ADVANCES IN MANDARIN BROADCAST SPEECH RECOGNITION

Mei-Yuh Hwang[1], *Wen Wang*[2], *Xin Lei*[1], *Jing Zheng*[2], *Ozgur Cetin*[3], *Gang Peng*[1]

[1]Univ. of Washington, Dept. of Electrical Engineering, Seattle, WA 98195 USA
[2]SRI International, Menlo Park, CA, 94025 USA
[3]International Computer Science Institute, Berkeley, CA, 94704 USA

{mhwang,leixin,gpeng}@ee.washington.edu, {wwang,zj}@speech.sri.com,
ocetin@icsi.berkeley.edu

ABSTRACT

We describe our continuing efforts to improve the UW-SRI-ICSI Mandarin broadcast speech recognizer. This includes increasing acoustic and text training data, adding discriminative features, incorporating frame-level discriminative training criterion, multiple-pass acoustic model (AM) cross adaptation, language model (LM) genre adaptation and system combination. The net effect without LM adaptation was a 24%-64% relative reduction in character error rates (CERs) on a variety of test sets. In addition, LM adaptation gave us another 6% of relative CER reduction on broadcast conversations.

Index Terms— Mandarin, character error rates, discriminative features, discriminative training, LM adaptation.

1. INTRODUCTION

In previous work [1], we built a competitive speech recognizer for Mandarin broadcast news, which achieved 6%-16% of CER on various test sets. This paper describes our recent improvements along a number of dimensions, from front end to acoustic models, to language models, and to system decoding architecture which takes advantage of AM cross adaptation and system combination. We will also present the effect of LM adaptation for the genre that does not have enough LM training data.

In Section 2, we first summarize the data used in training, development, and testing. Then the description of our development work starts with lexicon and language models in Section 3, followed by Section 4 describing acoustic modeling with a novel combination of discriminative MLP features, minimum phone error (MPE) [2] model training, and feature-based MPE (fMPE) transform [3]. Section 5 illustrates our decoding structure and how cross-adaptation and confusion network combination (CNC) [4] are used in the overall system. Finally in Section 6, we present our recent attempt in adapting the language models for conversational speech. Section 7 summarizes our improvements and discuss further refinement in the future.

2. TRAINING DATA

2.1. Acoustic data

Our acoustic data are from various LDC Mandarin corpora. Table 1 lists the sources of data used to train our AMs. In contrast with [1] where only 97 hours of speech were used for training, there are in total 465 hours of data: 313 hours of broadcast news (BN) and 152 hours of broadcast conversations (BC). BC is a genre defined to include broadcasts characterized by spontaneous conversations, such as talk shows and call-in shows.

Data	Years	Programs	BN	BC
hub4	1997	cctv, voa, kaznAM	29.6	
tdt4	2000-2001	cctv, voa, cnr	89.1	
gale	2004-2006	cctv, voa, phoenix	194.6	152.3
Total			313.3	152.3

Table 1. Acoustic training data, in hours.

GALE (Global Autonomous Language Exploitation) is a new DARPA-sponsored program initiated in 2006, aiming at translation into English text. We used all the GALE Year-1 Q1 and Q2 acoustic data that come with human quick transcriptions. In the first year of the program, there were much more BN data than BC data. During development, we focused on the broadcast news genre because we believe a fundamental improvement in BN will also result in better BC performance. On the other hand, our LM adaptation was focused on adapting the well-trained BN LMs to BC-genre speech.

Our development data included the 2004 development set for the EARS program, dev04, and the evaluation set for that year, eval04. After we finished tuning our system based on dev04 and eval04 we then applied the system to the GALE extended dryrun evaluation set (ext06) and the BC development set defined by Cambridge University (dev05bc). These data sets also contain some non-Mandarin speech. Naturally, these sets are excluded from either AM or LM training.

2.2. Text data

Table 3 lists all the text data used in LM training and development. The previous system [1] used a subset of the top five corpora, in the

Data	Years	Programs	BN	BC
dev04	2003	cctv	0.5	
eval04	2004	cctv, ntdtv, rfa	1.0	
ext06	2005	phoenix	1.0	
dev05bc	2005	phoenix, voa		2.7

Table 2. Acoustic development (dev04, eval04) and evaluation data (ext06, dev05bc), in hours.

