

Multifactor Adaptation for Mandarin Broadcast News and Conversation Speech Recognition

Wen Wang, Arindam Mandal, Xin Lei, Andreas Stolcke, Jing Zheng

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA

{wwang, arindam, leixin, stolcke, zj}@speech.sri.com

Abstract

We explore the integration of multiple factors such as genre and speaker gender for acoustic model adaptation tasks to improve Mandarin ASR system performance on broadcast news and broadcast conversation audio. We investigate the use of multifactor clustering of acoustic model training data and the application of MPE-MAP and fMPE-MAP acoustic model adaptations. We found that by effectively combining these adaptation approaches, we achieve 6% relative reduction in recognition error rate compared to a Mandarin recognition system that does not use genre-specific acoustic models, and 5% relative improvement if the genre-adaptive system is combined with another, genre-independent state-of-the-art system.

Index Terms: large vocabulary automatic speech recognition, broadcast news, broadcast conversation, genre classification, MAP adaptation, MPE-MAP, fMPE-MAP

1. Introduction

Building a large vocabulary automatic speech recognition (ASR) system for a new domain is a difficult task. One major reason is that collection of the training data for acoustic and language model development is expensive and very labor intensive. Acoustic model (AM) and language model (LM) adaptation attempt to adjust parameters of a well-trained existing model from one domain to a new domain, by exploring statistics obtained from a small in-domain seed corpus. There has been much work on AM adaptation approaches for improving ASR performance. In this paper, we investigate AM adaptation for a Mandarin large vocabulary ASR system for transcribing broadcast news (BN) and broadcast conversation (BC) audio under the Defense Advanced Research Projects Agency (DARPA) Global Autonomous Language Exploitation (GALE) program.¹ These two genres are significantly mismatched. The BN genre consists of “talking head” style broadcasts, i.e., generally one person reading a news script. The BC genre, by contrast, is more interactive and spontaneous, referring to free speech in news-style TV and radio programs and consisting of talk shows, interviews, call-in programs, live reports, and round-tables. It is important to note that the classification of a program as BN or BC is a non-trivial task. For its data collection efforts, the Linguistic Data Consortium (LDC) defined the genre of a show as the dominant genre of the program. Still, both genres can occur within a single program without indication of which portion is BN and which is BC. As a genre relatively new to the ASR community, BC recognition is much more challenging than BN. Besides being collected from various media programs/formats and featuring spontaneous speech phenomena, a closer look at

the Mandarin BC programs shows that they also cover a wide range of political, economical, civil, and cultural topics in China and around the world. Also, many BC audio files feature significant background noise, low recording quality, music background, accented Mandarin, and overlapped speech. There has been much effort on employing LM adaptation to improve ASR performance on BC including our effort [1, 2, 3], but less on AM genre adaptation. This paper focuses on exploring AM adaptation to improve ASR accuracy on BC data within a state-of-the-art Mandarin ASR system.

2. Speech and Text Corpora

In the work described here, we use about 1642 hours of BN and BC speech data collected by LDC for training our acoustic models. The training data includes 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. For testing, we use two different test sets for system development: the GALE 2007 development set (dev07)² and the GALE 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. LDC also provided manual snippet segmentations for both test sets, with 124 snippets for dev07 and 50 snippets for dev08. Note that with this manual segmentation, one snippet is a contiguous segment from a broadcast audio file and there is no mixture of BN and BC segments within one snippet.

Our text corpora comprise 1.4 billion words. In addition to transcripts of the acoustic training data, we add the LDC Chinese Gigaword corpus, all GALE-related Chinese web text releases, web text downloaded and released by National Taiwan University, Cambridge University, and the University of Washington, and the Mandarin conversational telephone speech recognition LM training data [4]. A maximum likelihood (ML) word segmentation algorithm [4] is used to segment the training texts into multi-character words. We applied data-driven lexicon correction and expansion algorithms [5] to refine our lexicon and defined the decoding vocabulary including 64K words.

3. SRI ASR System Architecture

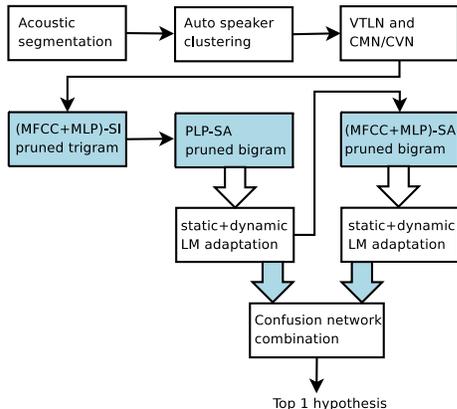
3.1. Two Acoustic Systems

The decoding architecture of SRI’s Mandarin ASR system is shown in Figure 1. As illustrated in the figure, our system employs semi-cross-adaptation and system combination between two complementary sub-systems. Three discriminative techniques that are effective in reducing recognition errors are used:

¹The goal of the GALE program is to develop computer software techniques to analyze, interpret, and distill information from speech and text in multiple languages.

