Improving Lightly Supervised Training for Broadcast Transcription

Y. Long, M.J.F. Gales, P. Lanchantin, X. Liu, M.S. Seigel, P.C. Woodland

Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.

{yl467,mjfg,pkl27,xl207,mss46,pcw}@eng.cam.ac.uk

Abstract

This paper investigates improving lightly supervised acoustic model training for an archive of broadcast data. Standard lightly supervised training uses automatically derived decoding hypotheses using a biased language model. However, as the actual speech can deviate significantly from the original programme scripts that are supplied, the quality of standard lightly supervised hypotheses can be poor. To address this issue, word and segment level combination approaches are used between the lightly supervised transcripts and the original programme scripts which yield improved transcriptions. Experimental results show that systems trained using these improved transcriptions consistently outperform those trained using only the original lightly supervised decoding hypotheses. This is shown to be the case for both the maximum likelihood and minimum phone error trained systems.

Index Terms: lightly supervised training, speech recognition, confidence scores

1. Introduction

In order to robustly estimate the acoustic model parameters for large vocabulary continuous speech recognition (LVCSR) tasks, a large amount of training audio data along with accurate transcriptions is required. Obtaining accurate manual transcriptions is an expensive task. Therefore it is desirable to use manual transcripts which are only partially correct, such as closed captions or subtitles. In order to use this type of data, a range of techniques have been proposed. In conventional lightly supervised training [1], a biased language model (LM) trained on the closed-captions is used to recognise the training audio data. The recognition hypotheses are then compared to the close-captions. Matching segments are filtered to be used in re-estimation of the acoustic model parameters. This entire process is carried out iteratively, until the amount of training data obtained converges. A range of techniques have since been proposed along this line to improve upon this lightly supervised training method. Different data filtering methods were investigated for discriminative training in [2, 3]. Alternative representations of biased LMs that aim to capture the deviation of imperfect transcriptions from the correct ones were proposed in [4, 5]. Word level consensus networks (CN) were used to improve transcription quality in the regions where mismatches between the imperfect transcriptions and the biased LM decoding hypotheses occur in [6].

There are three issues with the conventional lightly supervised training approaches. First, as the original imperfect transcriptions deviate more from the correct ones, the constraints provided by the biased LM are increasingly weakened. This leads to an enlarged 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. Second, 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 [7]. Finally, most lightly supervised training research has been focused on improving only the quality of the training transcriptions. It is assumed that the correct transcriptions are available for test data used in performance evaluation. However, for many practical applications, such as the broadcast lecture transcription task considered in this paper, 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 [8, 9].

This paper investigates improving lightly supervised acoustic model training in the context of an archive of broadcast lectures. The original transcriptions (`OrigTrans`) provided for this task are not true manual transcriptions, but are programme scripts that contain no time-stamp information. Their quality therefore varies with respect to the extent that individual speakers deviate from the scripts. When the actual speech deviates significantly from the original programme scripts, the quality of standard automatically derived lightly supervised decoding hypotheses using a biased LM can be unreliable. In order to address the issues mentioned above, 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 are presented. These are able to leverage differences in the characteristics of the two forms of transcription information, and yield improved combined transcriptions. Since accurate verbatim transcripts are unavailable for the test data, an evaluation technique based on ranking systems using imperfect reference transcripts is used to evaluate system performance in this work.

The rest of the paper is organised as follows. In section 2, the conventional lightly supervised training is reviewed. section 3 presents the transcription combination schemes. The database of broadcast lectures is described in section 4. Experiments and results are presented in section 5, followed by the conclusion in section 6.

2. Lightly Supervised AM training

Several steps are taken in this work to do the lightly supervised acoustic model (AM) training. First, text normalisation is applied to the original transcriptions to correct obvious transcription errors. Then, the automatic transcriptions are generated for

The research leading to these results 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.

the audio data, before these two sets of transcriptions are combined. Finally, these combined transcriptions are used for AM training.

The automatic transcriptions of the audio data are the lightly supervised decoding hypotheses generated using an approach similar to that used for lightly supervised training and unsupervised training on broadcast news data in [2] and [10]. A biased LM was built and used to generate transcriptions.

