Transcribing Mandarin Broadcast News

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,gauvain,lamel}@limsi.fr

ABSTRACT

This paper describes improvements to the LIMSI broadcast news transcription system for the Mandarin language in preparation for the DARPA/NIST Rich Transcription 2003 (RT'03) evaluation. The transcription system has been substantially updated to deal with the varied acoustic and linguistic characteristics of the RT'03 test conditions. The major improvements come from the use of lightly supervised acoustic model training in order to benefit from unannotated audio data, the use of source specific language models, and MDI adaptation to tune the language models for sources with limited amounts of training data. The character error rate on the development data has been reduced from 34.5% with the baseline system to 22.6% with the evaluation system.

1. INTRODUCTION

In this paper improvements to the LIMSI broadcast news transcription system for the Mandarin language [2] which was evaluated in the NIST RT'03 evaluation are described. The acoustic models in the baseline system were trained on about 24 hours of data from the 1997 Hub4 Mandarin corpus available via LDC (LDC98S73). The language models were trained on the transcriptions of the audio training data and Mandarin Chinese News Corpus containing about 186 million characters. This system was evaluated on the 1997 NIST Hub4 Mandarin evaluation data containing 1h of speech from the same sources as the training data [10], and had a character error rate of 18.1%.

The RT'03 Mandarin broadcast news development and test data are part of the TDT4 Mandarin BN audio collection, and come from five different sources from Mainland China (CNR, CTV and VOA) and from Taiwan (Central Broadcasting Station (CBS) and CTS). Since only two of these sources (VOA and CTV) are also in the 1997 Hub4 Mandarin corpus, there is a mismatch between the acoustic training and test data. An additional acoustic mismatch is due to the audio signal quality, as the data from the Taiwan sources was collected in a compressed format over the Internet. There are other challenges arising from the different linguistic natures of the data from Mainland and Taiwan as well as different accents. As all of the text sources available from the LDC being from Mainland China, additional text data from Taiwan were kindly shared with us by BBN.

Although the TDT4 corpus has not been manually transcribed for acoustic model training, quite a bit of raw audio data are available, along with closed-captions or approximate transcripts. In order to make use of this data lightly supervised acoustic model training [7] was carried out, in which our existing Mandarin recognizer was used to automatically transcribe the audio data using a biased language model, i.e., a language model trained on the closed captions for the particular show covering the data epoch.

The remainder of the paper is as follows. The next section gives an overview of the speech recognizer, highlighting some specificities for the Mandarin system. The following three sections describe in more detail the acoustic and language models, and the pronunciation lexicon. This is followed by a summary of the experimental results on the development data and for the official evaluation.

2. SYSTEM OVERVIEW

The LIMSI Mandarin broadcast news transcription system is essentially the same as that used to transcribe American English and other languages, with 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 audio partitioning procedure (segmentation and labeling) is identical to the one used in the LIMSI American English system, however, for the Taiwan sources the partitioner the speech-in-noise GMM is replaced by a GMM trained on a portion of the TDT4 data from these sources. The partitioning procedure first detects and rejects non-speech segments using GMMs. Then an iterative maximum likelihood segmentation/clustering procedure is then applied to the speech segments using GMMs and an agglomerative clustering algorithm [6]. The procedure is controlled by 3 parameters: the minimum cluster size (10s), the maximum log-likelihood loss for a merge, and the segment boundary penalty. The algorithm stops when no merge is possible and the segment boundaries are refined (within a 1s interval) so as to locate the segment boundaries within

silence portions thus avoiding cutting words. Speaker-independent GMMs corresponding to wideband speech and telephone speech (each with 64 Gaussians) are then used to label telephone segments. This is followed by segment-based gender identification, using 2 sets of GMMs with 64 Gaussians (one for each bandwidth). The result of the partitioning process is a set of speech segments with cluster, gender and telephone/wideband labels.

The LIMSI BN speech recognizer [6] uses 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.5kHz for telephone data) every 10ms. For each 30ms frame the Mel scale power spectrum is computed, and the cubic root taken followed by an inverse Fourier transform. Then LPC-based cepstrum coefficients are computed. The cepstral coefficients are normalized on a segment-cluster basis using cepstral mean removal and variance normalization. Thus each cepstral coefficient for each cluster has a zero mean and unity variance. The 39-component acoustic feature vector consists of 12 cepstrum coefficients and the log energy, along with the first and second order derivatives.

