Lightly Supervised Acoustic Model training Using Consensus Networks

Langzhou Chen, Lori Lamel and Jean-Luc Gauvain

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
{clz,lamel,gauvain}@limsi.fr

ABSTRACT

This paper presents some recent work on using consensus networks to improve lightly supervised acoustic model training for the LIMSI Mandarin BN system. Lightly supervised acoustic model training has been attracting growing interest, since it can help to substantially reduce the development costs for speech recognition systems. Compared to supervised training with accurate transcriptions, the key problem in lightly supervised training is getting the approximate transcripts to be as close as possible to manually produced detailed ones, i.e. finding a proper way to provide the information for supervision. Previous work using a language model to provide supervision has been quite successful. This paper extends the original method presenting a new way to get the information needed for supervision during training. Studies are carried out using the TDT4 Mandarin audio corpus and associated closed-captions. After automatically recognizing the training data, the closed-captions are aligned with a consensus network derived from the hypothesized lattices. As is the case with closed-caption filtering, this method can remove speech segments whose automatic transcripts contain errors, but it can also recover errors in the hypothesis if the information is present in the lattice. Experiment results show that compared with simply training on all of the data, consensus network based lightly supervised acoustic model training based results in about a small reduction in the character error rate on the DARPA/NIST RT'03 development and evaluation data.

1. INTRODUCTION

Accurately transcribed speech is essential for acoustic model training. But accurate transcriptions are not always available, since producing them is costly. In contrast a much larger quantity of audio data may be useful for acoustic model training if the more easily available approximate transcriptions or related texts can be efficiently used. For example, for the development of our Mandarin BN system there are only about 24 hours of data with accurate transcriptions. However, an additional 120 hours of TDT4 Mandarin BN audio data are available for which there are only closed-captions (as part of the TDT4 text collection). Given the high proportion of data for which only approximate transcripts are available, the performance of the recognition system depends heavily on lightly supervised acoustic model training.

Erroneously transcribed speech data is even a problem for supervised acoustic model training. Sometimes transcription errors lead to the failure of Viterbi forced alignment of the transcript and the signal. Failure to segment is one way in which transcription errors can be removed. Pitz et al. [1] investigated different criteria to detect such transcription errors and found that manual correction of automatically detected errors could improve system performance.

Since in general closed-captions are much less precise than detailed manual transcriptions, previous work at LIMSI on light supervision used the closed-captions to provide indirect supervision via the language model [2] as opposed to trying to directly align the captions in place of reference transcriptions. Using the language model (LM) trained with the captions, a large corpus of unannotated audio data are transcribed automatically. The resulting recognition hypothesis are then used in forced alignment prior to carrying out standard EM training. While these experiments in [2] demonstrated that the framework of lightly supervised training works well, there is still some room for improvement.

A major problem with using the closed-captions for indirect supervision, is that while it reduces problems due to errors in the captions, acoustic model training assumes that the erroneous hypotheses are truth. Kemp and Waibel [3] reported that using lattice-based confidence measures to remove probable recognition errors could improve the performance of unsupervised training. Wessel and Ney [4] explored the use of confidence measures to filter automatic transcripts in order to remove portions of the data which were likely to be erroneous. An alternative is to align the closed-captions with the automatic transcriptions, and to keep only portions that agree [2]. The use of filtering with the closed-captions, which is essentially a perfect confidence measure, was found to slightly reduce the word error rate. An implicit assumption is that it is extremely unlikely that the recognizer and the closed-caption both have the same error.

Aligning the closed-captions with the automatic transcriptions detects mismatches which are potential recognition er-

rors. Some of the discrepancies however are due to errors in the captions. NIST reported the disagreement between the closed-captions and the manual transcripts to be on the order of 12% [5] on a 10 hours of TDT2 data. Removing all the speech segments with mismatches results in clean training data at the cost of losing a significant part of the training material. In fact, this lost data may be more important than the preserved data because this data contains recognition errors, while the preserved data has been recognized correctly. In comparing regions that disagree, it appears that when the captions are correct the phone sequence of the hypothesis is close to what would be expected, even though the words are wrong. When the captions and hypothesis are very different phonemically, the captions are likely to be incorrect. This implies that such regions of mismatch do not necessarily correspond to recognition errors.

