Unsupervised Training with Directed Manual Transcription for Recognising Mandarin Broadcast Audio

K. Yu, M.J.F. Gales and P.C. Woodland

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK
Email: {ky219, mjfg, pcw}@eng.cam.ac.uk

ABSTRACT

The performance of unsupervised discriminative training has been found to be highly dependent on the accuracy of the initial automatic transcription. This paper examines a strategy where a relatively small amount of poorly recognised data are manually transcribed to supplement the automatically transcribed data. Experiments were carried out on a Mandarin broadcast transcription task using both Broadcast News (BN) and Broadcast Conversation (BC) data. A range of experimental conditions are compared for both maximum likelihood and discriminative training using directed manual transcription. For BC data, using fully unsupervised discriminative training, only 17% of the reduction in character error rate (CER) from supervised training is obtained. By automatically selecting 18% of the data for manual transcription yields 50% of the CER gain from supervised training. The directed approach to selecting data outperforms the use of a random set of data for manual transcription.

***Index Terms*—** Speech Recognition, Unsupervised Training

1. INTRODUCTION

A recent trend in building high performance speech recognition systems is to use very large training sets for parameter estimation of the acoustic models. For some tasks, such as the automatic transcription of broadcast data, it is fairly easy to obtain thousands of hours of audio data from radio and television shows. However, in order to train the acoustic models word level transcriptions are also required. Hence the major cost of using large amounts of broadcast data is the provision of accurate manual transcriptions.

In some cases, approximate manual transcriptions, such as closed captions, are available, These approximate transcriptions can be used with lightly-supervised training [1] in which a biased language model is created from the approximate transcriptions and used to recocognise the audio data. This leads to low error rate semi-automatic transcripts which yield good performance for hidden Markov model (HMM) parameter estimation using both maximum likelihood (ML) and discriminative parameter estimation [2]. However, in other cases, for instance for broadcast news transcription from Arabic and Mandarin, even approximate transcriptions are not available. In these scenarios, *unsupervised training* techniques need to be used.

Two strategies may be used for unsupervised training. The standard approach is to automatically recognise the audio using a seed model. A method of data selection can be applied to remove data that is believed to be poorly transcribed. The selected data are added to the original training dataset to update the acoustic model and optionally the language model [3, 1, 4]. An alternative strategy, based on the theory of active learning [5], is to automatically select a small amount of data which is believed to be poorly recognised data for manual transcription to supplement the fully automatic transcripts: this is referred to here as *directed manual transcription*. The underlying assumption is that the inclusion of correctly transcribed high-error rate data is likely to be more useful for improving the quality of the acoustic model.

State-of-the-art speech recognition systems make use of discriminative objective functions such as the Minimum Phone Error (MPE) criterion [8]. However most previous studies of unsupervised training have examined only ML estimation of HMM parameters using the standard unsupervised training approach, although unsupervised discriminative training has been investigated in [9, 4]. Since, discriminative training aims to reduce the difference between the recognised output and the (assumed) "correct" transcription, it is unsurprising that it is far more sensitive to the accuracy of the transcriptions than ML training [4]. This sensitivity may restrict the range of data types which may be successfully used. In the Mandarin transcription task, since Broadcast Conversation (BC) data is more poorly transcribed by the seed model than the Broadcast News (BN) data, the reduction in character error rate (CER) relative to that from supervised training is far smaller for BC data than BN data [4].

This paper examines the use of the directed manual transcription strategy for unsupervised discriminative training for the transcription of Mandarin BN and BC data. Section 2 gives details of the unsupervised training procedures, Section 3 describes the experimental setup and then Section 4 gives experimental results for both ML and MPE trained models with varying amounts of fully unsupervised and manually transcribed data.

2. DIRECTED MANUAL TRANSCRIPTION

This section briefly describes the general procedure used for unsupervised training with directed manual transcription. The general setup is based on that described in [4] which includes: initial segmentation of the unsupervised data; automatic transcription generation; data selection; and use in ML and MPE training.

The segmentation/clustering stage creates and clusters speech segments from the raw audio for recognition and unsupervised adaptation. The procedure used is the same as described in [10]. First, commercials are removed by detecting repeated blocks of audio data. The data is segmented into homogeneous speaker/acoustic condition blocks and any audio labelled as music discarded. Finally, gender

This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. Many thanks go to X.A. Liu for training some of the LMs used in the experiments.

detection and speaker clustering are performed.

