

Development of the 2008 SRI Mandarin Speech-to-text System for Broadcast News and Conversation

Xin Lei¹, Wei Wu², Wen Wang¹, Arindam Mandal¹, Andreas Stolcke¹

¹SRI International, Menlo Park, CA 94025 USA

²Univ. of Washington, Dept. of Electrical Engineering, Seattle, WA 98195 USA

{leixin,wwang,arindam,stolcke}@speech.sri.com, weiwu@ee.washington.edu

Abstract

We describe the recent progress in SRI's Mandarin speech-to-text system developed for 2008 evaluation in the DARPA GALE program. A data-driven lexicon expansion technique and language model adaptation methods contribute to the improvement in recognition performance. Our system yields 8.3% character error rate on the GALE dev08 test set, and 7.5% after combining with RWTH systems. Compared to our 2007 evaluation system, a significant improvement of 13% relative has been achieved.

Index Terms: lexicon expansion, language model adaptation, Mandarin speech recognition

1. Introduction

Recent advances have been made in our Mandarin broadcast news (BN) and conversation (BC) speech-to-text (STT) system, developed under the DARPA Global Autonomous Language Exploitation (GALE) program. We demonstrated that the core technologies for English STT such as discriminative training [1] and acoustic adaptation [2] are also applicable to the Mandarin language. Mandarin-specific components can be used to exploit the characteristics of the Chinese language. We have investigated lexical word segmentation, tone modeling, and pronunciation lexicon designs in [3, 4] to further improve system performance. In this paper, we give an overview of our Mandarin STT system for the GALE 2008 evaluation, and also describe our recent progress on automatic lexicon expansion and language model (LM) adaptation.

The remainder of this paper is organized as follows. In Section 2, we begin with a description of the acoustic and text data used in training and testing the system, as well as the automatically expanded lexicon. In Section 3, the acoustic and language models and the decoding architecture are described. Section 4 presents progress on LM adaptation. Section 5 demonstrates the recognition performance improvements via experiments. Finally, we summarize the findings and propose future work in Section 6.

2. Corpora and Lexicon

2.1. Acoustic Data

As shown in Table 1, we use about 1642 hours of BN and BC speech data collected by the Linguistic Data Consortium (LDC), including Mandarin Hub4 (30 hours), GALE Year 1 (747 hours), GALE Year 2 (394 hours), GALE Year 3 (317 hours), and GALE Year 4 Release 1 (154 hours) corpora. All the acoustic data have manual transcriptions from LDC. There are around 1.35 million audio segments in the acoustic training data, with an average length of 4.4 seconds per segment.

Corpora	BN	BC	Total
Hub4	30	–	30
GALE Year 1	355	392	747
GALE Year 2	192	202	394
GALE Year 3	139	178	317
GALE Y4R1	51	103	154
All	767	875	1642

Table 1: Acoustic training data in hours.

The amount of BN data is close to that of BC data. Training data is collected from broadcast stations such as CCTV, RFA, NTDTV and PHOENIX, spanning 1997 through April 2008. Comparing to the December 2007 evaluation, we have removed 89 hours of TDT4 data that is relatively noisy and hurts performance slightly, and added 230 hours of Year 3 and Year 4 data.

We test our system on two development test sets: DARPA GALE 2007 development set (dev07) and 2008 development set (dev08). The dev07 test set contains 2.4 hours of speech data from 74 broadcast shows, with 1.0 hour in BN and 1.4 hours in BC. The dev08 test set has 1.0 hour of speech from 44 shows, with 0.5 hour in both BN and BC.

2.2. Text Data

Our text corpora comprise 1.4 billion words, including transcriptions of the acoustic training data, the LDC-released Mandarin Gigaword corpus, GALE-related web text data, and conversational telephone speech text data. A few passes of cleaning are applied to the source text to remove HTML tags, punctuation, and corrupted GB2312 codes, and to normalize numbers from digits to spoken forms. Word fragments, laughter, and background noise transcriptions are mapped to a special garbage word.