²The dev07 set used here is the IBM-modified version, not the original LDC-released version.

Figure 1: System decoding architecture. The Block arrows represent N-best hypotheses.



multi-layer perceptron (MLP) features [6], minimum phone error (MPE) discriminative Gaussian training [7, 8], and MPE-trained discriminative feature transforms (fMPE) [9]. Following the findings in [10], we chose combinations of two out of these techniques: MLP+MPE and fMPE+MPE, with the most effective technique, MPE training, always applied. For the MFCC system, 32-dimensional MLP features are appended to the feature vector. The PLP system uses 42-dimensional PLP+ F_0 features, with fMPE transform applied. Except for the Hub4 data where we use the hand-labeled speaker information, automatic speaker clustering is applied in all other training corpora to cluster utterances into *pseudo speakers*. Then vocal tract length normalization (VTLN) and cepstral mean and variance normalization (CMN/CVN) are applied for each pseudo speaker. All triphone models are first ML trained, with 3500 states clustered by decision tree and 128 Gaussian components per state, followed by MPE training. The within-word model, with the highly pruned trigram LM, is used only at Step (MFCC+MLP)-SI for fast decoding. All later steps use cross-word (*cw*) models trained with constrained maximum likelihood linear regression (CMLLR) in feature space (speaker adaptive training). These *cw* models are then adapted on hypotheses from earlier stages using maximum likelihood linear regression (MLLR). The adapted models (denoted MLP-SA and PLP-SA) are used for search and for generating bigram lattices. Static and dynamic topic-adaptive trigram LMs were built for both front-ends and used for interpolating with the generic trigram LMs and rescored the two sets of lattices [3]. N-best lists were then extracted from rescored lattices, and we combined the two front-ends by a character-level confusion network combination (CNC) with rescored weights optimized for minimizing character error rate (CER).

4. Acoustic Model Adaptation

4.1. Automatic Genre and Gender Clustering

To adapt AMs on genre data, we need to label our 1642 hours of AM training data with BN and BC labels. The audio file for each show released by LDC is assigned to BN or BC, but we noticed that many shows are mixtures of BN and BC segments. We experimented with genre-adapting AMs at the show level using LDC labels, but observed no gains in accuracy. We then explored genre clustering of finer-grained units, i.e., snip-

pets within a show, such that no significant genre mixing occurs within a unit. However, no snippet-like segments were defined for AM training data. Hence, we developed a rule-based approach to segment each show into *pseudo snippets*, by combining information from the average length of test set snippets defined by LDC, the boundaries of automatic speaker clusters, and the duration of gaps between two consecutive segments. We segmented the 1642 hours of audio data into 806,854 pseudo snippets. We then followed LDC’s practice and assigned a single genre label to each pseudo snippet.

For automatic clustering of pseudo snippets into BN and BC classes, we explored various classifiers including Naive Bayes, K-nearest neighbor, and support vector machine (SVM). For selecting events of the probabilistic model, we investigated using words, documents, or a combination of documents and words. The classifiers used input features consisting of lexical N-grams, and feature selection was performed using occurrence counts and selecting words that had the highest mutual information with the class variable. The genre clustering procedure operated in a semi-supervised fashion. The genre classifier was initialized and trained on the combination of the dev07 and dev08 test set snippets with the LDC-provided snippet genre labels. Cross-validation experiments on this data set showed that the Naive Bayes model produced the best classification accuracy, with the SVM a close second. Hence, we used the Naive Bayes classifier for the following experiments. We employed self-training to bootstrap the genre model to label all AM training data pseudo snippets. For each iteration of self-training, we held out a small subset of the unlabeled pseudo snippets, used the current genre classifier to label them, filtered out samples with high ambiguity (i.e., similar posteriors between the two classes), and added the remaining samples into the training pool for the genre classifier. The procedure was repeated until all pseudo snippets were labeled with genre labels.