A language model is initially trained using all of the original transcriptions. This LM is then interpolated with a generic language model, with the highest interpolation weight on the component trained on the original transcriptions which results in an interpolated LM biased to the original in-domain transcripts. Interpolation with a generic background LM is necessary so that reasonable language model scores are assigned when the speakers deviate from the original transcripts.

The automatic transcriptions were generated using an acoustic model, trained on accurate transcriptions of other broadcast data. A standard two-pass (P1-P2) recognition framework [11, 12], with the second pass including unsupervised adaptation, was used to decode of the automatically segmented and speaker-clustered training audio data using the interpolated biased LM. The output lattices generated in the second pass (P2 stage) when generating the 1-best hypotheses are used to estimate the confidence scores for both the automatic transcriptions and the original transcriptions in Section 3.2.

3. Transcription combination

Segment and word level combination approaches that are used to improve the transcription quality for lightly supervised training are presented in this section.

3.1. Segment-level combination

As discussed in section 1, it is useful to exploit phone level mismatch information when the original and automatically decoded transcriptions disagree significantly. Mismatch information at this level is useful as the goal is to use the combined transcriptions to train the context-dependent triphone acoustic models. In the segment level transcription combination method considered in this paper, the segment level phone difference rate (PDR) is used to select the segments in the original transcriptions that can be combined with the automatically derived hypotheses (`AHyp`) outputs. The traditional segment-level phone error rate is calculated, but this is a PDR as there are no accurate transcriptions. The entire combination process takes the following steps:

- **Step 1**: map the episode or show level original transcriptions (`OrigTrans`) into each of the `AHyp` segments using a standard dynamic programming alignment. Unmapped words are discarded.
- **Step 2**: force-align the mapped `OrigTrans` and `AHyp` transcriptions to obtain the phone sequences.
- **Step 3**: calculate the PDR between the above two phone sequences, if both exist (some segments may not have a mapped `OrigTrans`).
- **Step 4**: select the segments from `OrigTrans` which have PDR values less than a threshold optimized on a held-out dataset; fill in the remaining segments with `AHyp` to yield the transcriptions for the full training data set.

3.2. Word-level combination

As discussed in section 1, when the quality of the original imperfect transcriptions deteriorates, the mismatch between the original transcripts and the biased LM decoding hypotheses is large. This results in a reduction in the amount of usable training data after filtering is applied. One technique to address this issue is to use word level consensus networks (CN) in the regions where mismatches occur [6]. It is assumed that the imperfect transcription is always present in the biased LM CN network. This means that a different word with a confidence score above a certain threshold generated by the biased LM decoding, or the word given by the imperfect transcription, should be selected when performing the combination. However, for the broadcast lecture transcription task considered in this paper, this assumption can be too strong. When the search constraints used by the biased LM are weakened, the original transcription is no longer guaranteed to be present in the biased LM output lattices. To handle this issue, a modified word level CN based transcription combination scheme is 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. The algorithm consists of the following four steps:

- **Step 1**: the `OrigTrans` transcriptions are first mapped into the `AHyp` segments, as was carried out for the segment-level combination.
- **Step 2**: using the lattices generated in Section 2 to obtain `AHyp`, the lattice arc posterior ratio (LAPR) presented in [13] is calculated as the confidence score (CS) for each word in `AHyp`.
- **Step 3**: a "virtual" confidence score based on a 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.
- **Step 4**: after all words in both `AHyp` and `OrigTrans` are assigned confidence scores, ROVER [14] is used, taking the confidence scores into account, to do the transcript combination, yielding the final set of "best" word sequences for each segment. This is then used to train the acoustic model.

In step 3, hard scores are described as "virtual" confidence scores because they are not confidence scores in the usual sense. They are used to represent whether or not we are prepared to accept the words in `OrigTrans`. If such a word has any acoustic evidence, as indicated by occurring in the recognition lattice, it is accepted as being correct, with competing words in the lattices considered to be recognition errors. This is the basis for the assignment of a confidence score larger than any possible competing score.

4. Corpora: The BBC Reith Lectures

In a collaboration with the British Broadcasting Corporation (BBC) Research and Development, we have already investigated the automatic transcription of broadcast materials across multi-genre media archives in [15]. In this paper, an archive of the BBC's Reith Lectures are investigated for our lightly supervised acoustic model training technique.

