Accent Detection and Speech Recognition for Shanghai-Accented Mandarin

Yanli Zheng[⊕], *Richard Sproat*[⊕], *Liang Gu*[†], *Izhak Shafran*[‡],
Haolang Zhou[‡], *Yi Su*[‡], *Dan Jurafsky*[⋆], *Rebecca Starr*[⋆], *Su-Youn Yoon*[⊕]

University of Illinois[⊕], IBM T. J. Watson Research Center[†]
Johns Hopkins University[‡], Stanford University[⋆]
{zheng3,rws,syoon9}@uiuc.edu[⊕], lianggu@us.ibm.com[†],
{zakshafran,haolangzhou,suy}@jhu.edu[‡], jurafsky@stanford.edu[⋆], starr@sundell.net[⋆]

Abstract

As speech recognition systems are used in ever more applications, it is crucial for the systems to be able to deal with accented speakers. Various techniques, such as acoustic model adaptation and pronunciation adaptation, have been reported to improve the recognition of non-native or accented speech. In this paper, we propose a new approach that combines accent detection, accent discriminative acoustic features, acoustic adaptation and model selection for accented Chinese speech recognition. Experimental results show that this approach can improve the recognition of accented speech.

1. Introduction

Accent is by far the most critical issue for the state-of-the-art Chinese automatic speech recognition (ASR) systems. This is because Chinese is a language with so many dialects including Mandarin, Wu (spoken by Shanghainese), Yue (spoken by Cantonese), Min (spoken by Taiwanese), etc. Although the official spoken language is *Putonghua* (also called *Standard Mandarin* or *Mandarin* in the speech recognition literature), it is spoken extremely differently by speakers living in different dialectal regions of China. As a result, current ASR systems trained on Putonghua or Standard Mandarin often experience a dramatic accuracy loss for speakers with strong accents.

Active research has been carried out on dialectal or foreign accented speech recognition during the past few years. The proposed methods vary from simply collecting data in that accent and training a recognizer, to various ways of adapting recognizers trained on unaccented speech. Wang, Schultz, and Waibel [1] investigated German-accented English speakers in the VERBMOBIL (conversational meeting planning) task. Tomokiyo and Waibel [2] examined the task of recognizing Japanese-accented English in the VERBMOBIL domain. In both tasks, it was found that training on non-native speech data, achieves the most obvious gains in performance on accented data. The simplest use of adaptation was merely the direct use of MLLR (Maximum Likelihood Linear Regression) to adapt individually to each test speaker. In [3], in order to recognize Shanghainese-accented Putonghua, Huang et al. applied standard speaker MLLR adaptation to a Microsoft Whisper system that had been trained on 100,000 sentences from 500 speakers living in the Beijing area. In [1, 2], MLLR was adapted not just to the single accented test speaker, but to a larger number of accented speakers. Research in [1, 2, 3] shows the effectiveness of MLLR or MAP (Maximum A Posteriori) adaptation on accented speech. but it did not report whether combining MLLR and MAP could be helpful for accented ASR.

While some promising results have been published on accented speech recognition using the above approaches, the recognition accuracy on accented speech is still low and definitely needs further improvement. In particular, some research issues remain open. First, more sophisticated forms of MLLR or MAP may be applied, such as MLLR using phone-specific transforms rather than a single global transform. Furthermore, our research shows that current adaptation schemes have varied performance on different groups of speakers. Second, the effect of combining MLLR and MAP in accented ASR needs to be explored. Third, the accent of each speaker should be treated as a matter of degree. Previous work on accented Chinese speech recognition [3, 4, 5, 6] typically treats speakers from a given dialectal region as a single class. In reality, these speakers clearly have different degrees of accent.

In this work we optimized the MAP/MLLR combination for our task. Then, building on this optimal MAP/MLLR combination, we developed new approaches to detecting and utilizing degree of accent in accented ASR. A series of new algorithms is proposed: phoneme-based automatic accent detection, formant-augmented acoustic features for accented speech, and accent-based model selection during acoustic model decoding.

For the sake of simplifying our experiments we focus here only on one form of accented Putonghua, namely the accent of people of Shanghai whose native language is Shanghainese, which belongs to the Wu dialect group, a group with 87 million speakers. Thus all experiments in this paper were performed on Wu-accented conversational Chinese speech. Nevertheless, we believe that our proposed approaches will also be helpful for accented speech with other Chinese dialects or in other languages.

