UNSUPERVISED TRAINING FOR MANDARIN BROADCAST NEWS AND CONVERSATION TRANSCRIPTION

L. Wang, M.J.F. Gales and P.C. Woodland

MIL, Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK
Email: {lw256, mjfg, pcw}@eng.cam.ac.uk

ABSTRACT

A significant cost in obtaining acoustic training data is the generation of accurate transcriptions. For some sources close-caption data is available. This allows the use of *lightly-supervised* training techniques. However, for some sources and languages close-caption is not available. In these cases *unsupervised* training techniques must be used. This paper examines the use of unsupervised techniques for discriminative training. In unsupervised training automatic transcriptions from a recognition system are used for training. As these transcriptions may be errorful data selection may be useful. Two forms of selection are described, one to remove non-target language shows, the other to remove segments with low confidence. Experiments were carried out on a Mandarin transcriptions task. Two types of test data were considered, Broadcast News (BN) and Broadcast Conversations (BC). Results show that the gains from unsupervised discriminative training are highly dependent on the accuracy of the automatic transcriptions.

***Index Terms*—** Speech Recognition, unsupervised learning

1. INTRODUCTION

For some tasks, such as Broadcast News (BN) transcription, audio data can be easily collected from radio and television shows. Thus, it is possible to collect thousands of hours of audio data. However to build Speech-to-Text (STT) systems, in addition to the audio data, transcriptions are required. Generating accurate manual transcriptions for this data is highly expensive. For some sources, closed-captions and television transcripts may be available. These approximate transcriptions have been successfully used in lightly-supervised training techniques [1, 2]. However, for some tasks and languages, approximate transcriptions are not available. To make use of this audio data *unsupervised training* techniques may be used.

Most previous studies of unsupervised training [3, 4, 1, 5] have examined Maximum Likelihood (ML) estimation techniques and concentrated on English BN transcription. These schemes have found that the use of iterative additions of the data and the use of confidence scores can yield improvements in performance, especially when limited amounts of manually generated transcriptions are available. The majority of state-of-the-art speech recognition systems make use of discriminative training approaches such as Maximum Mutual Information (MMI) [6] and Minimum Phone Error (MPE) [7]. Recently unsupervised discriminative training has been used with large amounts of data [8] for both English and Arabic BN transcription. It was shown that MMI-based discriminative training could be effectively used. This paper examines in detail the gains and possible limitations of applying unsupervised discriminative training techniques to Mandarin BN and Broadcast Conversation (BC) transcription.

Discriminative approaches are expected to be more sensitive to the accuracy of the transcriptions than ML-based approaches. In particular as the schemes attempt to improve the recognition of the "correct" sequence from all others, if the "correct" sequence is wrong then little improvement may be obtained. There is also the issue that if a STT system is used to generate the transcription then the competing paths tend to be "closer" to the transcription than if manual had been used. Both of these issues are examined in this paper. Unsupervised performance gains from additional data are compared to those obtained if supervised transcriptions had been used.

An additional problem is that for unsupervised training the shows will not be checked to see whether they are suitable for target application, or indeed are in the target language. For example in Mandarin broadcasts there are some shows that have significant levels of English. Rather than relying on the use of segment and word-level confidence scores to remove these large segments of data, it may be preferable to eliminate whole shows. In this work, a dual language system, Mandarin and English, is used that allows the detection of large percentages of English in shows. In addition, due to acoustic mismatches with Mandarin, it also allows the detection of large percentages of other non-Mandarin languages in shows.

2. UNSUPERVISED TRAINING

This section briefly describes the unsupervised training procedure used in this work. The general procedure is similar to the approaches used for both lightly supervised and unsupervised training [1, 2], but with the need to generate *denominator lattices* for discriminative training [6].