Data	BN	BC
(1) tdt+	17.7M	
(2) gale	3M	2.7M
(3) Giga-cna	451.4M	
(4) Giga-xin	260.9M	
(5) Giga-zbn	15.8M	
(6) NTU-Web	95.5M	2.1M
(7) CTS		159M
dev06	34.1K	

Table 3. LM training and development text data, in number of words.

total of 420M words. Tdt+ includes TDT4, Hub4, and MTC corpora. Our improved system used all the top six corpora to separately train six general LMs, to be interpolated to obtain the single static LM for general dictation.

The GALE text data, in the second row, include all the transcriptions of the GALE acoustic data listed in Table 1, plus the GALE web transcription (closed-caption like). These web transcriptions are more similar to speech test data, as they correspond to real speech rather than written articles.

The Gigaword corpus contains articles from three newswire and newspaper sources: Central News Agency from Taiwan, Xinhua newspaper from China, and Zaobao from Singapore.

National Taiwan University (NTU) downloaded news articles and conversation transcriptions from CCTV, PHOENIX, and VOA web sites (dated before February 2006) to cover some of the sources missing from the GALE data. These data do not necessarily correspond to speech. Yet they are more like GALE data than the Gigaword corpus, as they are from the same broadcast sources, rather than from newswire articles.

In addition, the EARS Mandarin conversational telephone speech (CTS) LM training data described in [5] were added later for BC LM adaptation. All together, there were more than 1G words of training text.

we designate the GALE 2006 BN development set (dev06) as our LM tuning set, for the linear interpolation of the above six LMs. Dev06 is a superset of dev04 and eval04. It also contains the 2003 NIST Rich Transcription BN evaluation set and some new data from the GALE Year-1 BN audio transcript release.

Note that in choosing the LM training data, all text data from the same months as contained in dev06 were excluded, to avoid memorization of the test data. However, because we had little BC data, we only excluded text data within a one-week window centered on dates covered by dev05bc.

3. LANGUAGE MODELING

3.1. Improved lexicon

To do Chinese word segmentation, we started from the 49K-word unigram LM trained in [1], and then added a few thousand Chinese words from the LDC lexicon and another few thousand names automatically identified in the TDT4 corpus. The added new words were given a constant unigram probability. We then used this expanded unigram to perform maximum likelihood word segmentation on all text data. The most frequent 60K words in the training text were then chosen as our decoding vocabulary. This vocabulary included 1760 English words.

3.2. Five static language models

Five LMs were trained and used in various stages of decoding: qLM_2, qLM_3, LM_3, LM_{5a}, and LM_{5b}. To obtain each LM, six N-gram LMs were independently trained according to the first 6 corpora in Table 3, with modified Kneser-Ney smoothing [6], and then interpolated to maximize the likelihood of dev06.

qLM_2, a quick bigram, was highly pruned for fast decoding in the first recognition pass when cross-word triphone AMs were used. qLM_3, highly pruned from LM_3, was used for fast search when within-word triphone AMs were used. LM_3, a full trigram, was used in trigram lattice expansion and N-best generation. LM_{5a} and LM_{5b} were full 5-gram LMs designed for N-best rescoring. LM_{5a} was an interpolation of one word-based 5-gram and two class-based 5-gram LMs. LM_{5b} used count-based Jelinek-Mercer smoothing [7, 6] on the union of all training data counts.

3.3. Perplexity reduction

To understand the contribution of LM training data and lexicon, we used a subset of dev06 to compare the perplexities of different LMs. This subset of dev06 did not contain out-of-vocabulary (OOV) words in either the 49K or 60K lexicon, in order to avoid any confounding effects due to the special nature of OOV. This subset had about 29.4K words. Table 4 shows the word perplexity of a few LMs. It also shows the utterance-level log LM likelihood to better compare LMs whose lexicon sizes are different. With the improved lexicon and added training text, the word perplexity was reduced by 21% from the 49K-lexicon full 4-gram to the 60K-lexicon full trigram. Another dramatic perplexity reduction was achieved by expanding into 5-gram.