2. Data Collection and Transcription

In this paper, we use spontaneous speech data collected from 50 male and 50 female speakers of Wu-accented Standard Chinese, henceforth *Putonghua*. The spontaneous speech consisted of free-form monologues where the speaker was asked to discuss a topic of their choice from a small set of predetermined topics. It was recorded at 16KHz using a head-mounted microphone. The data were orthographically transcribed, and phonetically transcribed into syllable *initials* (onsets) and *finals* (nucleus+coda) — henceforth "IF" refers to *initial-final* phoneset. Canonical *pinyin* transliterations were also derived from the orthographic transcription. Finally, speakers were classified by experts into their "Putonghua level" on a 6-point scale ranging from 1A (most standard) to 3B (least standard); all of our speakers fell in the range 2A–3B. Of the 100 speakers, 80 were used for training data and the remaining 20 for test data. Further details on the data can be found in [7].

3. Features of Accentedness

One of the phonetic properties of Shanghai-accented Mandarin is the tendency to replace the standard retroflex fricatives and affricates *sh/ch/zh* with their alveolar equivalents *s/c/z*. In a sociolinguistic study [8], Starr and Jurafsky have shown that the amount of retroflexion correlates with various socioeconomic factors, such as age, gender and education level. Higher amounts of retroflexion (i.e., more standard pronunciation) are found in younger speakers, female speakers and more educated speakers; see Figure 1. Starr and Jurafsky argue for economic development being the primary factor underlying the shift towards more standard pronunciation. Given this work, it turns out that simple but robust measures of degree of accentedness are the proportion of alveolar affricates and fricatives in the speech of a given speaker, or in other words:

$$\frac{C(s)}{C(s)+C(sh)},\ \frac{C(z)}{C(z)+C(zh)},\ \frac{C(c)}{C(c)+C(ch)}$$

where $C(l)$ is the count of the label l in the transcription for each speaker, $l = s, z, c, sh, zh, ch$.

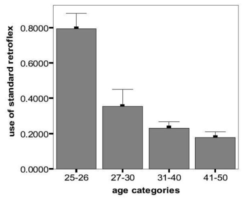

Figure 1: Retroflex in spontaneous speech by age category

Other similar measures also correlate with accentedness. Thus, for example, Shanghai speakers often substitute the coda *eng* for *en* and *ing* for *in*, and vice versa; the alveolar-to-dental substitution is about twice as common as the dental-to-alveolar substitution in our data. Clusters based on retroflex/alveolar ratios plus nasal count ratios—see Section 4—correlated somewhat better with human judgments of accentedness than did clusters based on retroflex/alveolar ratios alone. However, the inclusion of the nasal data did not correlate with better ASR performance, so we will not report on this further here.

4. Automatic Identification of Accentedness

In this section, we explore the methods of detecting accentedness degree. For simplification, we classified the speakers into two groups: "more standard" with "Putonghua level" 2A–2B and "more accented" with "Putonghua level" 3A–3B. Previous work on accentedness detection such as [5] has mostly focused on using MFCC coupled with GMM's or other classifiers to classify a speaker's utterances as belonging to a specific accent type. We experimented with using GMM classifiers with MFCC plus F_0 (pitch) as the input features. On the test speakers, the accuracies are 69% for the "more standard group" and 86% for the "more accented" group.

In our approach, we use the alveolar/retroflex proportions discussed in the previous section. Since we cannot assume manual transcriptions for target speakers, the ratios are computed from decoding lattices generated using our baseline MBN model (introduced in Section 6.1). The single best transcription is very errorful, but in previous work [9, 10] it has been shown that if one computes counts for strings over a lattice rather than over the single best path, one can generally improve one's estimate of the population statistics. Following [10], counts are computed for a segment α by summing the probability of each path π in the lattice, multiplied by the number of times α occurs on π. Counts for each speaker are then derived in the obvious way by summing the thus-derived lattice counts. Thus we construct a "count" $C(\alpha|L)$ for a given label α on a path π in a lattice L as follows:

$$C(\alpha|L) \quad = \quad \sum_{\pi \in L} p(\pi) C(\alpha|\pi)$$

where $C(\alpha|\pi)$ is the number of times α is seen on path π. In this way, phone population estimates can be derived for each lattice and hence for each speaker, and the requisite ratios computed.