The initial transcription for the data without manual transcripts is generated using a multi-pass decoding framework with MPE models trained on about 186 hours of data. In these training data, Mandarin BN data predominates, and about 11 hours of English data is included which along with an appropriate language model and word list allows the system to generate both Mandarin and English output as appropriate. A two-pass (P1-P2) decoding setup is used which is the same overall design as for the lattice generation stage in the full CU multi-pass recognition system [10], The P1 stage generates an initial transcription which is used in the generation of adaptation transforms for each of the segment clusters. Lattices are generated using the adapted models with a trigram language model. These lattices are then rescored using a 4-gram language model and confusion network decoding used to generate the final output and a confidence score for each recognised word. It is worth noting that in this work, 36 component MPE models trained on the 186 hours of data are used in the P1-P2 decoding. This is slightly different from [4], where 16 component models were used.

To perform lattice-based MPE training, it is necessary to generate lattices corresponding to both the “correct” transcription (“numerator” lattices) and “denominator” lattices that correspond to output of a recognition system. The denominator lattices are generated using the ML model trained on all data including the unsupervised data and a heavily pruned bigram model.The numerator lattices are generated by aligning data with the same ML model to the initial automatic transcription. Note that it is the use of a non-adapted ML model and a simple language model that yields a difference in recognition performance between the assumed correct recognition output and the 1-best string from the denominator lattices. For the case of directed manual transcription, the denominator and numerator lattices for the selected data are regenerated before MPE training.

To select which shows to manually transcribe, the show-averaged confidence scores are used. First any audio broadcast which is believed to not contain Mandarin is detected based on the show-averaged confidence score and the percentage of English words recognised [4] and removed. The show-averaged confidence scores are calculated by averaging the word level confidence scores:

$$C_{\mathcal{S}} = \frac{\sum_{\mathcal{W}\in\mathcal{S}} C_{\mathcal{W}} T_{\mathcal{W}}}{\sum_{\mathcal{W}\in\mathcal{S}} T_{\mathcal{W}}}$$

where $C_{\mathcal{S}}$ is the show level confidence score of show $\mathcal{S}$, $C_{\mathcal{W}}$ is the word level confidence score of word $\mathcal{W}$, $T_{\mathcal{W}}$ is the duration of the word. A threshold on $C_{\mathcal{S}}$ is set to split the unsupervised data into two parts: shows with $C_{\mathcal{S}}$ lower than the threshold are selected to be manually transcribed, otherwise the automatic transcription is used.

For all experiments on HMM training, the 36 component ML system trained on 186 hours of data is used as the starting point. The extra data (either in fully unsupervised mode or with directed manual transcriptions) are then added to the original training data to update acoustic model parameters. ML training is then performed which also acts as the initialisation for subsequent MPE training.

3. EXPERIMENTAL SET-UP

3.1. Baseline acoustic model

The baseline acoustic model, **S0**, is the same as the baseline used in [4], which was trained on the 186.4 hours of manually transcribed data. This consists of 155.6 hours of Mandarin BN data, 19.6 hours of BC data, and 11.1 hours of English data including 10 hours of randomly selected TDT4 English data. The basic acoustic features for all the recognition system were 13 cepstral coefficients (including energy) derived from MF-PLP analysis and segment level CMN. The static cepstra were appended with 1st, 2nd and 3rd order derivatives to form a 52-dimensional feature vector and then projected using a HLDA transform to 39-dimensions. Pitch was extracted, and added to the feature vector along with its 1st and 2nd order derivatives. State-clustered triphone HMMs, with 6K distinct states and an average of 36 Gaussian components per state were used. The same decision tree and HLDA transform was used for all systems in this paper. The total number of Gaussian components per state was also kept fixed at an average of 36.