For these reasons, simply discarding the data in mismatched areas is probably too simplistic. Mismatched regions are obtained by comparing only the best recognition hypotheses to the captions. A potential improvement can be gained by using a less strict criterion in which additional information is gleaned from the recognition word lattices. Many alternative word candidates are in the lattice and one of these words may agree with the caption. Therefore in this paper, the original strict agreement criterion is relaxed and all segments where the words in the caption are found on a path through the word lattice are kept. The search is carried out via a consensus network [6], a multiple alignment network in which all the word hypotheses in lattice are ordered. Consensus networks were proposed as a means to minimize the word errors instead of the overall sentence error. In this work the multiple alignment structure of the consensus network is used to easily select different word candidates according to their position.

The remainder of this paper is organized as follows. The next section provides an overview of the original lightly supervised acoustic model training method proposed in [2]. Section 3 describes the proposed lightly supervised acoustic model training method based on the consensus network and Section 4 overviews the Mandarin BN system used for experimentation. The last two sections provide some experimental results and conclusions.

2. LIGHTLY SUPERVISED TRAINING

The differences between the accurate transcriptions of the speech data and closed-captions necessitates a modified approach to acoustic model training with respect to supervised training. Some of the main differences are:

- The closed-captions do not have any indicators of non-speech events or disfluencies, such as breath noise, filler words or fragments
- The captions contain little or no time information
- The closed-captions have no explicit speaker or acoustic information, such as speaker identities and genders, speaker turns, nor indications of background noise or music.
- The closed-captions are imperfect and have transcription errors (substitutions, insertions, deletions).

Conventional HMM training relies on an alignment between the audio signal and the phone models, usually derived from a careful orthographic transcription of the speech and a good phonemic lexicon. The manual transcripts also provide time information for which words belong to which speech segment. In order to train acoustic models on audio data that only have closed-captions, the training procedure must be modified. There are two main differences. First, the audio data needs to be partitioned into homogeneous segments, automatically producing some of the missing information that is missing in the closed-captions, such as speaker, gender and bandwidth. Second, for the reasons discussed in the previous section, the closed-captions cannot be aligned directly with the audio data. One approach is to train a biased LM on the closed-captions and to use this LM in the recognizer which is used to generate the automatic transcripts. These are in turn used in the standard AM training procedure [2].

In this work, aimed at improving the LIMSI Mandarin BN system in preparation for DARPA/NIST Rich Transcription 2003 (RT'03) evaluation, a modified procedure is used from that described in [2]. The initial set of acoustic models used to decode the unannotated TDT4 data are trained on 24 hours of data from 1997 Hub4 Mandarin corpus distributed by LDC. These models are much better than those used as seed models in [2], and therefore the hypothesized transcripts are also better and there is less of a need for iteration. Since the closed-captions for the TDT4 Mandarin data match the spoken language quite well, a strongly biased LM was estimated on only the closed-caption data. Source specific LMs were used to transcribe the TDT4 training data used in the experiments reported here.

3. USING CONSENSUS NETWORKS

As mentioned above, recognition errors in the automatic transcripts of the audio training data will result in less accurate acoustic models. Removing segments for which the captions and the hypotheses disagree can throw away over 30% of the data, and in addition the kept data is representative of data that is already well recognized by the system. On the other hand, the unfiltered hypotheses potentially contain many transcription errors which can limit the performance of the models. In this work we aim to find a trade-off between these two options.

The proposed approach tries to align the closed-captions with the word lattices created by the decoding procedure, and if the closed-caption can match one of the paths trhough the word lattice, then the corresponding speech segment is

retained. The basic idea is that even if a path does not correspond to the best hypothesis, if it can survive beam pruning and at the same time match the closed-caption, it is still likely to correspond to the correct transcript of the speech. Unfortunately, exploring all possible paths in a lattice is quite costly, so this work proposes to use consensus networks to solve the problem.

A consensus network [6] is a multiple string alignment of a word lattice, which incorporates all lattice hypotheses into a single linear alignment. It can be used to minimize the word errors by selecting the word with the highest posterior probability in each position. In this work, a consensus network is generated from each word lattice, where there is one lattice per speech segment. The closed-caption is then aligned with the consensus network using a dynamic programing algorithm. For every position in the aligned closed-caption and consensus network, if the posterior probability of best candidate in the consensus network is higher than a threshold (in this work the threshold is 0.9) it is kept at this position. In this case we consider that the best candidate is probably right, even though it does not match the corresponding word in the caption. If the posterior probability is below the threshold, we look for another candidate in the same position which matches the corresponding word in closed-caption. If one is found, independent of its probability, this word is chosen as the best candidate.