Given the counts for each speaker, we used Cluto 2.1.1 [11] to decide upon two "clusters" each (more accented, more standard) for the training and testing data. By default Cluto's *vcluster* uses a repeated bisections method with a cosine distance measure. The agreement between the "clusters" and human assignment of accentedness is fairly good: on the test speakers, the accuracy of the clustering is 78% for the "more standard group" and 72% for the "more accented" group. Corresponding ASR results will be shown and discussed in Section 6.4.

5. Model selection based on accentedness

To make use of the prior knowledge of accentedness, we propose a model-selection algorithm. Suppose that there are M different acoustic models, θ_1, θ_2, ... , θ_M, given observation x, we want to find the best acoustic model according to Eq. 1,

$$\begin{aligned} \theta_{MAP} &= \underset{k=1,2,\cdots,M}{argmax}\ p(\theta_k|x) \\ &= \underset{k=1,2,\cdots,M}{argmax}\ \sum_a \underbrace{p(\theta_k|a)}_{\theta_k \perp x|a} \quad \underbrace{p(a|x)}_{\text{accentedness classifier}} \end{aligned} \qquad (1)$$

where a is the accentedness variable. For a binary case classification in Section 4, we have M=2,

$$a = \begin{cases} 1 & \text{if the speaker is "more standard"} \\ 2 & \text{if the speaker is "more accented"} \end{cases}$$

and

$$p(\theta_k|a) = \delta(k - a)$$

To make Eq. 1 work, first, we need a reliable accentedness classifier, as described in the previous section; second, we need to find the acoustic model θ_k, which is most appropriate for the degree of accentedness. In Section 6.2 and 6.3, we show how to find two acoustic models that favor different accent groups. And Section 6.4 reports the results of *model-selection* experiment, showing the effectiveness of the accentedness classifier.

6. Experiments

6.1. Baseline system

A word bigram language model is used in all the experiments. The test and training corpora were segmented using a maximum matching algorithm using a fixed dictionary consisting of 50,647 entries developed at Tsinghua University. Language model training corpora consisted of the following conversational Putonghua data with 1.22 million characters:

- Mandarin HUB5 (200 telephone conversations of up to 30 minutes each)
- 100 hours of conversational Putonghua speech collected by Hong Kong University of Science and Technology.
- The transcriptions from the 6.3 hours of training data from our Wu-accented speech corpus.

Standard MFCC-based acoustic models with 14 mixtures per state were constructed using HTK version 3.2 [12]. Two baseline acoustic model training sets were used:

- MBN: 1997 Mandarin Broadcast News corpus (Hub-4NE), consisting of 30 hours of speech from mostly trained speakers.
- WU: 6.3 hours of Wu-accented training data.

The MBN data was chosen since it matches our data in one respect, namely that it is wideband recording. The baseline result for the MBN acoustic model was 61.0% *Character Error Rate* (CER — the standard measure of performance in Chinese speech recognition). For the Wu-accented training data, the CER was 44.2%.

6.2. Adaptation of Acoustic Models using Standard MAP/MLLR

Previous research [1] suggests that MLLR can be used on groups of speakers in a training set to help adapt acoustic models to foreign accent. However, applications of MLLR in this multi-speaker adaptation environment have been limited to a *single global transform*. Huang et al. [3] used MLLR with 65 phone-based transforms on individual test speakers, but they turned off the MLLR in their standard baseline system.

In this section, we explore adaptation techniques in both speaker independent (SI) and speaker dependent (SD) systems. We first show that combining MLLR with multiple transforms and MAP can improve the recognition performance. We then show that the gain we get from speaker independent adaptation can be further improved with speaker dependent adaptation. The experiments contrast two types of adaptation: adapting out-of-domain acoustic model (MBN) to indomain (Wu) data; and adapting in-domain (Wu) models on speakers with varying accent.

We experimented with various supervised adaptation techniques on the training set. Results are show in Table 1. This table shows that *IF-60*, the MLLR with 60 phone-based transforms, is significantly better than *Auto-60* which is the MLLR of 60 transforms by data-driven clustering. Assigning transforms to each IF allows the acoustic model to capture systematic variations associated with accent, while the data-driven regression tree cannot take advantage of this prior knowledge. By applying MAP on top of both of the MLLRs, the gap is narrowed. We also found that the best combined system is 1.7% absolute better than applying MAP alone.