LM	word perplexity	log likelihood
49K full 4-gram	243.8	-75332
60K qLM_2	359.5	-79499
60K qLM_3	228.7	-73390
60K LM_3	193.0	-71094
60K LM_{5a}	77.9	-58830

Table 4. Word perplexity of different LMs on a subset of dev06 which did not contain OOV.

4. ACOUSTIC MODELING

4.1. Pronunciations

We developed a Chinese character pronunciation dictionary with multiple pronunciations per character. This single-character dictionary contained around 8000 entries, including almost all the possible simplified Chinese characters, with the first pronunciation being the most common pronunciation for a given character. To obtain the pronunciation of a new word (in the 60K lexicon, but not in the 49K lexicon), we simply concatenated the most common pronunciations for each character in the word. This allowed us to quickly build the new lexicon. The phone set was the BBN-based main-vowel phone set, which included 70 tonal phones plus one silence and one noise phone. For the GALE acoustic training data, we did not spend time transcribing English words with the Mandarin phone set. Instead, we simply set their pronunciations as a sequence of noise phones. The length of the sequence was made proportional to the length of the spelling. Once the pronunciation dictionary was constructed in this fashion, we could start training acoustic models.

For decoding test data, we adopted a different methodology to obtain pronunciations for English words. As mentioned earlier, there were more than 1700 English words in the decoding lexicon. To be able to recognize these English words, we simply mapped the SRI English pronunciation phone set into our Mandarin phone set.

4.2. Discriminative front end

Two front ends were used in our improved system, for the purpose of cross-adaptation and system combination. The first one computed 13-order MFCC and spline smoothed pitch [8], plus their first and second order derivatives, resulting in a feature vector of 42 dimensions.

The second front end computed MLP-based phoneme posteriors [9]. The targets of these MLPs were the 72 phones mentioned above. Following [9], the two-stage MLP-based phoneme posteriors were combined with the 9-frame one-stage MLP-based phoneme posteriors via inverse weighted entropy. Finally the combined 72 phoneme posteriors were reduced to 32 discriminative features via principal component analysis and appended into the 42-dim MFCC feature. That is, the second front end was a superset of the first front end.

Like the previous system, we applied vocal tract length normalization and feature mean and variance normalization to each "pseudo" speaker. A clustering algorithm based on the mixture weights of an MFCC-based Gaussian mixture model was used to group all utterances within the same broadcast show into acoustically homogeneous "pseudo" speaker clusters.

4.3. Multiple discriminative training criteria

Instead of one single AM in the previous system, two sets of AMs were trained in the new system, each with a different front end as described above.

The AMs used in the final system were all gender-independent, MPE trained [10] with fMPE feature transforms. For the MFCC-feature front end, there were 3000 decision-tree clustered states with 128 Gaussians per state. Crossword triphones were used in the MFCC system with feature-space speaker adaptive training (SAT), via single-class constrained MLLR [11].

For the MLP-feature front end, we did not have enough time to train an equally complex system as with the MFCC-feature system. Instead, we trained 3000*64 Gaussians for within-word triphones without speaker adaptive training. For a detailed account of our innovative combination of MLP features, fMPE transforms, and MPE training, please refer to [12].

5. DECODING

Before decoding, input broadcast shows were segmented into short utterances of average 6 seconds per utterance, and were subjected to pseudo-speaker clustering within the same show, consistent with how the training data had been processed. Speaker-based vocal track length normalization and utterance-based mean and variance normalization were then executed.

In the previous system, there was only one-pass self adaptation. Our new decoding structure consisted of two iterations of cross-adaptation, as illustrated in Figure 1. The first iteration of cross-adaptation is illustrated in the first column of the figure, where speaker-independent quick trigram decoding was performed using the MLP AM. The top hypothesis was used to cross adapt the MFCC AM.

The second iteration of cross adaptation is shown in the second column of the figure, where the adapted AMs generated separate N-best lists for 5-gram rescoring. The two 5-gram scores and the AM scores are finally combined via a character-level confusion network to dump the single best character sequence as the final output.