A maximum likelihood (ML) word segmentation algorithm [3] is used to segment the training texts into multi-character words. Like many other systems, our Mandarin automatic speech recognition (ASR) system is based on “word” recognition with phone-based subword units rather than character-based recognition. Word-based ASR has the advantage that longer units lead to less acoustic confusability and longer character context in the LM. The potential disadvantage is a possible higher out-of-vocabulary (OOV) rate. To ameliorate the OOV problem, we add single-character words for all the characters that occur in the training data.

2.3. Pronunciation Lexicon

A well-designed decoding lexicon, in terms of both vocabulary selection and pronunciation dictionary, is an essential component for good ASR performance. Typically, the ASR lexicon is constructed manually using linguistic resources. The lexicon needs to be updated or expanded over time, especially for Mandarin broadcast speech where new words are constantly being created. In our 2008 system, we have adopted a data-driven framework to expand the lexicon with multiple pronunciations and new vocabulary words to improve Mandarin ASR performance [5]. The pronunciation variants of a Chinese word are enumerated by combining the pronunciation variants of each constituent character. All the variants are added to the dictionary and used to perform forced alignment of the acoustic training data. Multiple pronunciations are then selected from the forced alignment statistics. To extract new vocabulary words, an efficient algorithm using the forward and backward N-gram statistics is adopted. Additional details of this data-driven lexicon expansion approach can be found in [5].

3. Models and Decoding Architecture

3.1. Two Acoustic Systems

A key component of our system is cross-adaptation and system combination between two complementary subsystems. In [6] it was shown that three discriminative techniques are effective in reducing recognition errors: multi-layer perceptron (MLP) features, minimum phone error (MPE) training criterion, and feature space MPE (fMPE) transform. In practice, we have found that the gains from three discriminative techniques are additive, under a speaker-independent (SI) recognition setup. However, after unsupervised speaker adaptation (SA), combining all three techniques yields minimal further improvement compared with combining only two of them. Therefore, in designing our two acoustic models (AMs), we decided to choose the most effective combinations of two techniques: MLP+MPE and fMPE+MPE, where the most effective technique, MPE training, is always applied. To ensure diversity in ASR subsystems, we use different cepstral features (MFCC and PLP) for the two systems, and design a new pronunciation phone set with more diphthongs and neutral tones to improve performance on BC speech [3] for the PLP system.

Since Mandarin is a tone language, we have adopted an embedded tone modeling technique in both acoustic systems. Tonal phonetic units are used and 3-dimensional smoothed pitch features [4] are appended to the cepstral features in the front end. For the MFCC system, 32-dimensional MLP features are also appended to the feature vector, resulting in a feature vector of 74 dimensions. More details of the MLP features are discussed in [3]. The PLP system uses 42-dimensional PLP+ F_0 features, with fMPE transform applied. Except for the Hub4 data where we use hand-labeled speaker information, automatic speaker clustering is applied to cluster utterances in all other training corpora into *pseudo* speakers. Then vocal tract length normalization (VTLN) and cepstral mean and variance normalization (CMN/CVN) are applied for each pseudo speaker.

All HMMs have the standard 3-state Bakis topology without skipping arcs. Triphone models are trained first using the ML criterion, with 3500 states clustered by a decision tree and 128 Gaussian components per state. The Gaussian means are then re-estimated using the MPE criterion, with phone lattices generated by running recognition with a unigram LM and ML-trained AM on the training data. The within-word mod-

els are used for generating initial hypotheses for adaptation and speaker adaptive training (SAT) based feature transforms. The crossword models are trained with SAT transformed features, and used for decoding in later passes.

3.2. Language Models

BN and BC LMs are trained with modified Kneser-Ney smoothing on the expanded 64K-word decoding vocabulary (see Section 5 for details) and interpolated to optimize the perplexity on a heldout set. Pruned bigram and trigram LMs are used in the first-pass decoding. The full trigram is used in subsequent lattice rescoring if not applying LM adaptation, in which case BN and BC component trigram LMs are used. The full trigram LM has 82M bigram entries and 481M trigram entries. We have also tried using higher-order LMs for rescoring, but no significant improvement has been achieved. Both static and dynamic topic-adaptive language models are used in rescoring, as described in Section 4.