Baseline (no adapt.)	+ MAP	+ MLLR (Auto-60)
61.0%	45.4 %	51.2%
+ MLLR (IF-60)	+MLLR+MAP(Auto-60)	+MLLR+MAP(IF-60)
47.8%	44.5%	**43.7%**

Table 1: CER (%) Comparison of varies types of adaptation to baseline acoustic models trained on MBN corpus

It has been reported [3, 2] that speaker dependent MLLR adaptation is very useful for accented or non-native speech. We performed speaker-dependent adaptation on both MMIF-60 and WU baseline models, where MMIF-60 represents the best model of +MLLR+MAP (IF-60) in Table 1. Two global transforms are used in our experiment, one for the silence model and one for speech models. The results in Table 2 shows that we can get about 3% absolute gain after speaker adapatation.

Table 2 also shows the speaker averaged CER for "more standard" group and "more Accented" group, which have been defined by retroflex ratio-based classifier from Section 4. It can be observed from the table that MMIF-60 favors "more standard" speakers, and WU favors "more accented" speakers. For comparison, the results of speaker independent systems for the same groups of speakers are also listed in Table 2.

	Speaker-indep.		Speaker-dep.	
Speaker Group	WU	MMIF-60	WU	MMIF-60
more standard	39.6	**37.5**	36.5	**34.7**
more accented	**49.0**	50.3	**46.0**	47.0

Table 2: Speaker averaged CERs (%) of speaker dependent (SD) and speaker independent (SI) systems

6.3. Study of accent discriminative acoustic features

The results in Table 2 show that there is an approximately 10% (absolute) gap between "more accented" and "more standard" speakers for all the SI and SD models. In this section we present methods for improving the performance of "more accented" speakers so that the gap can be narrowed.

In [13], Liu and Fung show that besides energy, formant frequency and pitch are also helpful in a task for accent classification. It is reasonable to assume that some acoustic features, such as formant parameters, pitch, word-final stop closure duration etc., might be more discriminative for accented speech. Therefore it may be helpful to add some of these features to the accented speech recognizer. To test this assumption, we carried out preliminary experiments by appending formant parameters to MFCC features. The formant parameters were estimated automatically using the formant tracking algorithm in [14].

In our experiment, we choose first three formants ($F_1^3 = [F_1\ F_2\ F_3]$) and their amplitudes ($\eta_1^3 = [\eta_1, \eta_2, \eta_3]$) as the accent related features. The detailed definition and estimation formulas of η are given in [14]. Two acoustic models were trained by appending F_1^3 and η_1^3 to the 39 dimensional MFCC vectors respectively.

The results are given in Table 3. We observed that the model with η_1^3 was able to improve 5 out of the 11 speakers in the "more accented" group; and the model with appended F_1^3 was only able to improve 2 out of the 11 speakers in the "more accented" group. The performance was degraded for speakers in the "more standard" group for both models.

The above experiment shows that formant amplitudes η_1^3 might contain extra information for accent discrimination. We therefore constructed a new *accent favorable* model $\mathbf{WU}_\eta$ by finding the best path in the union of the two decoding lattices from the Wu baseline model and the new model with extra feature dimensions η_1^3. As shown in Table 3, compared to the WU baseline model, the overall CER for this group is reduced to 48.2%, and the CERs were reduced for 8 out 11 speakers in the "more accented" group.

A similar experiment was done for speaker dependent system, where two models (WU and $MFCC + \eta_1^3$) were adapted for each individual test speaker and a WU_η was obtained for each test speaker. Compared to the WU baseline model test speaker adaptation, the CERs were reduced for 9 out 11 speakers in the "more accented" group.

	MFCC+F_1^3	MFCC+η_1^3	WU_η
SI	49.4	48.9	**48.2**
SD	-	46.1	**45.6**

Table 3: Average CER (%) of more accented speakers by modeling both MFCC and formant parameters

6.4. Experiment on Model Selection

In this section, we use the following model selection strategies:

$$\theta = \begin{cases} \theta_{MMIF-60} & \text{if the speaker is in cluster 1} \\ \theta_{WU} \ or \ \theta_{WU_\eta} & \text{if the speaker is in cluster 2} \end{cases} \quad (2)$$

Table 4 shows the results of model selection between WU or WU_η and MMIF-60 models based on automatic accent detection results in Section 4. The results show that by using the ratio of counts of particular fricatives and affricates as the input to the accent classifier, we were able to improve the WU baseline by 1% absolute in both SI and SD cases. Furthermore, formant amplitude η is useful to discriminate "accented speakers".