Segmentation: The initial stage for unsupervised training is similar to that used for recognising broadcast data. The same procedure as that described in [9] was used. First, advert removal is run. Here the arithmetic harmonic sphericity distance is used to detect repeated blocks of audio data, for example jingles or commercials. Acoustic segmentation is performed. The data is then split into wide-band and narrow-band speech. Sections of music are discarded. Finally gender detection and speaker clustering are run.

Transcription generation: Initial transcriptions are generated using good acoustic models, MPE trained in this work, and a multi-pass framework. Is this paper, the P1-P2 framework, normally used as the initial lattice generation stage in the CU multi-pass recognition

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.

framework [9], is run. The two stages are:

- **P1**: gender-independent models are used to generate initial transcriptions using a trigram language model and relatively tight beamwidths.
- **P2**: the 1-best hypothesis from the P1 stage is used to generate adaptation transforms. Here least squares linear regression and diagonal variance transforms are estimated. Using the adapted models lattices are generated using a trigram language model. These lattices are then rescored using a 4-gram language model.

In contrast to the normal numbers quoted for the P1-P2 stage, the Viterbi 1-best is used for the transcription as this is felt to give a better balance between deletions and insertions.

Denominator lattice generation: All the descriminative training schemes implemented in this paper make use of denominator lattices as a compact representation of all possible word sequences [6]. The standard approach is to use a weakened language model, normally a heavily pruned bigram or unigram. The rationale for this is that the weakened language model increases the number of confusions in the data, hence improving the generalisation of discriminative training to unseen data. In unsupervised discriminative training it serves an additional important task. If the same language model and acoustic models were used as for the transcription generation, then the best path in the numerator (transcription) and denominator must, by definition, be the same. This limits possible gains from discriminative training. The difference between the numerator and denominator 1-best performance is further increased as ML acoustic models (trained on all data including the unsupervised data) are used in the generation of the numerator lattices. In contrast adapted MPE models trained on only the manually transcribed data are used to generate the transcriptions.

Data selection: Two forms of data selection are implemented in this work. The first makes use of a dual language, Mandarin and English, system. In Mandarin BN and BC shows, English often appears. For carefully selected data, such as that used for the 2003 and 2004 NIST Mandarin BN transcription evaluations, the percentage of English data is typically in the range of 1-2%. However for some shows this percentage is significantly higher. It would be useful to select and eliminate these shows from the training set as they are unlikely to be appropriate for training Mandarin acoustic models [1]. Detection of non-Mandarin shows is performed by setting a threshold on the percentage of English word recognised for that show and on the overall show confidence. The show level confidence scores are estimated by first generating confusion networks and mapping the resulting confidence scores [10]. Though the dual language system was based on English and Mandarin, it is found to detect other non-Mandarin data such as shows containing large percentages of German. After removing "non-Mandarin" shows, further data selection is performed at the segment level. Segment confidence scores are estimated in a similar fashion to the show confidence scores. Those segments that fall below a set threshold are removed. This is similar to the approaches adopted in [2, 8].

Model training: Using the baseline ML system as the starting point, in this case a 16 component system, initially ML training is performed using the transcriptions and the average number of components per state increased to 36. Then either MMI or MPE discriminative training is performed.

3. RESULTS

3.1. Experimental Set-Up

The baseline system, **S0**, was trained on data with manual and approximate "closed-caption" transcriptions (these will both be referred to as "manual" to clearly distinguish from the unsupervised data). This is similar to the basic system described in [9]. The data consists of 155 hours of BN data and 19 hours of BC data. In addition 10 hours of English data, randomly selected from the TDT4 English data, was used. The basic acoustic features for all the recognition system were 13 Cepstral coefficients (including energy) and their derivatives, 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 appended to the features 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 was also kept fixed at an average 36 for all systems.