3.2. Baseline language model

The baseline language model was the same as used in [4]. All text was processed using a simple character to word segmenter based on a longest-first match. The multi-character word-list for this consisted of about 51K words. The total word-list, including single-character Mandarin words and the 10K most frequent English words, was 68K in size. The language models used in these experiments were trained using various sources including the LDC Chinese Gigaword release; web download data and audio data transcripts used for the **S0** model. Three separate LM components were built and interpolated to construct the baseline language model. The first trained on Gigaword and broadcast news sources used about 1074M words of text, and was interpolated with a general English LM in a ratio of 9:1. A BC component, comprising only of the transcriptions for the 19 hours of BC data was trained on 0.24M words. Finally an additional component built using web-data from Phoenix TV (PHX) [1] was built which contained 64M words after ensuring there was no overlap with any of the test or unsupervised data. This data was found to be suitable for both BN and BC transcription. Word-based trigram and 4-gram LMs were then trained for each source and interpolated and merged to form the final model.

3.3. Test datasets

The test sets used to evaluate the systems are `bnmdev06` and `bcmdev05`. `bnmdev06` comprises 3.6 hours of data taken from a range of BN sources. It includes some of the standard existing test sets described in [10], `dev04f`, `eval03m` and `eval04`. In addition a more recent set of 4 shows taken from July-October 2006 were also included. The test data for BC, bcmdev05, comprises 2.5 hours of data taken from 5 BC shows broadcast during March 2005.

3.4. Unsupervised training dataset

The dataset used for unsupervised experiments, after segmentation and Mandarin show detection, consists of 318.1 hours of data: 185.6 hours of BN and 132.5 hours of BC. For these data, LDC provided (quick) manual transcriptions although these were not used in pure unsupervised training experiments, but some of these transcripts were used in the experiments on directed manual transcription and the contrasts with supervised training. Including the non-Mandarin shows, there was a total of 327.6 hours of data: 187.6 hours of BN and 131.1 hours of BC.

3.5. Data Selection

The baseline models, **S0**, were used to generate the initial transcriptions with confidence scores. To examine the difference between BC

[1] The Phonenix data was kindly made available by SRI-UW-NTU

and BN data, with a similar decoding configuration, the **S0** baseline models were used to recognise `bnmdev06` and `bcmdev05`. This should give an indication of expected accuracy of the automatic transcriptions for the unsupervised data. It has been found that the performance on `bcmdev05` had approximately double the error rate of `bnmdev06` [4]. Therefore, the automatic transcription of the unsupervised BC data should be significantly worse than for the BN data: this can also be demonstrated by comparing the confidence scores of BC and BN data.

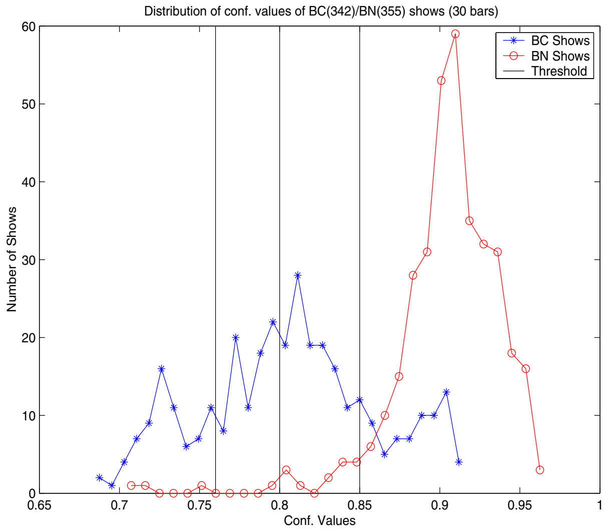

Fig. 1. Confidence score distribution of BC and BN data.

Figure 1 shows the confidence score distribution for the recognised BC and BN data. The distribution is clearly bimodal, and the BC data generally has lower confidence scores than the BN data. As described above, show based confidence scores are used to select the data to be manually transcribed. Table 1 gives the different thresholds used in the experiments[2].

Conf.	Man. Trans. (hr)			Auto. Trans. (hr)		
Thresh.	BC	BN	All	BC	BN	All
0.00	0.0	0.0	0.0	132.5	185.6	318.1
0.76	28.3	0.5	28.8	103.7	184.9	288.7
0.80	58.1	0.8	58.9	73.3	184.6	257.9
0.85	101.9	5.4	107.3	28.8	180.1	208.9
1.00	131.1	187.6	327.6	0.0	0.0	0.0

Table 1. Summary of data selection

The different thresholds for manual transcription, and the corresponding amounts of manually and automatically transcribed data are shown in Table 1. Note that the contrasts with no manually transcribed data (threshold=0.0) and all added data is manually transcribed (threshold=1.0) are also shown.