		WU+MMIF-60		WU_η + MMIF-60	
		GMM	SCZ	GMM	SCZ
SI	more acc.	-	49	-	**48.2**
SI	speaker avg.	44.4	43.8	44.3	**43.4**
SD	more acc.	-	46	-	**45.6**
SD	speaker avg	41.3	40.9	41.2	**40.7**

Table 4: CER (%) for model selection based on accent detection. "WU+MMIF-60" is selection between WU and MMIF-60, and "WU_η + MMIF-60" is selection between WU_η and MMIF-60 according to Eq. 2. "SCZ" is selection based on accent detection using the retroflex count ratio. "GMM" is selection based on GMM-based accent detection.

7. Conclusion

We report the approach of combining accent detection, accent discriminative acoustic features, acoustic adaptation and model selection to the problem of accented Chinese speech recognition. Experimental results show an 1.0∼1.4% absolute reduction of character error rate over the most state-of-the-art acoustic modeling techniques on Wu-accented Chinese speech. We show that accent classification followed by model selection can significantly improve performance when the degree of accent varies significantly. The accent classification can be further enhanced by using techniques such as in [15] and the models by using accent-specific decision trees. A future area will be to investigate replacing our hard decision on accenting with soft model selection integrated into ASR.

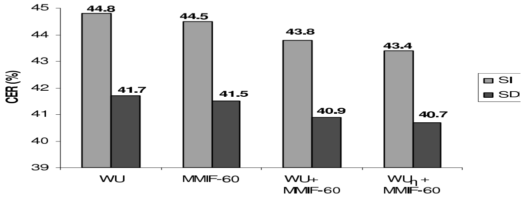

Figure 2: Summary of results

8. Acknowledgments

This material is based upon work supported by the NSF under Grant No. 0121285. We thank Stavros Tsakalidis and other members of the Workshop 2004 Chinese Dialectal Speech Recognition team, and the faculty and staff of the CLSP at JHU.

9. References

[1] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *ICASSP 2003*. IEEE, 2003, pp. 540–543.

[2] L. Mayfield Tomokiyo and A. Waibel, "Adaptation methods for non-native speech," in *Proceedings of Multilinguality in Spoken Language Processing*, Aalborg, 2001.

[3] C. Huang, E. Chang, J. Zhou, and K.-F. Lee, "Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition," in *ICSLP 2000*, Beijing, 2000, pp. 818–821.

[4] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian mixture models," in *IEEE Workshop on ASRU*, Italy, 2001.

[5] C. Huang, T. Chen, and E. Chang, "Accent issues in large vocabulary speech recognition," *International Journal of Speech Technology*, vol. 7, pp. 141–153, 2004.

[6] Y. Liu and P. Fung, "Partial change accent models for accented mandarin speech recognition," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, U.S. Virgin Islands, 2003.

[7] R. Sproat, T. F. Zheng, L. Gu, D. Jurafsky, I. Shanfran, J. Li, Y. Zheng, H. Zhou, Y. Su, S. Tsakalidis, P. Bramsen, and D. Kirsch, "Dialectal chinese speech recognition: Final technical report," 2004, http://www.clsp.jhu.edu/ws2004/.

[8] R. Starr and D. Jurafsky, "Phonological variation in Shanghai Mandarin," in *NWAV 33*, Ann Arbor, September 2004.

[9] M. Bacchiani, M. Riley, B. Roark, and R. Sproat, "MAP stochastic grammar adaptation," 2005, to appear *Computer Speech and Language*.

[10] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *HLT-NAACL 2004: Main Proceedings*, D. M. Susan Dumais and S. Roukos, Eds. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2 - May 7 2004, pp. 129–136.

[11] G. Karypis, *Cluto: A Clustering Toolkit*, University of Minnesota, Department of Computer Science, Minneapolis, MN, 2003, http://www-users.cs.umn.edu/ karypis/cluto/.

[12] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2.1)*, Cambridge University, Cambridge, 2002, http://htk.eng.cam.ac.uk/.

[13] W. K. Liu and P. Fung, "Fast accent identification and accented speech recognition," in *ICASSP*, 1999.

[14] Y. Zheng and M. Hasegawa-Johnson, "Stop consonant classification by dynamic formant trajectory," in *ICSLP*, Jeju Island, Korea, 2004.

[15] I. Shafran, M. Riley, and M. Mohri, "Voice signatures," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, U.S. Virgin Islands, 2003.