The speaker gender information is labeled in the transcripts released by LDC. However, preliminary results showed that automatically labeled gender outperformed the manual labels for adapting AMs on gender-clustered data. Hence, in this work, we followed our usual practice and used a gender-specific Gaussian Mixture Model (GMM) models trained on sampled gender-labeled AM training data, to predict gender labels for both training and test data. Then we performed AM adaptation on the automatically labeled gender data, using MPE-MAP and fMPE-MAP as described below.

4.2. MPE-MAP and fMPE-MAP+MPE-MAP

Maximum a posteriori (MAP) adaptation and its discriminative variants, such as MMI-MAP (maximum mutual information MAP) and MPE-MAP (minimum phone error MAP), have been widely applied to AM adaptation. Besides MLE-MAP, MMI-MAP, and MPE-MAP, the work in [11] introduced fMPE-MAP, which is an extension to the original fMPE algorithm, with an enhanced ability to port Gaussian models and fMPE transforms to a new domain. For fMPE-MAP, when a new transform is estimated, we run an MLE-MAP update of model parameters, instead of the MLE update in fMPE training. Experiments on the 2007 SRI-ICSI meeting recognition system showed that combining fMPE-MAP and MPE-MAP produced significant gain over MPE-MAP. In this work, for both MFCC+MLP and PLP front-ends, we investigate MPE-MAP adaptation on genres (MPE-MAP was also used for gender adaptation). In particular, for the PLP front-end, since the original model before adaptation was trained with fMPE, we also investigate

fMPE-MAP in combination with MPE-MAP. When applying fMPE-MAP + MPE-MAP on genre or gender data, we initialize the fMPE transform with the transform trained for the gender/genre-independent model and then apply fMPE-MAP to train gender/genre-dependent fMPE transforms and models in a few iterations.

5. Experimental Results and Discussions

We evaluated both AM genre adaptation (denoted **GR-A**) and a combination of gender adaptation (denoted **GD-A**) and GR-A. In this paper, we denote gender- and genre-independent AMs as **GD-I** and **GR-I**, respectively. Also, **GR-I** for rescoring weight optimization denotes optimizing the weights on the complete test set, whereas **GR-A** for rescoring weight optimization denotes splitting the test set into the BN and BC subsets and optimizing the weights on the two subsets separately. The splitting is based on the LDC-provided genre labels on the snippet level, which are verified to be accurate based on manual inspections.

We first investigated genre adaptation on the 2007 GALE evaluation system within-word PLP model trained with MPE on top of fMPE (denoted fMPE+MPE). Table 1 shows the CER on the dev07 test set from the baseline within-word fMPE+MPE-trained PLP model without gender adaptation or genre adaptation. For testing the MPE-MAP GR-A model (without gender adaptation, i.e., GD-I), we compared two conditions. One is to use LDC-provided genre labels and employ the corresponding genre adapted models for decoding a certain snippet. The other condition is to classify each snippet as BN or BC using the automatic genre classifier described in Section 4.1 that is trained on all AM training data, and then to decode the snippet with the corresponding genre adapted models. We found that for GR-A models, the automatically classified snippet genre labels work better than the LDC-provided genre labels, probably due to better matching on genre classification between training and testing. Hence, in the following experiments, we used our genre classifier for generating genre labels for snippets. The genre-adapted acoustic model yielded 0.6% absolute gain on CER, 0.3% on BN and 0.6% on BC. Note that although we used automatically classified genre labels to employ GR-A AMs, for easy comparisons with related work, the BN and BC breakdown of CERs for each test set are still computed using LDC-provided genre labels.

Table 1: CER (%) on dev07 from within-word fMPE+MPE trained PLP models trained on 1,500 hours of audio data, 1-best decoding with a pruned trigram LM.

	dev08 CER (%)		
	BN	BC	All
baseline	5.4	20.2	13.9
GD-I, GR-I			
GD-I, GR-A	5.2	19.7	13.4
LDC genre labels			
GD-I, GR-A	5.1	19.6	13.3
Automatic genre labels			

Next we compared the MPE-MAP GD-A, GR-I AMs to their GD-I, GR-I counterparts and observed gain from gender adaptation (0.5% on within-word MPE-trained MFCC+MLP model on the dev08 test set). We then investigated the effect of combining genre adaptation and gender adaptation in the SRI Mandarin ASR system described in Section 3, where the AMs

were trained on 1642 hours of audio data. To maintain diversity between the two sub-systems as much as possible, we adapted the models for the two front-ends differently. The MFCC+MLP within-word and cw MPE-trained models were adapted on gender data only using MPE-MAP. For the PLP cw fMPE+MPE-trained model, we compared three different conditions, namely, adapting it on the gender data only, on the genre data only, and combining the gender and genre adaptation for this model. This way, we can investigate the effect of adaptations on the PLP sub-system performance as well as on the system combination performance. The results are shown in Table 2. Note that **2g** and **3g** denote the pruned word bigram and trigram LMs, and **trigram-LMA** denotes the trigram topic and dynamically adapted language model for lattice rescoring.