From Table 1, it can be seen that for the selection thresholds used, BC data dominates. This is consistent with Figure 1 as BC data normally have lower confidence scores. Therefore, confidence score based selection implicitly performed BC/BN show selection.

4. RECOGNITION RESULTS

All recognition results in this section use either ML or MPE trained acoustic models with unadapted single pass decoding.

4.1. Unsupervised Acoustic Model Training

Table 2 shows the performance of various systems using ML and MPE training. The baseline language model was used in decoding.

Conf.	`bnmdev06`		`bcmdev05`	
Thresh.	ML	MPE	ML	MPE
S0	15.1	13.6	29.3	25.4
0.00	13.8	12.0	27.7	24.8
0.80	13.5	11.7	26.9	23.6
1.00	12.8	10.5	26.1	21.8
Rand.	13.5	11.6	27.2	23.8

Table 2. %CER of Unsupervised and directed manual transcriptions. `Rand` indicates that 58.5 hours of data was randomly selected for manual transcription.

From Table 2 it can be seen that, with ML training, complete use of automatic transcriptions (thresh=0) led to similar absolute reductions in CER for both BN and BC data: 1.3% for `bnmdev06` and 1.6% for `bcmdev05` compared to the **S0** performance. In contrast, for MPE training, the gain on `bcmdev05` was only 0.6%, which is less than half of that on `bnmdev06` (1.6%). This demonstrates that the higher error rate for BC data can significantly affect the performance of unsupervised discriminative training. For MPE, the proportion of the reduction in CER when using full manual transcription (thresh=1.0), on BN is 57%, while on BC only 17%. If manual transcripts for 58.9 hours are used (thresh=0.8) i.e. 18% of the data, then the proportions become 61% for BN and 50% for BC. Note that the performance on BC data outperforms the use of a randomly selected subset of data of approximately the same size (`Rand.` in Table 2), while the performance on BN data is very similar with either the directed set or the random set which shows that adding manual transcriptions for lower confidence shows is preferable to random selection.

Conf.	`bnmdev06`		`bcmdev05`	
Thresh.	ML	MPE	ML	MPE
0.00	13.8	12.0	27.7	24.8
0.76	13.6	11.8	27.5	24.3
0.80	13.5	11.7	26.9	23.6
0.85	13.3	11.5	26.7	22.7
1.00	12.8	10.5	26.1	21.8

Table 3. %CER of Unsupervised and directed manual transcriptions with varying levels of directed manual transcription.

Table 3 compares the use of thresholds of 0.76 (28.8 hours) and 0.85 (107.3 hours) with directed manual transcription in addition to the 58.9 hours. It can be seen that the reductions in CER increase as more data is added at the cost of producing the additional transcripts. Taking systems with complete automatic transcription as the baseline, adding 9% manual transcripts yields 17% of the CER reduction from using the full set of manual transcripts for MPE models on BC; 18% selected data yields 40% and 33% manual yields 70%. Note

[2] One show was automatically detected as being primarily non-Mandarin speech. This show was removed from the data used for automatic transcription.

that since the automatic selection favours BC data, the performance on BN data does not increase so quickly.

4.2. Alternative Training Strategies

S0 + (hrs)		bnmdev06		bcmdev05	
Man.	Auto.	ML	MPE	ML	MPE
0.0	0.0	15.1	13.6	29.3	25.4
0.0	257.9	13.6	12.0	27.9	24.8
0.0	318.1	13.8	12.0	27.7	24.8
58.9	0.0	14.7	12.7	26.7	23.1
58.9	257.9	13.5	11.7	26.9	23.6

Table 4. Comparison between training strategies when adding 58.9 hours of directed manual transcripts

Table 4 gives a comparison between different strategies of training using confidence score based selection with threshold 0.80. The first row gives the baseline S0 performance, and the third row use of completely unsupervised data. The second line is a show-level version of the segment-level selection strategy in [4]. For the ML systems the removal of the low-confidence shows, mainly BC data, improves the BN performance but degrades the BC performance. For the MPE systems the use of the low-confidence shows gave no difference in performance, The fourth row *only* added the data with directed manual transcription (i.e. no fully unsupervised data), and the final row adds in both the manual and automatically transcribed data. For BN data it is clear that adding the automatically transcribed data is beneficial while for BC data it is slightly better to only add in the manually transcribed data: this again illustrates the issues of unsupervised discriminative training with fairly high error rate transcriptions.