All text was processed using a simple characters to word segmenter based on a longest-first match algorithm. The multi-character word-list for this consisted of about 51K words. Any Chinese character that wasn't present in the word-list was processed as an individual word. 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 giga-word released and the web download data. In addition the audio data transcriptions for the baseline system were used. Three separate LM components were built and interpolated. The first, BN, component used about 1074M "words" of text. Note the BN component was interpolated with a general English LM in a ratio of 9:1 for the interpolation weights. The BC component, comprising only of the transcriptions for the 19 hours of BC data, had 0.24M words of training data. Finally a component using web-data from Phoenix TV (PHX) [2] was built. 64M words of text were retained 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 trained for each source and interpolated.

Two test sets were used to evaluate the systems, 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 [9], dev04f, eval03m and eval04. In addition a more recent set of 4 shows taken from July-October 2006 were used. The evaluation data for BC, bcmdev05, comprises 2.5 hours of data taken from 5 BC shows during March 2005.

System	Data	Transcription	Data Sel.	Size
S0	baseline	manual	—	185 hr
S1	+subset1		—	504hr
S2a	+subset1	auto	—	503hr
S2b			CN08	408hr
S3a	+subset1	auto	—	955hr
S3b	+subset2		CN08	752hr

Table 1. The training sets for acoustic models.

[1] Though some English data is included in the standard training data [9] the percentage is relatively small. Furthermore, the accuracy of transcribing English segments of data is relatively poor.

[2] Thanks to SRI and the Nightingale team for making this data available.

For the unsupervised training experiments two additional subsets of acoustic data were used. The first subset (subset1), after segmentation and Mandarin show selection, consists of 317 hours of data, 185 hours of BN data and 132 hours of BC data. For subset1 quick transcriptions were also available, comprising 319 hours of data, 183 hours BN and 136 hours BC. This allows a contrast of the use of unsupervised training with supervised training. The second subset (subset2) has 451 hours of data, 301 hours of BN and 150 hours of BC. For this data no transcriptions were available. Table 1 shows the amount of data and the acoustic models that made use of the data.

3.2. Baseline Performance and Data Selection

LM	Interp. Weights			Perplexities	
	BN	BC	PHX	bnmdev06	bcmdev05
v0	1.00	0.00	0,.00	246.2	379.3
v1	0.51	0.17	0.32	254.7	272.4

Table 2. 4-gram perplexities of BN and BC test sets

As both BN and BC data is to be recognised, it is interesting to examine the differences between the two sets of data. Table 2 shows the perplexities when using only the BN element of the language model and the final interpolated LM with BN, BC and PHX components. It is clear that in terms of the text there is, as expected, a large difference between the BC and BN test sets. Without the BC and PHX components there is a difference of over 130 points in the perplexities between BN and BC. For all experiments in this paper, including transcription generation, the **v1** language model was used.

System	bnmdev06	bcmdev05
S0	12.4	25.0

Table 3. %CER of P1-P2 stages of baseline on bnmdev06 and bcmdev05

The baseline system, **S0**, was used to generate the initial transcriptions and confidence selection. Table 3 shows the P1-P2 performance[3] of these baseline models on the BC and BN tests sets. These should give an indication of the accuracy of the transcriptions that were generated. The performance on the BC test set, bcmdev05, has approximately twice the error rate of the BN data, bnmdev06. Thus the transcriptions for the BC data should be significantly worse than those of the BN data.

Two thresholds were used in the show level selection. The first was a show-level confidence score of 55%, the second a threshold of 20% for the percentage of English. Four four shows were detected as non-Mandarin and were listened to. Three of the shows contained large amounts of English interviews and the other show contained only songs. The amount of data removed using this show selection approach depends significantly on the care taken in selecting the sources and shows recorded. In previous work on the BN data released under the EARS program, a far larger percentage of shows were detected as English.

After detecting the "Mandarin" shows, segment level confidence scores were applied. In this case a threshold of 80% confidence (CN08) was used. The total amount of data retained is shown in table 1. For example of the 317 hours of data in subset1 95 hours were removed. An interesting aspect of this data selection was that far more BC data was removed than BN. In subset1 only 65.4 hours of BC data, approximately half, was retained. This shows that confidence scores used are reasonable as the recognition performance on the BC data is worse than that on the BN data.