Another problem arise that is specific to Chinese being a character-based language. The recognizer uses a word based lexicon, and as a result the consensus network is a word based network. So when dealing with Mandarin data, the problem of word segmentation of the closed-captions needs to be considered (see [7].

4. SYSTEM DESCRIPTION

The LIMSI Mandarin broadcast news transcription system [7] is essentially the same as that used to transcribe American English and other languages [8], with the models (lexicon, acoustic models, language models) trained for Mandarin Chinese. The overall computation time is about 10xRT for the two-step decoding procedure, including the audio partitioning process and unsupervised acoustic model adaptation. The result of the partitioning procedure is a set of speech segments with cluster, gender and telephone/wideband labels. Word recognition is performed in two steps: 1) initial hypothesis generation used for cluster-based acoustic model adaptation, 2) word lattice generation and lattice rescoring.

The acoustic training data consist of about 24 hours broadcast news data in Mandarin with accurate time-aligned transcriptions (1997 Hub4-Mandarin) and about 120 hours of data from the TDT4 corpus distributed by LDC with closed-captions. The 1997 data come from 3 sources: VOA, CCTV and KAZN-AM. The TDT4 data come from 5 sources: CTV, CNR and VOA (Mainland style); and CBS and CTS (Taiwan style). The acoustic models are sets of gender-dependent, position-dependent triphones with 11k tied states built using both MAP adaptation of SI seed models for each of wideband and telephone band speech.

Four-gram language models are obtained by interpolation of backoff n-gram language models trained on a variety of text corpora, divided in three parts. The first part consists of the text data distributed by the LDC prior to the 1997 evaluation. The second part contains additional texts (closed-captions and transcripts) from the TDT2, TDT3 and TDT4 corpora. The third part consists of additional Mainland texts from the People Daily newspaper, and two sources from Taiwan that were shared with us by BBN. A single language model was built interpolating all component models. Source-specific language models were built by chosing mixture weights using the transcripts of the development data.

The Mandarin lexicon developed and distributed by LDC for use in the Hub5 task served as the basis for our pronunciation lexicon. The original lexicon contains 44,405 items. A lightly supervised iterative procedure was used to collect the new words from new training data and resulted in a new vocabulary containing 57700 words. Pronunciations are represented using 61 phones, of which 4 symbols represent silence, filler words, and breath noises. The phone set contains 24 consonants and 11 vowels, each having 3 tones.

5. EXPERIMENTS AND RESULTS

These experiments make use of the development and evaluation data from the DARPA RT'03 benchmark evaluation, taken from the same 5 sources found in the TDT4 training data. The development data, provided by BBN, are the last show of each source from the end of December 2000.

The light supervision experiments have 3 iterations:

1. Initial gender and bandwidth dependent AMs were trained on 24 hours of 1997 Hub4 Mandarin data.
2. The initial AMs and a LM trained on the TDT4 closed-captions, were used to transcribe all of the TDT4 training data. The original 24 hours of manually transcribed data were pooled with the automatically transcribed data from TDT4. Wideband models were trained on the 1997 Hub4 data and TDT4 Mainland style sources (CTV, CNR, VOA). Narrowband models were trained on the 1997 Hub4 data and all the TDT4 sources excepting CTS, since the character error rate (CER) of CTS source is particularly high. These acoustic models were used in the LIMSI RT'03 system.
3. The AMs from the second iteration were used to retranscribe all of the TDT4 data. From these automatic transcripts three sets of models were built. One set of models is trained directly on Hub4 1997 data pooled with new 1-best transcripts of the TDT4 data as in the previous iteration. The second set of models was trained by pooling the

Models *Show*	*iter 1* *Initial*	*iter 2* *Eval'03*	*iter 3a* *1-best*	*iter 3b* *Consensus*
CNR	13.1	11.3	9.2	9.1
CTV	14.1	-	11.1	10.2
VOA	12.5	11.1	9.8	9.9
CBS	33.4	27.7	23.5	23.2
CTS	63.5	59.1	47.5	46.7

Table 1: Character error rates on the RT'03 development data.