For the PLP fMPE+MPE-trained cw model, gender adaptation and genre adaptation in this experiment employed MPE-MAP adaptation. For combining GD-A and GR-A, we first employed MPE-MAP adaptation on the male and female audio data, generating two gender-adapted models. Then for each gender-adapted model, we further applied MPE-MAP adaptation on the automatically clustered BN and BC data, for the audio data by this gender, creating corresponding BN and BC adapted models (denoted **GD-GR-A**). This two-stage adaptation procedure resulted in four adapted PLP cw fMPE+MPE models, namely, male-BN, male-BC, female-BN, and female-BC. These four models, after CMLLR and MLLR, were used for search and generating bigram PLP lattices and then rescored by trigram-LMA before system combination. As can be seen in the table, the PLP cw GR-A model produced 0.2% absolute gain over the GD-A model. We investigated two variations of GD-GR-A. The **GD-GR-A I** model adapted the GD-I, GR-I PLP cw first on gender partition and then on genre partition, with the I-smoothing model for the second stage genre adaptation set to the gender-adapted model. The **GD-GR-A II** model used the GD-I, GR-I generic model as the I-smoothing model for the second stage genre adaptation. We hypothesized that this approach of using a more generic I-smoothing model might reduce the impact of data fragmentation from the two stages of partitions. For all adaptation experiments, we optimized the number of adaptation iterations. Results shown in Table 2 demonstrated that both GD-GR-A I and GD-GR-A II models outperformed GD-A and GR-A models, and GD-GR-A II produced 0.2% absolute gain over the GD-GR-A I model. These results show that combining genre and gender adaptation can yield additional gain over either gender or genre adaptation, yet we need to find approaches to address the issue of data fragmentation. Table 2 also shows that with genre-adapted AMs, GR-A rescoring weight optimization works better than GR-I, and the best configuration of training the MPE-MAP GD-GR-A II model and conducting GR-A rescoring weight optimization yielded a final CER of 7.9%, a 0.4% absolute gain over the baseline 8.3% CER.

Next, we compared fMPE-MAP followed by MPE-MAP to MPE-MAP only adaptation in the two-stage adaptation framework. For fMPE-MAP followed by MPE-MAP, we optimized the values of adaptation weights, the number of iterations for fMPE-MAP and MPE-MAP, and the global learning factor for fMPE. The results are shown in Table 3. As can be seen, with fMPE-MAP followed by MPE-MAP (again, I-smoothing for the second-stage genre adaptation used the generic model) and GR-A rescoring weight optimization, the final CER on dev08 is 0.1% lower than that from MPE-MAP only. Hence the final gain on the SRI-only system from combining gender and genre adaptation over gender adaptation only is 0.5% abs. reduction

Table 2: CER (%) on dev08 from two sub-systems and system combination in the SRI-only system. All adaptations were conducted using MPE-MAP. The system architecture is described in Figure 1.

	Model	dev08 CER (%)		
		BN	BC	All
MFCC	a. within-word, GD-A, 3g	4.9	16.0	10.6
MFCC	b. cw, GD-A, 2g	4.0	15.0	9.7
	c. b+trigram-LMA	2.8	13.9	8.5
PLP	d. cw, GD-A, 2g	4.2	15.2	9.9
	e. d+trigram-LMA	3.0	14.2	8.8
	f. cw, GR-A, 2g	4.1	14.9	9.7
	g. f+trigram-LMA	3.0	13.9	8.6
	h. cw, GD-GR-A, 2g, I	4.0	14.7	9.5
	i. h+trigram-LMA	2.8	13.7	8.4
	j. cw, GD-GR-A, 2g, II	4.0	14.5	9.3
	k. j+trigram-LMA	2.7	13.5	8.2
Comb	l. c combines e, GR-I rescore	2.6	13.5	8.3
	m. c combines g, GR-I rescore	2.5	13.3	8.2
	n. c combines g, GR-A rescore	2.5	13.2	8.1
	o. c combines i, GR-I rescore	2.4	13.2	8.1
	p. c combines i, GR-A rescore	2.4	13.1	8.0
	q. c combines k, GR-I rescore	2.4	13.1	8.0
	r. c combines k, GR-A rescore	2.4	13.0	7.9

on CER (8.3% to 7.8%), a 6% relative improvement.