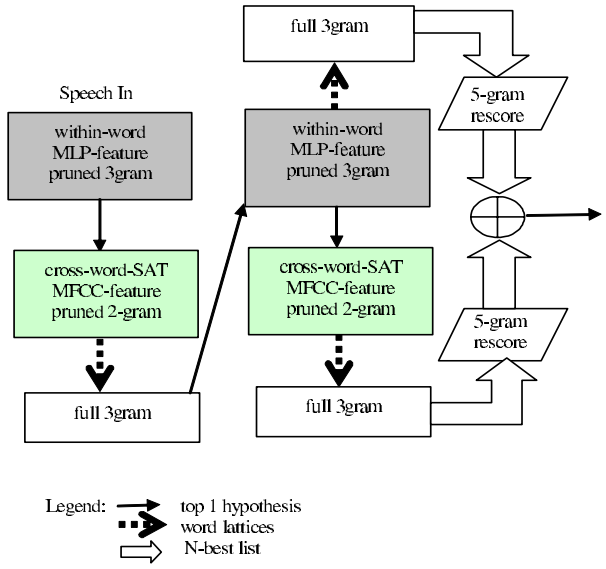

Fig. 1. Improved decoding architecture. Two iterations of cross-adaptation between MFCC and MLP acoustic models were applied.

The first row in Table 5 corresponds to the previous best system described in [1] on dev04 and eval04. The second row of the table demonstrates the effect of adding more acoustic data. To do a fair comparison, the same adaptation hypotheses used in Table 2 of [1] were used here for adaptation on the new crossword+SAT MPE trained model, and the same 4-gram word lattices were used to search for the best word sequence after adaptation. The small improvement could be blamed to the tight search space constrained by the lattices generated by the 49K lexicon and the inferior 4-gram.

The studies in [12] showed 8%-15% relative CER reduction using each of the three discriminative techniques (MLP features, MPE models, and fMPE transforms) alone, and 23% relative CER reduction with all three techniques combined, in speaker-independent bigram decoding. With all discriminative trainings incorporated into the two front-end AMs, the final output in Figure 1 achieved 3.7% CER on dev04 and 12.2% on eval04, as indicated in the last row of Table 5. The relative CER reduction was 38% and 24% respectively compared with the initial system.

AM	LM	Lex	AM	dev04	eval04
97 hr	420M wd	49K	CW+SAT MPE	6.0%	16.0%
465 hr	420M wd	49K	CW+SAT MPE	5.3%	15.1%
465 hr	850M wd	60K	MFCC fMPE+MPE || MLP fMPE+MPE	3.7%	12.2%

Table 5. CERs using different AMs, lexicons, and LMs.

Finally we compared the performance of the initial system and our improved system independently on ext06 and dev05bc in Table 6. The relative improvements were even bigger: 64% and 35.6% relatively.

AM	LM	Lex	AM	ext06	dev05bc
97 hr	420M wd	49K	CW+SAT MPE	15.0%	34.0%
465 hr	850M wd	60K	MFCC fMPE+MPE || MLP fMPE+MPE	5.4%	22.5%

Table 6. CERs of the initial vs. the improved system.

6. LANGUAGE MODEL ADAPTATION

Since there is little BC training text and conversational speech is spontaneous and difficult for speech recognition, the CER on BC genre test data is usually a few times higher than that of BN data. Therefore our goal here is to use the available training text and information from the test data to create a better LM that is more robust against BC genre.

We first separated the BN vs. BC training text as Table 3 indicates. The GALE BC text and the NTU-Web BC text, all together around 5M words, were used to train a set of BC specific LMs, denoted as BC_LM. The rest of the LM training data in Table 3, along with the CTS text, were pooled together to train another set of LM, BN_LM. Notice each of BC_LM and BN_LM contains five LMs (quick 2-gram, quick 3-gram, full 3-gram, word 5-gram, and count 5-gram as described in Section 3).

