of-domain data to improve performance. Results consistently show that our Multi-Level Adaptive Networks scheme results in substantial gains over over other systems including a PLP-based baseline, in-domain tandem features and the best out-of-domain tandem features. Future work will investigate further transcription combination approaches and testing schemes with imperfect transcription references. We also plan to investigate the MLAN technique in an HMM-GMM system that also incorporates speaker-adaptive training and fMPE transforms and to adapt the method for use in a hybrid DNN system. Finally the proposed approaches will be conducted on larger datasets such as Archives and TV-1week.

6. References

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

[2] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An audio indexing system for election video material," in *Proc ICASSP*, 2009, pp. 4873–7876.

[3] R. C. van Dalen, J. Yang, and M. J. F. Gales, "Generative kernels and score-spaces for classification of speech: Progress report," in *Tech. Rep. CUED/g-infeng/th.676, Cambridge University Engineering Department*, 2012.

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

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

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

[7] H. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. ICASSP*, vol. 1, 2004, pp. 737–740.

[8] L. Mathias, G. Yegnanarayanan, and J. Fritsch, "Discriminative Training of Acoustic Models Applied to Domains with Unreliable Transcripts," in *Proc. ICASSP*, vol. 1, 2005, pp. 109–112.

[9] B. Lecouteux, G. Linares, P. Nocera, and J. Bonastre, "Imperfect transcript driven speech recognition," in *Proc. InterSpeech'06*, 2006, pp. 1626–1629.

[10] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, V. Gadde, and J. Zheng, "An Efficient Repair Procedure for Quick Transcriptions," in *Proc. ICSLP*, 2004.

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

[12] 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.

[13] H. Hermanksy, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1630.

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

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

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

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

[18] L. Chen, L. Lamel, and J.-L. Gauvain, "Lightly supervised acoustic model training using consensus networks," in *Proc. ICASSP*, vol. 1, 2004, pp. 189–192.

[19] M. Seigel and P. Woodland, "Combining information sources for confidence estimation with CRF models," in *Proc. Interspeech*, 2011, pp. 905–908.

[20] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. ASRU Workshop*, 1997, pp. 347–352.

[21] B. Strope, D. Beeferman, A. Gruenstein, and X. Lei, "Unsupervised Testing Strategies for ASR," in *Proc. Interspeech*, Florence, Italy, 2011.

[22] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[23] A. Mohammed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[24] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, pp. 105–108.