Each phone model is a tied-state left-to-right CDHMM with Gaussian mixtures. The triphone-based context-dependent phone models are word-independent but position-dependent. The tied states are obtained by means of a decision tree. Word recognition is performed in two steps: 1) initial hypothesis generation , 2) word lattice generation and lattice rescoring. The computation time is about 1.4xRT for the first pass and 8.4xRT for the second pass.

Step 1: Initial Hypothesis Generation: This step generates lattices and initial hypotheses which are then used for cluster-based acoustic model adaptation. This is done via one pass cross-word trigram decoding with gender-specific sets of position-dependent triphones (5500 tied states) and a trigram language model (8M trigrams and 8M bigrams). Band-limited acoustic models are used for the telephone speech segments. The trigram lattices are rescored with a 4-gram language models.

Step 2: Word Lattice Generation Unsupervised acoustic model adaptation is performed for each segment cluster using the MLLR technique [8]. A word lattice is generated for each segment using a bigram LM and position-dependent triphones with 11500 tied states (16 Gaussians). The 2-gram word lattice which is then expanded with a 4-gram LM. The posterior probabilities of the lattice edges are estimated using the forward-backward algorithm. The 4-gram lattices are converted to a confusion network with posterior probabilities by iteratively merging lattice vertices and splitting lattices edges until a linear graph is obtained. This procedure gives comparable results to the edge clustering algorithm proposed in [9]. The words with the highest posterior in each confusion set are hypothesized.

3. ACOUSTIC MODELING

The acoustic models were trained on about 27 hours of 1997 Hub4-Mandarin training data with accurate time-aligned transcriptions and about 120 hours of data from the TDT4 corpus distributed by the LDC with closed-captions. The baseline models were trained on only the manually transcribed Hub4 Mandarin broadcasts recorded from 3 sources: Voice of America (VOA), People's Republic of China Television (CCTV) International News programs and Commercial radio based in Los Angeles (KAZN-AM). In order to be robust with respect to the varied acoustic conditions, the acoustic models were trained on all data types: clean speech, speech in the presence of background noise or music, or transmitted over noisy channels. Although the baseline acoustic models used in [2] were only wideband and gender-independent, gender-dependent and bandlimited models were also trained on the same manually transcribed data.

Since time-aligned transcripts are not available, the TDT4 data from the Mainland China sources (CNR (China National Radio, Beijing), CTV and VOA) and the CBS (Central Broadcasting Station) Taiwan source were transcribed with the baseline recognizer using acoustic models estimated on the manually transcribed Hub4 Mandarin data and with source-specific language models estimated on the TDT4 closed captions for each source.[1] Wideband models were trained by pooling the Hub4 Mandarin data with the TDT4 data from Mainland China. Bandlimited models were trained on the same sources pooled with the Taiwan CBS data. Twenty of the CBS shows (6 hours) were manually segmented in order to roughly align the closed-captions. These 20 shows were added to the pooled data. The estimated character error rate on a selected portions from the CBS shows (about 2 hours) used for lightly supervised training is about 8%.

The acoustic models are sets of position-dependent triphones with tied states obtained using a divisive decision tree based clustering algorithm [6]. The set of 230 questions used concern the phone position, the distinctive features (and identities) of the phone and the neighboring phones, as well as specific questions about the base phone and tones for vowels. Two sets of gender-dependent acoustic models were built using both MAP adaptation of SI seed models for each of wideband and telephone band speech. The acoustic models used in the first decoding pass cover 5500 contexts, with 5500 tied-states with 16 Gaussians per state. Larger models are used in the second decoding pass, covering about 21k triphone contexts with 11500 tied-states, and a total of about 180k Gaussians.

[1] Due to the poor acoustic quality and the corresponding high error rate on the CTS (Chinese Television Service, Taiwan) data, these shows were not used for lightly supervised acoustic model training.