3.3. Decoding Architecture

Figure 1 illustrates the flowchart of our recognition system. The audio data is first segmented into small utterances of a few seconds long. Then speaker clustering is applied using Gaussian mixture models of static MFCC features and K-means clustering, the same as in training. VTLN is performed for each pseudo speaker, followed by utterance-based CMN/CVN on all features.

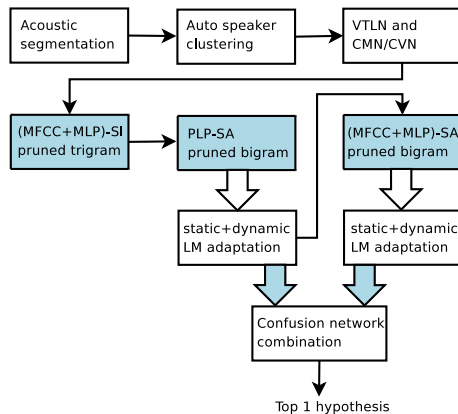


Figure 1: System decoding architecture. White block arrows represent lattices, and shaded block arrows represent N-best hypotheses.

As shown in the three shaded boxes in Figure 1, the decoding consists of three passes:

1. (MFCC+MLP)-SI: We begin with a fast search using a speaker-independent within-word triphone MPE-trained MFCC+MLP model and the aggressively pruned trigram LM. This gives us a good initial adaptation hypothesis quickly.
2. PLP-SA: Next we use the top hypothesis from the last step to learn the speaker-dependent SAT feature transform and perform MLLR adaptation per speaker, on the crossword triphone SAT+fMPE MPE trained PLP model. The pruned bigram is used in decoding to generate lattices. Then LM adaptation is performed to rescore

the lattices. The top hypothesis and an N-best list are generated.

- (MFCC+MLP)-SA: Similar to PLP-SA, we first run cross-adaptation, using the top hypothesis from the last stage to adapt the crossword triphone SAT MPE-trained MFCC+MLP model. The same LM adaptation is performed after the bigram decoding.

Finally, the N-best lists from PLP-SA and (MFCC+MLP)-SA stages are combined using a confusion network to generate the best hypothesis.

4. LM Adaptation

LM adaptation has been applied successfully to improve the performance of our Mandarin speech recognition system. LM adaptation is performed at the snippet level: for each snippet, the LM is adapted according to the first-pass decoding hypothesis to match its topic and genre. The adapted LM is used for rescoreing the bigram lattices. In our system, both static and dynamic LM adaptation are implemented.

4.1. Static LM Adaptation

In static LM adaptation, a number of static topic LMs are first built to cover different topics. Different from [7], where a manually constructed topic tree is used, we perform unsupervised clustering on the training corpora using the open source toolkit CLUTO [8] with TF-IDF features. A cosine distance is adopted to measure the similarity between two documents in the training corpora. Bisecting k -way clustering is performed to divide the training corpora into k topic clusters. A static topic LM is trained for each topic cluster.

To make the system more up-to-date, we only use data from the most recent 10 years (1997-2006) in Gigaword XIN, Gigaword ZBN, and Gigaword AFP news corpora to do unsupervised topic clustering. The number of topic clusters is set to 64 (a larger number of clusters did not further improve performance).

4.2. Dynamic LM Adaptation

Dynamic LM adaptation employs information retrieval technology to search for topic-matched documents for each snippet to build LMs. The advantage is that it divides the language model training corpora into finer granularity than static language model adaptation, and thus can build more specific LMs for each snippet.

The document ranking algorithm in our system is the traditional vector space model (VSM) for information retrieval. The bigram decoding hypotheses of all the utterances in each snippet are concatenated and used as a query, with stopwords removed. The top n documents with the highest ranking are grouped together to train a dynamic topic LM for the corresponding snippet. In our experiments, n is set to 120.