In the 2008 GALE evaluation system, we perform system combination with RWTH Mandarin ASR systems[3]. Results from combining the best configuration of the SRI-only system, as configuration (i) in Table 3, with RWTH’s ASR output, are shown in the second row of Table 4 for both dev07 and dev08 test sets, denoted *RWTH + SRI GD-GR-A*. The baseline system in the table used GD-A, GR-I AMs in the SRI system. In the end, genre AMs achieved 5% relative reduction on CER on dev08 and 4% relative reduction on dev07 over the baseline. We also investigated estimating pronunciation probabilities separately for BN and BC and added pronunciation scores as an extra knowledge source for GR-A knowledge source combination. However, this addition did not bring improvement on CER.

In conclusion, we investigated the use of multifactor clustering of acoustic model training data and the application of MPE-MAP and fMPE-MAP+MPE-MAP AM adaptations. We found that by effectively combining these adaptation approaches, we can achieve 4% to 5% relative reduction in the overall recognition error rate of SRI/RWTH’s GALE evaluation Mandarin ASR system.

6. 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 of SRI for help during the system development and Bin Zhang and Mari Ostendorf of University of Washington for discussions on genre classification.

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.

Table 3: CER (%) on dev08 from two sub-systems and system combination in the SRI-only system, comparing fMPE-MAP+MPE-MAP and MPE-MAP only. The system architecture is described in Figure 1.

	Model	dev08 CER (%)		
		BN	BC	All
MFCC	a. within-word, GD-A, 3g	4.9	16.0	10.6
MFCC	b. cw, GD-A, 2g	4.0	15.0	9.7
	c. b+trigram-LMA	2.8	13.9	8.5
PLP	d. cw, GD-GR-A, 2g	4.0	14.5	9.3
	MPE-MAP II e. d+trigram-LMA	2.7	13.5	8.2
	f. cw, GD-GR-A, 2g	3.9	14.1	9.0
	fMPE-MAP+MPE-MAP g. f+trigram-LMA	2.6	13.1	8.0
Comb	h. c combines e, GR-A rescore	2.4	13.0	7.9
	i. c combines g, GR-A rescore	2.4	12.9	7.8

Table 4: CER (%) on dev07 and dev8 from system combination of SRI and RWTH ASR systems. The second row is from combining the RWTH ASR output with the updated SRI system with gender and genre adaptations.

System	dev07 CER (%)			dev08 CER (%)		
	BN	BC	All	BN	BC	All
Baseline	2.2	12.3	8.0	2.6	12.2	7.6
RWTH + SRI GD-GR-A	2.0	11.8	7.7	2.2	11.7	7.2

7. References

- [1] D. Mrva and P. Woodland, “Unsupervised language model adaptation for Mandarin Broadcast Conversation transcription”, in *Proceedings of ICSLP*, pp. 1961–1964, 2006.
- [2] W. Wang and A. Stolcke, “Integrating map, marginals, and unsupervised language model adaptation”, in *Proceedings of Interspeech/Eurospeech*, pp. 618–621, Antwerp, Belgium, August 2007.
- [3] X. Lei, W. Wu, W. Wang, A. Mandal, and A. Stolcke, “Development of the 2008 SRI Mandarin speech-to-text system for broadcast news and conversation”, in *Proceedings of Interspeech*, 2009.
- [4] 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.
- [5] X. Lei, W. Wang, and A. Stolcke, “Data-driven lexicon expansion for Mandarin broadcast news and conversation speech recognition”, in *Proceedings of ICASSP*, 2009.
- [6] B. Chen, Q. Zhu, and N. Morgan, “Learning long-term temporal features in LVCSR using neural networks”, in *Proceedings of ICSLP*, 2004.
- [7] D. Povey and P. Woodland, “Minimum phone error and I-smoothing for improved discriminative training”, in *Proceedings of ICASSP*, 2002.
- [8] J. Zheng and A. Stolcke, “Improved discriminative training using phone lattices”, in *Proceedings of Interspeech*, pp. 2125–2128, 2005.
- [9] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, “fMPE: Discriminatively trained features for speech recognition”, in *Proceedings of ICASSP*, 2005.
- [10] 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 *Proceedings of ICASSP*, pp. 633–636, 2007.
- [11] J. Zheng and A. Stolcke, “fMPE-MAP: Improved discriminative adaptation for modeling new domains”, in *Proceedings of Eurospeech*, 2007.