3.3. Unsupervised Training

System	Trans.	Data Select.	CER (%)	
			BN	BC
S1	manual	—	22.2	42.4
S2a	auto	—	17.9	32.4
S2b		CN08	14.0	21.5

Table 4. %CER of 1-best outputs of denominator lattices versus numerator transcriptions on subset1 data

After data selection denominator lattices were generated as described in the previous section. In order for discriminative training to operate well the numerator and denominator 1-best hypotheses should be different. Table 4 shows the CER for scoring the 1-best denominator output against the transcription (numerator 1-best), for the **S1** system built using the manual transcriptions for the subset1 data and the **S2a/b** systems built using the automatic transcriptions. As expected the manual transcriptions have a greater mismatch with the denominator 1-best than the automatic transcriptions. If segment-based selection is applied, the difference becomes even larger. This shows an important play-off for unsupervised discriminative training. Selecting those segments for which there is high confidence in the transcriptions, reduces the opportunity of discriminative training to reduce the error rate.

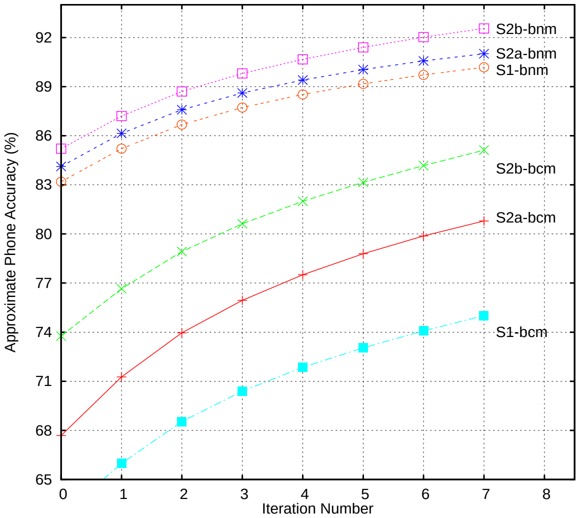

Fig. 1. Approximate phone accuracy of for MPE training iteration.

Figure 1 shows the approximate phone accuracy during MPE training. The trends correspond closely with those illustrated at the CER in table 4. For the manual transcriptions the MPE score is consistently lower than for the automatic transcriptions.

[3] All other results are based on unadapted single-pass decodes.

System	Data Select.	bnmdev06		
		ML	MMI	MPE
S0	—	15.5	14.3	13.6
S1	—	13.8	12.6	11.5
S2a	—	14.5	13.4	12.7
S2b	CN08	14.2	13.3	12.6
S3a	—	14.4	—	12.3
S3b	CN08	14.2	—	12.2

Table 5. Unadapted %CER on bnmdev06

Table 5 shows the performance of the various unadapted single-pass systems using ML, MMI and MPE training on the BN test set. A number of interesting trends can be observed. For ML training the use of segment-level data selection (CN08), shows small consistent gains over using all the data. Moreover with data selection the gains from using subset1 over the baseline (**S0**) in unsupervised ML training (**S2b**), 1.3% absolute, is over 75% of the gains using the manual transcriptions (**S1**), 1.7%. For ML training no additional gain was obtained from using subset2. For both forms of discriminative training (MMI and MPE) the gains from using segment-level data selection are reduced. Using subset1 unsupervised MPE training only obtain about 50% of the gains obtained from using the manual transcriptions in MPE training. However discriminative training still gave improvements over ML training and MPE training was consistently better than MMI training. The use of additional data, subsets2, did give small additional gains for MPE training. Overall the use of 770 hours of untranscribed data followed by data selection gave a 1.4% absolute reduction in CER.