The corpora for retrieval in dynamic LM adaptation include GALE manual transcripts, GALE web data, NTU web data, UW web data, and data from the most recent 10 years (1997-2006) in Gigaword XIN and ZBN.

4.3. Interpolation Method

For each snippet, we put the 64 static topic trigram LMs and the corresponding dynamic trigram LM together to form a pool of LMs. To reflect the genre variation among different snippets, we also include the component BC LM that is trained with all

Table 2: *Lexicon expansion CER results (%)*.

Lexicon	dev07			dev08		
	BN	BC	All	BN	BC	All
baseline	4.9	20.6	13.9	6.2	19.2	12.9
+ pron	4.7	20.1	13.5	5.7	19.0	12.6
+ vocab	4.8	19.3	13.1	5.5	18.3	12.1

BC style data. The three LMs with the lowest perplexity on the bigram decoding hypothesis are selected. They are then linearly interpolated with the component BN LM to minimize the perplexity on the corresponding bigram decoding hypothesis.

5. Experimental Results

In this study, we focus on evaluating the efficacy of two techniques: data-driven lexicon expansion and LM adaptation. We also show that the STT system performance can be further improved significantly when combining with output from RWTH systems [9].

5.1. Lexicon Expansion

We start with a training lexicon of approximately 71K words, and select the most frequent 60K words as the baseline decoding lexicon. This baseline lexicon has several hundred multiple pronunciations. For fast turnaround in experiments, we perform first-pass decoding with the ML-trained within-word models and a pruned trigram LM, as in the (MFCC+MLP)-SI stage shown in Figure 1.

Recognition results in terms of character error rate (CER) are listed in Table 2. New pronunciations and words are learned from all the acoustic training data and transcripts. Adding 800 multiple pronunciations (denoted as “+ pron”) brings significant improvements on both dev07 and dev08. On top of the dictionary with added pronunciations, adding around 4K new vocabulary words (denoted as “+ vocab”) achieves 0.7% absolute improvement on the BC subset of both dev07 and dev08. We may attribute this to the fact that conversational speaking style renders words more confusable, and that concatenating words into longer units makes these less confusable again [10]. There are very small improvements on BN subsets after vocabulary expansion. This means that the vocabulary expansion is more beneficial for the BC genre, possibly because pronunciations are more affected by co-articulation in that genre.

Finally, we have native speakers of Mandarin inspect the added vocabulary items in the final dictionary, and eliminate those that were not semantically plausible as word units. This effort only improves the CER results by 0.1% absolute. The final decoding lexicon has a 64K-word vocabulary.

5.2. LM Adaptation

The LM adaptation experiments are performed with our final system for GALE 2008 evaluation (but without RWTH’s systems for combination). Table 3 shows the results on static and dynamic LM adaptation and their combination. The baseline in the experiments uses the full trigram LM for lattice rescoreing. It is shown that the combined LM adaptation can consistently reduce the CER on both LDC dev07 and dev08 sets, and on both BN and BC genres.

We also compare the relative contribution of the static and dynamic LM adaptation. Table 3 shows that static LM adaptation alone gives only minor improvement on recognition per-

Table 3: LM adaptation CER results (%).

	dev07			dev08		
	BN	BC	All	BN	BC	All
baseline	2.7	14.4	9.4	2.9	13.7	8.5
static only	2.4	14.3	9.2	2.9	13.8	8.6
dynamic only	2.3	14.2	9.1	2.7	13.4	8.3
static + dynamic	2.3	14.2	9.1	2.6	13.5	8.3

Table 4: System combination CER results (%).