Table 7 lists the series of experiments conducted for LM adaptation on dev05bc. For fair comparison, the baseline static LM was created by interpolation of BC_LM and BN_LM, rather than the interpolation of six LMs described in Section 3. This actually yielded a small improvement on dev05bc, as shown in the first row of the table (21.9% vs. 22.5%). For the remaining three experiments, we always did unsupervised LM adaptation to maximize the likelihood of the first-pass decoding hypothesis (output of the first MLP-feature box in Figure 1). The quick 2-gram, quick 3-gram, full 3-gram and the word-based 5-gram were all adapted. Notice the count-based 5-gram was not adapted.

Adaptation Setup	First-pass	Final CER
Baseline (static)	24.9%	21.9%
$\lambda_1 BC\_LM + (1-\lambda_1) BN\_LM$	24.4%	21.2%
$\lambda_2 BC\_LM + (1-\lambda_2) BN\_LM^{'}$	24.3%	21.0%
$\lambda_3 BC\_LM + (1-\lambda_3) BN\_LM^{''}$	24.0%	20.6%

Table 7. CERs of dev05bc with LM adaptation.

As shown in the second row, dynamically interpolating BC_LM and BN_LM on a per-show basis and re-decoding from scratch improved the first-pass CER by 0.5% absolutely and the final CER by 0.7% absolutely.

Next we adapted BN_LM to $BN\_LM^{'}$ first, again based on the first-pass decoding hypothesis, using unigram marginals. Then BC_LM and $BN\_LM^{'}$ were interpolated using the initial hypothesis once more. We did not find further meaningful improvement, probably because the same hypothesis was used in both adaptation steps and therefore no new information was added.

Finally, we used the 2.7M-word GALE BC training text to adapt BN_LM ahead of time to $BN\_LM^{''}$, using unigram marginals. This made $BN\_LM^{''}$ more like BC genre and still well trained for BN. Then BC_LM and $BN\_LM^{''}$ were interpolated using the initial hypothesis. This gave us the best result on dev05bc at 20.6% of CER.

The same adaptation approaches applied to BN test set, eval04, did not yield further improvement. This implies that the static LM was well trained for BN genre and the dynamic LM adaptation was robust enough that no accuracy was lost even with the BC-adapted $BN\_LM^{''}$.

7. SUMMARY AND FUTURE WORK

We presented a state-of-the-art large vocabulary speech recognizer for Mandarin broadcast speech. Compared to its predecessor, it benefited from increased training data, improved lexicon and LMs, the combination of several discriminative acoustic training criteria, cross adaptation, system combination, and LM adaptation. All together, it brought down the CER of dev05bc from 34% to 20.6%, a 39% relative CER reduction.

We are currently investigating other approaches of LM adaptation. One approach is to adapt BN_LM with maximum a posterior learning using the in-domain GALE BC text, then dynamically interpolate the adapted BN_LM with BC_LM on a per-show or per-topic basis, based on the first-pass decoding hypothesis. We have promising results and are looking forward to further improvement.

8. ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

9. REFERENCES

[1] M.Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Proc. Interspeech*, 2006, pp. 1233–1236.

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

[3] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. ICASSP*, 2005.

[4] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer Speech and Language*, pp. 373–400, 2000.

[5] T. NG, M. Ostendorf, M.Y. Hwang, M. Siu, I. Bulyko, and X. Lei, "Web data augmented language models for Mandarin conversational speech recognition," in *Proc. ICASSP*, 2005, pp. 589–592.

[6] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Science Group, Harvard University, TR-10-98*, 1998.

[7] F. Jelinek and R. Mercer, "Interpolated estimation of markov source parameters from sparse data," in *Workshop on Pattern Recognition in Practice*, 1980.

[8] X. Lei, M. Siu, M.Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for Mandarin broadcast news speech recognition," in *Proc. Interspeech*, 2006.

[9] B. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks," in *Proc. ICSLP*, 2004.

[10] J. Zheng and A. Stolcke, "Improved discriminative training using phone lattices," in *Proc. Interspeech*, 2005, pp. 2125–2128.

[11] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[12] J. Zheng, O. Cetin, M.Y. Hwang, X. Lei, A. Stolcke, and N. Morgan, "Combining discriminative feature, transform, and model training for large vocabulary speech recognition," submitted to ICASSP 2007.