System	Data Select.	bcmdev05		
		ML	MMI	MPE
S0	—	29.2	26.7	25.3
S1	—	26.1	23.3	21.8
S2a	—	27.7	25.9	24.7
S2b	CN08	27.9	25.8	24.8
S3a	—	27.7	—	24.8
S3b	CN08	28.0	—	24.7

Table 6. Unadapted %CER on bcmdev05

The models were then evaluated on the BC test set, bcmdev05. the results are shown in table 6. For ML training the gains over the baseline (**S0**) system from using unsupervised transcriptions with subset1, 1.5% absolute, were less than 50% of the gains obtained with the supervised transcriptions, 3.1%. There was no consistent improvement using segment level data selection for any of the training schemes. The gains from using discriminative training were much reduced. MPE unsupervised training with subset1 gave less than 20% of the gains obtained with supervised MPE training over the baseline system. Only a 0.5-0.6% absolute reduction in error rate was achieved over the **S0** system. Furthermore the use of additional data, subset2, yielded no additional gains.

4. CONCLUSIONS

This paper has examined the use of unsupervised training, in particular when combined with discriminative training. The general process for unsupervised discriminative is similar to that used in unsupervised training with ML. The main difference is that in discriminative training it is necessary to also generate denominator lattices. When generating these lattices, in addition to the standard need to weaken the LM to improve generalisation, there is the need to make the system used to generate the denominator lattices different to that used for the numerator transcriptions. The paper has also described the use of a dual language system for data selection.

The performance of unsupervised training for ML, MMI and MPE systems was evaluated on a Mandarin transcription task. For this task both BN and BC test data was used. The nature of the two sets of data was found to be different, both in terms of the word sequences used and the audio data. For baseline recognition, the system used to generate the training transcriptions, the performance on BC had about twice the error rate of BN. For both tasks, ML unsupervised training, worked well. Even for the harder BC data the gains from unsupervised ML training were about 50% of those obtained for supervised training on the same data and for BN data about 75%. In contrast for discriminative training, though gains were obtained on BN, they were only about 50% of the supervised gains. Moreover on BC data the gains were less than 20% of the supervised gains. Adding additional unsupervised data gave no additional gains in performance on BC.

From these results, there appear to be limits to the gains from unsupervised discriminative training when there is little manually transcribed data for a particular, mismatched, data type. Whether this problem can be addressed by incrementally folding in additional data will be investigated in future work.

5. REFERENCES

[1] L. Lamel, J.L. Gauvian, and G. Adda, “Lightly supervised and unsupervised acoustic model training,” *Computer Speech and Language*, 2002.

[2] H.Y. Chan and P.C. Woodland, “Improving broadcast news transcription by lightly supervised discriminative training,” in *Proc. ICASSP 2004*, May, 2004.

[3] T. Kemp and A. Waibel, “Unsupervised training of a speech recognizer: recent experiments,” in *Proc. Eurospeech 99*, Sep. 1999, pp. 2725–2728.

[4] L. Lamel, J.L. Gauvian, and G. Adda, “Unsupervised acoustic model training,” in *Proc. ICASSP*, May, 2001.

[5] F. Wessel and H. Ney, “Unsupervised training of acoustic models for large vocabulary continuous speech recognition,” *IEEE Trans. on Speech and Audio Processing*, 2005.

[6] P. C. Woodland and D. Povey, “Large scale discriminative training of hidden Markov models in speech recognition,” *Computer Speech and Language*, 2002.

[7] D. Povey and P. C. Woodland, “Minimum phone error and I-smoothing for improved discriminative training,” in *ICASSP*, 2002.

[8] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, “Unsupervised training on large amount of broadcast news data,” in *Proc. ICASSP 2006*, May, 2006, pp. 1056–1059.

[9] R. Sinha, M.J.F. Gales, P.C. Woodland, and etc., “The cu-htk mandarin broadcast news transcription system,” in *Proc. ICASSP 2006*, May, 2006, pp. 1077–1080.

[10] G. Evermann and P. C. Woodland, “Posterior probability decoding, confidence estimation and system combination,” in *Proceedings Speech Transcription Workshop*, 2000.