4.1. Data description

The Reith Lectures [1] are a series of annual radio lectures on significant contemporary issues, delivered by leading figures from their relevant fields. There are only original scripts containing no time-stamps available from the BBC for these radio lectures. The original scripts are not verbatim transcriptions, and contain a number of errors (such as substitutions, deletions and insertions), which depend on the degree to which the lecturers deviated from their original prepared scripts during their speech.

The Reith Lectures used in this work consists of 155 episodes, covering the years from 1976 to 2010. Only one lecturer was invited to speak in most years, with the exception of several special years. Each lecturer had 3-6 episodes presented at different times. We manually labelled two regions within each episode: the lecture region given by the lecturer, and a non-lecture region which contained the introduction to the lecture by another presenter and, since 1988, a question and answer session after the main lecture. Only the lecture regions of the data were used for acoustic model training and evaluation. The duration of each episode ranged from 18-35 minutes, to give a total audio duration of 72 hours. Each episode was first automatically segmented, with STT-based speaker clustering performed as in [16]. Long silence and applause were discarded during the automatic segmentation, resulting in around 71.3 hours of lecture region data finally being retained. We divided this data into a training set of 68 hours (**rl.train**), a test set of 2.5 hours (**rl.eval**) and two episodes of with a 0.8 hour lecture region for which gold-standard transcripts (**rl.geval**) were created to guide our research on this data.

4.2. Initial investigation of the quality for `origTrans`

Before using the original transcripts of the Reith Lectures' to train acoustic models, a preliminary experiment was carried out to investigate the quality of `OrigTrans`. The automatic lightly supervised decoding hypotheses (`AHyp`) were compared with `OrigTrans` at the episode level, to calculate the word difference rate (WDR) in the lecture regions. Similar to the PDR defined in Section 3.1, the WDR is calculated in the same manner as the traditional word error rate. Results on several sample episodes are shown in Table 1.

Episodes	Sub	Del	Ins	WDR
1991+STEV+JON+THEX+LR2	2.3	0.7	0.5	3.5
2005+ALEC+BRO+THEX+LR2	2.7	0.9	0.9	4.5
2003+VSXX+RAM+THEX+LR2	9.5	4.2	1.8	15.5
2000+TOMX+LOV+RESP+LR2	8.9	0.8	7.0	16.7
2004+WOLE+SOY+THEX+LR2	10.7	17.6	7.7	36.0
2007+JEFF+SAC+BURS+LR2	12.5	35.9	5.6	54.0
Average	7.9	10.5	3.9	22.3

Table 1: Word difference rates on the lecture regions of several sample episodes , comparing `AHyp` against `OrigTrans` transcripts: substitution rate (%Sub), deletion rate (%Del), insertion rate (%Ins) and word difference rate (%WDR).

From the results in Table 1, it is clear from the range of different substitution, deletion and insertion rates that the mismatch between `AHyp` and `OrigTrans` varies for different episodes/lecturers. The difference in error rates for the sample episodes in the three blocks show that the similarity between `OrigTrans` and `AHyp` is either high, partial or minimal, respectively. This indicates that `OrigTrans` are imperfect, while they may occasionally be reliable, different speakers tend to deviate from their original scripts to different extents. A more in-depth comparison was undertaken by listening to the audio with potentially large mismatches between `AHyp` and `OrigTrans`. The results of this supported our hypothesis of mismatches due to script deviations.

[1] http://www.bbc.co.uk/programmes/b00729d9

4.3. Validation results on data set: rl.geval

In this section, the effectiveness of the segment and word level combination described in section 3 are validated on the **rl.geval** data set which has gold-standard reference (GRef) transcriptions. Taking GRef as reference, the traditional phone error rate (PER) and word error rate (WER) for the `AHyp`, `OrigTrans`, segment-level and word-level combined transcriptions are shown in the first, second and third block of Table 2. It can be seen that the word-level and the best segment-level combined transcriptions achieved similar significant reductions in PER and WER over the performance of `AHyp` and `OrigTrans`. This indicates that more accurate transcriptions can be obtained from the transcription combination, which should be useful for acoustic model training.