	Vocabulary size	*Average number of characters per word in TDT4*		
		Mainland sources	*Taiwan sources*	*Mainland+Taiwan sources*
iteration 1	50588	1.65	1.61	1.63
iteration 2	55283	1.68	1.67	1.68
iteration 3	57215	1.70	1.69	1.70
iteration 4	57700	1.71	1.70	1.71

Table 1: Vocabulary size and average word length in characters resulting from iterative word segmentation and new word collection using the TDT4 corpus.

4. LANGUAGE MODELING

N-gram language models are obtained by interpolation of backoff n-gram language models trained on a variety of text corpora which are 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 (Central Daily News and Chinese Television Service transcripts that were shared with us by BBN. The following text corpora were used for language model training:

Text resources available from LDC:

- TDT2, TDT3, TDT4 Mandarin transcripts (10.2M characters)
- People Daily newspaper 1991-1996 (85M characters)
- China Radio transcripts 1994-1996 (87M characters)
- Xinhua news 1994-1996 (22M characters)
- Acoustic training transcripts (0.43M characters)

Text resources shared by BBN :

- People Daily newspaper 1997,1999,2000 (39M characters)
- Central Daily News text 1997-2000 (61M characters)
- CTS transcripts 1997-2000 (14M characters)

All of the texts were normalized in a homogeneous manner for language model training. The following processing steps were carried out. First, formatting commands, unnecessary titles and symbols were removed from training data, then special symbols such as punctuation markers are processed according to the the way they are expressed in Chinese speech. After text normalization, the training data consists of clean Chinese character streams which can be used to train character based LMs directly [2].

The recognition vocabulary used in this work contains both words and characters, so it is necessary to segment the character stream into words. A lightly supervised iterative procedure consisting of word segmentation and new word collection is used to simultaneously define the recognition vocabulary and segment the character stream. Given an initial word list and character streams without word boundaries, word segmentation and new word collection creates an extended word list containing new words from new training data and adding word boundaries to the training texts. Our initial word list covering the first part of the training texts (i.e. those available for the 1997 evaluation) contains 50588 items [2]. Word segmentation makes use of the maximum match method which is widely used for Chinese word segmentation. It matches the text in a sentence with the longest item in the vocabulary list, so as to determine a complete segmentation of the sentence.

The procedure is as follows:

1. *Initialization:* Start with current word list and normalized texts.
2. *Word segmentation:* Using to current word list, segment the normalized texts (in this case the TDT2, TDT3 and TDT4 Mandarin transcripts) using the maximum match method.
3. *New word collection:* New word collection is based on the result of word segmentation. All neighboring items that satisfy the following conditions are selected as candidates for new words:
   1) the frequency of neighboring items is larger than a threshold; 2) the mutual information of neighboring items is higher than a threshold; 3) the neighboring items are single characters or word fragments (undeterminated words).
   Whether or not the new word candidates satisfying the 3 conditions are words or word-fragments is decided manually. A word candidate can be kept as a new word or as an allowable fragment (meaning that it is not a real word, but it can be used as part of a new word). Filtering of new words is the only manual step in this procedure.
4. Update the current word list and go to step 2.

This procedure was carried out iteratively and resulted in a new vocabulary list containing 57700 words as well as

the segmentation of the training data into words according to the new list. The resulting word list includes all frequent characters (there are about 8000 frequent characters in Mandarin) so there are essentially no OOV items.

Table 1 shows the results of word segmentation and new word collection. It can be seen that using the initial vocabulary, the average number of characters per word is larger for the Mainland data than for the Taiwan data. This can be expected since the initial vocabulary was determined using only text data from Mainland sources, so the resulting word segmentation for the Taiwan sources contains more single characters than are observed for the Mainland sources. After carrying out the word segmentation and new word collection procedure, new words from the Taiwan sources are selected. After a few iterations the average characters per word is the same in the Taiwan and Mainland sources.

The additional training data provided by BBN was only used for LM training and not for new word collection.

Different component LMs were trained on the text sources listed above and interpolated to form show-specific language models. For the Mainland shows (CNR, CTV and VOA), the mixture weights were chosen using the transcriptions of the dev03 data. For the Taiwan sources (CBS, CTS) for which there is less data, MDI adaptation was used to tune the LM to each source. First a common LM was trained on the available text data. Then, using the TDT4 CTS and CBS closed captions as adaptive data, MDI adaptation was carried out to create show dependent LMs [3].