Hub4 data with the TDT4 data using the modified hypotheses generated via the consensus network as described in Section 3. The third set of models uses the confusion network to modify the hypothesized transcripts, but keeps the original 1-best candidate if no match is found. Separate AMs were trained for the Mainland and Taiwan sources. Wideband and narrowband Mainland style models were trained on the 1997 Hub4 data plus the TDT4 Mainland sources.The Taiwan style models are trained on the 1997 Hub4 data pooled with all of the TDT4 sources.

The recognition character error rates (CER) on the development data are shown in Table 1 using source specific language models. The first and the second column give the results using the initial AM (iteration 1) and the first lightly supervised models (iteration 2). The lightly supervised models gave a large error reduction. Columns 3 and 4 give the CER of the 3rd iteration models, where 3a corresponds to the results of lightly supervised training based on 1-best hypotheses, and 3b represents the results of lightly supervised training based on the consensus network. Each successive iteration of training is seen to reduce the CER. There is a larger gain for the Taiwan sources than for the Mainland ones, which can be attributed to the fact that the initial model training data did not contain any Taiwan style sources. Since some of the development shows were included in the training of the iteration 2 models but not in the iteration 3 models, the performance of these models cannot be directly compared on this data set. A fair comparison can be made on the RT'03 evaluation data as shown in Table 2. The 1st column gives the CER of the 2nd iteration models which were used in the RT'03 eval system. The remaining columns compare results with 1-best (3a) and consensus network based training, with (3b) and without (3c) filtering. It can be seen that the consensus network based training gives a slight improvement with filtering, but without filtering has the same average performance as the 1-best.

6. CONCLUSIONS

This paper has reviewed our recent work in lightly supervised AM training using Mandarin broadcast news data and has presented a method based on consensus networks to select candidate words from the recognition word lattices. Word hypotheses are selected according to their posterior probability in the lattice and the closed-captions, instead of just using the 1best hypotheses as the automatic transcriptions for lightly supervised training. This method gets more audio training data than direct filtering with the closed-captions as proposed in [2] and at the same time creates more accurate transcriptions than without filtering. The experimental results show the gains achieved by lightly supervised acoustic training.

Models *Source*	*iter 2* *Eval03*	*iter 3a* *1-best*	*iter 3b* *Cons.(filt)*	*iter 3c* *Cons.(unfilt)*
CNR	6.1	5.5	5.4	5.6
CTV	8.0	7.2	7.1	7.4
VOA	11.6	11.8	11.5	11.5
CBS	24.5	24.1	23.1	23.3
CTS	54.8	50.5	49.8	51.1
avg	21.7	20.4	20.0	20.4

Table 2: CER on the RT'03 evaluation data.

7. ACKNOWLEDGEMENT

We are grateful to colleagues at BBN for sharing valuable resources and for the fruitful exchanges we had during system development.

REFERENCES

[1] M. Pitz, S. Molau, R. Schluter and H. Ney, "Automatic Transcription Verification of Broadcast News and Similar Speech Corpora," *Proc. 1999 DARPA Broadcast News Workshop*, 157-159, Herndon Va February-March 1999.

[2] L. Lamel, J.L. Gauvain and G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training *Computer, Speech and Language*, **16**(1):115-229, January 2002.

[3] T. Kemp and A. Waibel "Unsupervised Training of a Speech Recognizer: Recent Experiments," *Proc. ISCA Eurospeech'99*, 2725-2728, Budapest, September 1999.

[4] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *Proc. IEEE Automatic Speech Recognition and Understanding Workshop, ASRU'01*, Madonna di Campiglio, December 2001.

[5] J. Garofolo, C. Auzanne, E. Voorhees, and W. Fisher, "1999 TREC-8 Spoken Document Retrieval Track Overview and Results," *Proc. 8th Text Retrieval Conference TREC-8*, November 1999.

[6] L. Mangu, E. Brill and A. Stolke, "Finding Consensus Among Words: Lattice-Based Word Error Minimization," *Proc. ISCA EuroSpeech'99*, 495-498, Budapest, September 1999.

[7] L. Chen, L. Lamel and J.L. Gauvain, "Transcribing Mandarin Broadcast News" *Proc. IEEE Automatic Speech Recognition and Understanding Workshop, ASRU'03*, Virgin Islands, December 2003.

[8] J.L. Gauvain, L. Lamel and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, **37**(1-2):89-108, May 2002.