Transcription	PER	WER
`AHyp`	3.9	7.3
`OrigTrans`	4.1	5.4
SegComb[10]	3.5	6.0
SegComb[15]	3.3	5.3
SegComb[20]	3.0	4.6
SegComb[25]	2.9	4.5
SegComb[30]	3.0	4.5
SegComb[35]	3.2	4.7
WrdComb	3.2	4.7

Table 2: %PER (phone error rate) and %WER (word error rate) on the 0.8 hours **rl.geval** data set with different transcriptions: `AHyp` (automatic lightly supervised decoding hypotheses), `OrigTrans` (original transcriptions provided by BBC), SegComb[xx]: segment-level combined transcriptions using different PDR thresholds, WrdComb: word-level combined transcriptions.

5. Experiments and Results

The validation results from Table 2 showed that transcription combination worked well on **rl.geval** both at segment and word-level. It was therefore worthwhile investigating how the real speech transcription systems are affected by training acoustic models using the combined training data transcriptions.

5.1. Acoustic model training setup

Tied-state cross-word triphone HMM-GMM acoustic models were constructed using decision tree clustering for all of our systems. The basic features used were the perceptual linear prediction (PLP) coefficients with their 1st, 2nd and 3rd temporal derivatives, projected down to 39 dimensions with an HLDA transform. Both maximum likelihood (ML) and minimum phone error (MPE) training were performed. The baseline acoustic models (AMs) were trained on the 20.7 hours (**bbc.train**) BBC accurate multi-genre broadcast transcriptions which were used in [15]. 3000 tied-states with 12 Gaus-

sian mixture components per state were used for the baseline system, and 3600 tied-states with 16 components were used for all the systems trained on the total 88.7 hours data by adding the Reith Lectures training data into the **bbc.train** dataset.

The baseline MPE acoustic models trained on **bbc.train** were used to generate AHyp. The 65k word biased trigram LM was obtained by interpolating the LM trained on the Reith Lectures (origTrans), with a generic trigram LM trained on North American Broadcast News [12] using weights of 0.9 and 0.1 respectively.

5.2. Test setup

Single pass decoding systems without speaker adaptation were used for testing. The same generic trigram LM used for the biased LM interpolation was also interpolated with a LM trained on 0.63M words from both the lecture and non-lecture regions of the data using a weight of 0.74, to yield the final 65k LM for testing.

5.3. Relative measures

As discussed in section 1, for the broadcast lecture transcription task considered in this paper, accurate transcriptions that cover diverse target domains are costly to manually derive for both the training and test data. We only have a gold-standard reference (GRef) for the **rl.geval** set, and the BBC original transcriptions (OrigTrans) as reference for the **rl.eval** test set. Here, we investigated the reliability of a performance rank ordering given by 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 is then hoped that OrigTrans can be used for other larger sized test sets that don't have accurate transcripts. This approach is different from using the AHyp generated from a strong AM and LM as the assumed truth to achieve the relative measures in [8]. Here, OrigTrans is used as the reference instead of AHyp, as the AM used for lightly supervised decoding was not strong enough to generate a reliable version of AHyp, thus the AHyp is biased to the acoustic models used to generate the transcriptions. Whereas the BBC original transcriptions are not biased to any individual system.

5.4. Experimental results

Table 3 gives the unadapted single pass decoding results on the two episodes 0.8 hours **rl.geval** set for different training data sets. WERs and WDRs are computed by using the golden-standard GRef and original transcriptions (OrigTrans) in each case respectively. From the table, the system using 68.0h **rl.train** with automatically recognised AHyp transcriptions alone achieved much better performance than one using 20.7h multi-genre broadcast **bbc.train** with accurate transcriptions, for both ML and MPE models. This may be due to the acoustic and linguistic mismatch between the multi-genre broadcasts and the Reith Lectures. By adding the AHyp, segment-level (SegComb[25]) and word-level (WrdComb) combined transcriptions to the **bbc.train** data, further improvements are obtained due to a better acoustic match and the increased amount of in-domain training data. However, adding both the segment-level and word-level combined transcriptions only yield slightly improved performance over adding the original recognised transcriptions on these two episodes. Furthermore, it is observed that the same relative performance rank ordering is achieved by comparing the corresponding WER and WDR pairs derived from the hand labelled **GRef**, and the approximate reference (OrigTrans) for systems trained on different data. We therefore assume that OrigTrans can be used for evaluation of the larger **rl.eval** test set.