The development LMs were optimized using these text sources, predating the the dev03 data which were taken from the 2nd half of December 2000. The texts from the end of December and January were then included in the LM training data for the evaluation system.

The interpolation coefficients were chosen in order to minimize the perplexity a set of five dev03 shows and transcripts (one from each source) shared by BBN. The weight of the audio transcript component was set to 0.1 in all language models.

Because the amount of text data from Taiwan is quite small, MDI (Minimum Discrimination Information) adaptation [4] was used to train the Taiwan style LM more efficiently. MDI adaptation can be expressed as follows. Given a background model $P_b(h, w)$ and a adaptive corpus A, we want to find a model $P(h, w)$ satisfying a set of linear constraints for which the Kullback-Leibler distance between $P(h, w)$ and $P_b(h, w)$ is minimized. The MDI model is trained by using the GIS (Generalized Iterative Scaling) algorithm. In these experiments only the simplified MDI adaptation was carried, i.e. only the unigram model is considered, and only one iteration is performed [3]. The adaptive data consist of the TDT4 transcripts for CTS (0.66M characters) and CBS (0.46M characters).

Table 2 gives the word perplexities measure on the dev03 data for different language model configurations. The perplexities in first column were obtained with a common LM trained on all of the available text sources. The perplexity of the Taiwanese data is at least twice that of the Mainland data. The source specific LMs shown in the second column consistently reduce the perplexity for all sources, with the largest gains on the Taiwan data which are less well represented in the training corpus and the smallest gain for the VOA data. MDI adaptation gives an additional perplexity reduction of over 5% relative to the show-specific LMs.

Show	*TDT LM*	*Source LMs*	*MDI-adapt*
CTV	191	167	-
CNR	248	204	-
VOA	274	249	-
CBS	508	412	390
CTS	623	495	460
Avg.	351	282	-

Table 2: Language model word perplexities on the dev03 data.

5. RECOGNITION LEXICON

The Mandarin lexicon developed and distributed by LDC for use in the Hub5 LVCSR (Large Vocabulary Conversational Speech Recognition) task served as the basis for our pronunciation lexicon, with some modifications to the descriptive phone symbol set and additional lexical entries.[2] Pronunciations are represented using 61 phones, of which 4 symbols stand for for silence, filler words, and breath noises. The phone set contains 24 consonants and 11 vowels, where each vowel can have one of 3 tones. This is a simplified representation of tone where the 5 tones for vowels are collapsed into three: flat (tones 1 and 5), rising (tones 2 and 3), and falling (tone 4). A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. The 57k vocabulary contains 57707 words with 59152 phone transcriptions, so only about 2% of the entries have alternate pronunciations.

6. EXPERIMENTAL RESULTS

Experimental results are reported for the LIMSI Mandarin Chinese broadcast news transcription system on the development and evaluation data from the DARPA RT'03 benchmark evaluation. Table 3 summarizes the results on the development data for 6 configurations. The first column gives the results of our starting 3-pass system, which used a single set of speaker-independent, wideband acoustic models (trained on the 24 hours of Hub4 Mandarin data) and a language model trained on the Hub4 Mandarin 1997 texts. The character error rates are seen to be extremely

[2] The LDC lexicon contains a total of 44,405 words with phonemic transcriptions, tone markers (5 levels) and additional information on the morphology and frequency of occurrence in the Xinhua newswire texts and the Hub5 CallHome corpus.