4.3. Unsupervised Language Model Training

Once the transcriptions, either automatic or manual, are generated for the audio, they can also be used for language modelling. The basic procedure is to build separate language model components for the manual and/or automatic transcriptions and then interpolate among the newly built components and the three language model components described in section 3. Table 5 shows the performance of incorporating unsupervised data with 58.9 hours of directed manual transcriptions for both acoustic and language model building.

Update	bnmdev06		bcmdev05	
	ML	MPE	ML	MPE
—	15.1	13.6	29.3	25.4
AM	13.5	11.7	26.9	23.6
AM+LM	13.4	11.6	26.9	23.4

Table 5. %CER of Unsupervised Acoustic Model and Language Model Training for the 0.80 confidence threshold system.

Comparing the numbers in Table 5 to the corresponding numbers in Table 2, adding the unsupervised data to the language model gives further small improvements of up to 0.2%. Compared to traditional unsupervised training, on bcmdev05, with 18% data manually transcribed, it was possible to obtain 43% of the MPE gain when using the complete manual transcription.

5. CONCLUSIONS

In traditional unsupervised training, automatic transcription is used for both ML and discriminative training. However, the performance of unsupervised discriminative training can be poor if the initial recognition system does not have a low enough error rate. In this paper, directed manual transcription is used to partially address the problem. Show-level confidence scores are calculated and a small number of shows with low confidence scores are selected for manual transcription. The manual transcription is then used together with the automatically transcribed data for both ML and MPE training. The performance of direct manual transcription with ML and MPE training was evaluated on a Mandarin broadcast transcription task. Experiments showed that incorporating directed manual transcription can significantly increase the reduction in error rate for MPE estimated models for BC data compared to the traditional unsupervised approach. It is shown that the confidence score data selection can outperform random data selection. With more data manually transcribed, the MPE gains on both BC and BN increase although since the selection procedure automatically chooses more BC data the performance increase on BC is more rapid. It is also shown that for the BC test data it is slightly preferable to only include the additional manually selected data in training, although for BN including all the unsupervised data is clearly beneficial. Finally it is shown that small additional improvements in performance result from including the unsupervised and directed manual transcriptions in language modelling.

6. REFERENCES

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

[2] H.Y. Chan & P.C. Woodland, "Improving Broadcast News Transcription by Lightly Supervised Discriminative Training," *Proc. ICASSP 2004*, Montreal.

[3] T. Kemp & A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," *Proc. Eurospeech'99*, Budapest.

[4] L. Wang, M.J.F. Gales, & P.C. Woodland, "Unsupervised Training for Mandarin Broadcast News and Conversation Transcription," *Proc. ICASSP 2007*, Hawaii.

[5] D. Cohn, L. Atlas, & R. Ladner, "Improving Generalization with Active Learning," *Machine Learning*, Vol. 15, pp. 201–221, 1994.

[6] T.M. Kamm & G.G.L. Meyer, "Selective Sampling of Training Data for Speech Recognition," *Proc. Human Language Technology*, San Diego, 2002.

[7] G. Riccardi & D. Hakkani-Tur, "Active and Unsupervised Learning for Automatic Speech Recognition," *Proc. Eurospeech 2003*, Geneva.

[8] D. Povey & P. C. Woodland, "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," *Proc. ICASSP 2002*, Orlando.

[9] J. Ma, S. Matsoukas, O. Kimball, & R. Schwartz, "Unsupervised Training on Large Amounts of Broadcast News Data," *Proc. ICASSP 2006*, Toulouse.

[10] R. Sinha, M.J.F. Gales, P.C. Woodland et al. "The CU-HTK Mandarin Broadcast News Transcription System," *Proc. ICASSP 2006*, Toulouse.