Training data set	**GRef**-WER		OrigTrans-WDR	
	ML	MPE	ML	MPE
bbc.train	26.8	24.5	30.7	28.4
AHyp	24.5	22.0	28.4	25.9
bbc.train+AHyp	22.6	20.2	26.6	24.1
bbc.train+SegComb[25]	22.7	19.7	26.6	23.5
bbc.train+WrdComb	22.5	20.0	26.4	23.8

Table 3: WER/WDR in % on the 0.8 hours **rl.geval** test set for different training data sets.

Results in Table 4 on the 2.5 hours **rl.eval** test set reinforce the observation that performance is improved using AHyp transcriptions alone and when added to the small **bbc.train** data set. Interestingly, it is seen from this table that the combined transcriptions achieved much better performance than those originally recognised AHyp. This indicates that the **rl.eval** set benefits more from the inclusion of matched information captured from the partially correct Reith Lecture transcriptions than the **rl.geval** set. Furthermore, compared to adding the segment-level combined transcriptions into the **bbc.train** dataset, adding the word-level combined transcriptions obtained almost the same ML performance, but yielded larger relative performance gains for the MPE models. In addition, the MPE models achieved larger relative WDR reductions than ML models in adding both of the combined transcriptions compared to adding the AHyp. This indicates that MPE models are more sensitive to the accuracy of the training data transcriptions. Therefore, we believe that by combining the automatically recognised transcriptions with the original ones, more accurate transcriptions are obtained which allows us to train improved acoustic models.

Training data set	ML	MPE
bbc.train	27.1	24.9
AHyp	19.6	18.0
bbc.train+AHyp	19.3	17.4
bbc.train+SegComb[25]	18.7	16.8
bbc.train+WrdComb	18.6	16.3

Table 4: %WDR on the 2.5 hours **rl.eval** test set for different training data sets.

6. Conclusion

This paper has primarily focused on improving the transcription quality of acoustic model training data for the BBC lecture archive task. The combination at both the word and segment-level of the original transcriptions, with the lightly supervised transcription generated by recognising the audio using a biased language model has been presented. The results obtained in the validation experiments, as well as in the real transcription systems, show that both of the combination approaches investigated provide more accurate transcriptions than the original lightly supervised transcriptions, resulting in improved ML and MPE models. Further transcription combination approaches and testing schemes with imperfect transcription references will be investigated in future work.

7. References

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

[2] H.Y. Chan and P.C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training", in *Proc. ICASSP 2004*, pp. 737-740, Hong Kong.

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

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

[5] P. J. Moreno and C. Alberti, "A Factor Automaton Approach for the Forced Alignment of Long Speech Recordings", in *Proc. ICASSP 2009*, pp. 4869-4872, Taipei.

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

[7] B. Lecouteux, G. Linares, P. Nocera and J.F. Bonastre. "Imperfect transcript driven speech recognition", in *Proc. INTERSPEECH 2006*, pp. 1626-1629.

[8] B. Strope, D. Beeferman, A. Gruenstein and X. Lei, "Unsupervised Testing Strategies for ASR", in *Proc. INTERRSPEECH 2011*, pp. 1685-1688.

[9] J.D. Williams, I.D. Melamed, T. Alonso, B. Hollister and J. Wilpon, "Crowd-sourcing for difficult transcription of speech", in *Proc. ASRU 2011*, pp. 535-540.

[10] L. Wang, M.J.F. Gales and P.C. Woodland, "Unsupervised training for mandarin broadcast news and conversational transcription", in *Proc. ICASSP*, IEEE, 2007, pp. IV-353-356.

[11] G. Evermann & P.C. Woodland " Design of fast LVCSR Systems", in *Proc. ASRU Workshop 2003*. St. Thomas.

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

[13] M.S. Seigel and P.C. Woodland, "Combining Information Sources for Confidence Estimation with CRF Models", in *Proc. Interspeech 2011*, pp. 905-908.

[14] JG. 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.

[15] P.J. Bell, M.J.F. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski and P.C. Woodland, "Transcription of multi-genre media archives using out-of-domain data", in *Proc. SLT*, IEEE, 2012. pp. 324-329.

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