Show	*Initial 3-pass decoding* SI	*2-pass decoding*					
		Common TDT4 LM		*Source LMs + addl texts*			
		SI	*GD+wb/nb*	*SI*	*GD+wb/nb*	*TDT4 AMs*	*Eval03*
CTV	17.3	14.5	14.1	13.4	12.8	9.7	8.0
CNR	16.2	14.1	13.1	13.1	10.9	9.8	6.1
VOA	15.0	13.3	12.5	12.5	11.9	10.8	11.6
CBS	43.2	33.4	33.4	30.4	29.5	24.1	24.5
CTS	75.9	69.1	63.5	65.6	59.4	52.8	54.8
Avg.	34.5	30.1	28.3	28.0	25.8	22.6	21.7

Table 3: Comparison of recognition character error rates (CER) on the dev03 data for different system configurations. The 2-pass decoding results also make use of a modified partitioner for the CBS and CTS sources from Taiwan. The rightmost column reports the results on the Eval'03 data using the same TDT4 gender and bandwidth dependent acoustic models and source-specific language models trained on TDT4 data through the month of January.

high for the two Taiwan sources. The remaining entries all use a 2-pass decoding strategy and a modified partitioner for the Taiwan sources. The use of an updated language model (2nd and 3rd columns) with components trained on the TDT4 texts reduces the CER by over 10%, with an error reduction of 18% when gender-dependent and bandwidth dependent acoustic models are also used.

The next 3 columns give results with different acoustic models and source-specific language models. The error reduction is about 7% relative with speaker-independent models and 9% with gender and bandwidth dependent models. Training on the additional audio data in a lightly supervised manner is seen to significantly reduce the CER. The last column uses acoustic models trained on the 24 hours of Hub4 Mandarin manually annotated acoustic training pooled with the automatically determined transcripts of the TDT4 data. As a reminder the wideband models are trained only on the Mainland sources and the narrowband models are trained on bandlimited Mainland and the Taiwan CBS data. The use of narrowband models is seen to be particularly important for the data from Taiwan which is bandlimited. The CER on the Taiwan data still remains quite high, particularly for the CTS source which has been compressed for transmission over the Internet. Previous experiments with recognizing compressed data indicate that the best performance is obtained by using matching bandwidth models, and that training models on compressed data does not improve performance [1].

The rightmost column of the table gives the results on the RT03 source specific LMs trained on TDT4 data through the month of January. The difficulties of the test shows are generally in line with the development ones, however the CNR show seems to be somewhat easier.

7. CONCLUSIONS

In this paper we have summarized recent improvements to the LIMSI broadcast news transcription system for the Mandarin language. One of the challenges was to deal with a variety of audio sources, and with sources that were not represented in the manually annotated training data. The test data come from 5 audio sources, from Mainland China, from Taiwan, and from Voice of America. Large differences were observed in the linguistic and acoustic characteristics, as well as the dialect and pronunciations. These differences were partially compensated for by using adapted acoustic models and source-specific language models, but it was not possible to compensate for poor audio quality of the CTS source due to the high compression rate.

Lightly supervised acoustic model training was found to be quite effective, giving relative error reductions of over 10% on all the audio sources. This work has demonstrated that if fairly accurate closed captions or manual fast transcriptions are available, even without time markers, they can be successfully used to train language models which can then be used to generate automatic transcripts of the audio data for acoustic model training.

8. ACKNOWLEDGEMENT

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

REFERENCES

[1] C. Barras, L. Lamel and J.L. Gauvain, "Automatic Transcription of Compressed Broadcast Audio," *Proc. IEEE ICASSP'01*, **I**:265-268, Salt Lake City, UT, May 2001.

[2] L. Chen, L. Lamel and J.L. Gauvain, "Broadcast News Transcription in Mandarin," *Proc. ICSLP'2000*, Beijing, 1015-1018, October 2000.

[3] L. Chen, L. Lamel, J.L. Gauvain and G. Adda, "Unsupervised Language Model Adaptation for Broadcast News," *Proc. IEEE ICASSP'03*, Hong Kong, April 2003.

[4] M. Federico, "Efficient Language Model Adaptation through MDI Estimation," *Proc. ISCA EuroSpeech'99*, 1583-1586, Budapest, September 1999.

[5] J.L. Gauvain L. Lamel, "Fast Decoding for Indexation of Broadcast Data" *Proc. ICSLP'00*, 794-797, Beijing, October 2000.

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

[7] 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.

[8] C.J. Legetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**:171-185, 1995.

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

[10] D.S. Pallett, J.G. Fiscus, A. Martin and M.A. Przybocki, "1997 Broadcast News Benchmark Test Results: English and Non-English," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 5-12, Landsdowne, VA, February 1998.