System	dev07			dev08		
	BN	BC	All	BN	BC	All
RWTH only	3.1	13.6	9.2	3.9	13.1	8.7
SRI only	2.3	14.2	9.1	2.6	13.5	8.3
SRI + RWTH						
1) MFCC-SI	4.0	17.4	11.7	4.9	16.0	10.6
2) PLP-SA	2.6	15.0	9.7	3.0	14.2	8.8
3) MFCC-SA	2.5	14.6	9.4	2.8	13.9	8.5
4) + RWTH	2.2	12.3	8.0	2.6	12.1	7.5
2007 system	2.8	13.9	9.2	3.3	13.6	8.6

formance, and dynamic LM adaptation is as good as using static and dynamic LM adaptation together. This confirms our hypothesis that dynamic LM adaptation can provide a finer-grained adapted LM for rescoring. There is very little improvement or degradation on the BC genre when using static LM adaptation only. This may occur because the majority of the text training corpora is in the BN genre. The coverage of the BC genre is quite limited with static topic LMs. In addition, the relatively higher error rate in the BC first-pass decoding hypothesis makes it more difficult for either type of LM adaptation.

5.3. Combination with RWTH Systems

In the GALE 2008 evaluation, we perform system combination with RWTH Mandarin STT systems. Two systems from RWTH are used. Each system generates an N-best list for final confusion network based system combination. In addition, the RWTH top-1 hypothesis from combining the two RWTH systems is used for AM adaptation in the PLP-SA stage in Figure 1. The performance results of RWTH and SRI systems are shown in Table 4. We also show the results of the intermediate stages of the combined system. The two systems are comparable on both dev07 and dev08, yet with different error rates on BN and BC genres. This diversity in error patterns especially helps the combined BC performance. An improvement of 1.1% and 0.8% absolute is obtained from system combination on dev07 and dev08, respectively. Compared to our December 2007 evaluation system, the 2008 system reduces the CER on both test sets by around 13% relative.

6. Conclusions and Future Work

We have presented the recent progress in our Mandarin BN and BC ASR system. By porting the successful English ASR technologies, and exploiting Mandarin-specific characteristics, significant accuracy improvements have been made in our Mandarin recognizer. The more recent work on automatic lexicon expansion and LM adaptation contributes to the performance improvement.

The CER performance of our system on the BN genre is below 3% on the two development sets. However, although the

BN recognition is highly accurate, the error rate in the BC genre is still high because of its more conversational style. Future work includes investigating acoustic and LM adaptation toward the BC genre to improve BC recognition performance.

7. Acknowledgements

The authors thank our RWTH colleagues Christian Plahl and Ralf Schlüter for providing the RWTH system output for combination. Thanks also to Dimitra Vergyri and Jing Zheng at SRI for help during the system development.

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 (approved for public release, distribution unlimited). 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.

8. References

- [1] D. Povey and P. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, pp. 105–108.
- [2] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.
- [3] M. Hwang, G. Peng, M. Ostendorf, W. Wang, A. Faria, and A. Heidel, "Building a highly accurate Mandarin speech recognizer with language-independent technologies and language-dependent modules," to appear in *IEEE Trans. on Audio, Speech, and Language Processing*, 2009.
- [4] X. Lei, M. Siu, M. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for Mandarin broadcast news speech recognition," in *Proc. Interspeech*, 2006, pp. 1237–1240.
- [5] X. Lei, W. Wang, and A. Stolcke, "Data-driven lexicon expansion for Mandarin broadcast news and conversations speech recognition," in *Proc. ICASSP*, 2009, pp. 4329–4332.
- [6] J. Zheng, O. Cetin, M. Hwang, X. Lei, A. Stolcke, and N. Morgan, "Combining discriminative feature, transform, and model training for large vocabulary speech recognition," in *Proc. ICASSP*, 2007, pp. 633–636.
- [7] Q. Shi, S. Chu, W. Liu, H.-K. Kuo, Y. Liu, and Y. Qin, "Search and classification based language model adaptation," in *Proc. Interspeech*, 2008, pp. 1578–1581.
- [8] G. Karypis, "Cluto," <http://glaros.dtc.umn.edu/gkhome/views/cluto/>, 2002.
- [9] C. Plahl, B. Hoffmeister, M. Hwang, D. Lu, G. Heigold, J. Löff, R. Schlüter, and H. Ney, "Recent improvements of the RWTH GALE Mandarin LVCSR system," in *Proc. Interspeech*, 2008, pp. 2426–2429.
- [10] G. Saon and M. Padmanabhan, "Data-driven approach to designing compound words for continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 327–332, May 2001.