# ACCENT MODELING BASED ON PRONUNCIATION DICTIONARY ADAPTATION FOR LARGE VOCABULARY MANDARIN SPEECH RECOGNITION

*Chao Huang, Eric Chang, Jianlai Zhou, Kai-Fu Lee*

Microsoft Research, China
5F, Beijing Sigma Center, No. 49, Zhichun Road Haidian District
Beijing 100080, P.R.C.
i-chaoh@microsoft.com

## ABSTRACT

A method of accent modeling through Pronunciation Dictionary Adaptation (PDA) is presented. We derive the pronunciation variation between canonical speaker groups and accent groups and add an encoding of the differences to a canonical dictionary to create a new, adapted dictionary that reflects the accent characteristics. The pronunciation variation information is then integrated with acoustic and language models into a one-pass search framework. It is assumed that acoustic deviation and pronunciation variation are independent but complementary phenomena that cause poor performance among accented speakers. Therefore, MLLR, an efficient model adaptation technique, is also presented both alone and in combination with PDA. It is shown that when PDA, MLLR and PDA+MLLR are used, error rate reductions of 13.9%, 24.1% and 28.4% respectively are achieved.

## 1. INTRODUCTION

There are multiple accents in Mandarin. A speech recognizer built for a certain accent often obtains 1.5 ~ 2 times higher error rate when applied to another accent. The errors can be divided into two categories. One type of errors is due to misrecognition of confusable sounds by the recognizer. The other type of errors are those due to the speaker's own pronunciation errors. For example, some speakers are not able to clearly enunciate the difference between /zh/ and /z/. Error analysis shows that the second type of errors constitutes a large proportion of the total errors when a speech recognizer trained on Beijing speakers is applied to speech from Shanghai speakers. A key observation is that speakers belonging to the same accent region have similar tendencies in mispronunciations.

Based on the above fact, an accent modeling technology called pronunciation dictionary adaptation (PDA) is proposed. The basic idea is to catch the typical pronunciation variations for a certain accent through a small amount of utterances and encode these differences into the dictionary, called an accent-specific dictionary. The goal is to estimate the pronunciation differences, mainly consisting of confusion pairs, reliably and correctly. Depending on the amount of the adaptation data, a dynamic dictionary construction process is presented in multiple levels such as phoneme, base syllable and tonal syllable. Both context-dependent and context-independent pronunciation models are also considered. To ensure that the confusion matrices reflect the accent characteristics, both the occurrences of reference observations and the probability of pronunciation variation are taken into account when deciding which transformation pairs should be encoded into the dictionary.

In addition, to verify that pronunciation variation and acoustic deviation are two important but complementary factors affecting the performance of recognizer, maximum likelihood linear regression (MLLR) [6], a well-proven adaptation method in the field of acoustic model was adopted in two modes: separately and combined with PDA.

Compared with [1], which synthesizes the dictionary completely from the adaptation corpus; we augment the process by incorporating obvious pronunciation variations into the accent-specific dictionary with varying weights. As a result, the adaptation corpus that was used to catch the accent characteristics could be comparatively small. Essentially, the entries in the adapted dictionary consist of multiple pronunciations with prior probability that reflect accent variation. In [2], syllable-based context was considered. We extend such context from the syllable to the phone level, even the phone class level. There are several advantages. It can extract the essential variation in continuous speech from a limited corpus. At the same time, it can maintain a detailed description of the impact of articulation of pronunciation variation. Furthermore, tonal changes, as a part of pronunciation variation have also been modeled. In addition, the result we reported has incorporated a language model. In other words, these results could accurately reflect contribution of PDA, MLLR and the combination of two in the dictation application. As we know, a language model could help to recover from some errors due to speakers' pronunciation variation.

Furthermore, most prior work [1][2][5] uses pronunciation variation information to re-score the N-best hypothesis or lattices resulting from the baseline. However, we developed a one-pass search strategy that unifies all kinds of information, including acoustic model, language model and accent model about pronunciation variation, according to the existing baseline system.

In the following section, we will describe in details the PDA algorithm, including the construction of the accent-specific dictionary and its utilization in the search procedure. The experiments and result analysis are given in Section 3. Section 4 summarizes the advantages and weaknesses of this method and promising future directions.

# 2. ACCENT MODELING WITH PDA

Many adaptation technologies based on acoustic model parameter re-estimation make assumption that speakers, even in different regions, pronounce words according to a predefined and unified manner. Error analyses across different accent regions tell us that this is a poor assumption. For example, a speaker from Shanghai probably utters /shi/ as /si/ in the canonical dictionary (such as the official published one based on pronunciation of Beijing inhabitants). Therefore, a recognizer trained according to the pronunciation criterion of Beijing cannot recognize accurately a Shanghai speaker given such a pronunciation discrepancy. The aim of PDA is to build a pronunciation dictionary suited to the accent-specific group in terms of a "native" recognizer. Luckily, pronunciation variation between accent groups presents certain clear and fixed tendencies. There exist some distinct transformation pairs at the level of phones or syllables. This provides the premise to carry out accent modeling through PDA. The PDA algorithm can be divided into the following stages:

The first stage is to obtain an accurate syllable level transcription of the accent corpus in terms of the phone set of the standard recognizer. To reflect factual pronunciation deviation, no language model was used here. The transcribed result was aligned with the reference transcription through dynamic programming. After the alignments, error pairs can be identified. Here, we just consider the error pairs due to substitution error since insertion and deletion errors are infrequent in Mandarin because of the strict syllable structure. To ensure that the mapping pairs were estimated reliably and representatively, pairs with few observations were cut off. In addition, pairs with low transformation probability were also eliminated to avoid excessive variations for a certain lexicon items. According to the amount of accent corpus, context dependent or context independent mapping pairs with different transfer probability could be selectively extracted at the level of sub-syllable, base-syllable or tone-syllable.

The next step is to construct a new dictionary that reflects the accent characteristics based on the transformation pairs. We encode these pronunciation transfer pairs into the original canonical lexicon, and finally a new dictionary adapted to a certain accent is constructed. In fact, pronunciation variation is realized through multiple pronunciations with corresponding weights. Each dictionary entry can be a word with multiple syllables or just a single syllable. Of course, all the pronunciation variations' weights corresponding the same word should be normalized.

The final step is to integrate the adapted dictionary into the recognition or search framework. Much work makes use of PDA through multiple-pass search strategy [2][5]. In other words, prior knowledge about pronunciation transformation were used to re-score the multiple hypotheses or lattice obtained in the original search procedure. In this paper, we adopt a one-pass search mechanism as in WHISPER [3]. Equivalently, the PDA information was utilized at the same time as other information, such as language model and acoustic evaluation. This is illustrated with the following example.

For example: speakers with a Shanghai accent probably uttered "du2-bu4-yi1-shi2" from the canonical dictionary as "du2-bu4-yi1-si2". The adapted dictionary could be as follows:

| | | |
|---|---|---|
| … | | |
| shi2 | shi2 | 0.83 |
| shi2(2) | si2 | 0.17 |
| …. | | |
| si2 | si2 | 1.00 |
| …. | | |

Therefore, scores of the three partial paths yi1→shi2, yi1→shi2 (2) and yi1→si2 could be computed respectively with formulae (1) (2) (3).

$$Score(shi2 \mid yi1) = w_{LM} * P_{LM}(shi2 \mid yi1) + w_{AM} * P_{AM}(shi2) + w_{PDA} * P_{PDA}(shi2 \mid shi2) \tag{1}$$

$$\begin{aligned} Score(shi2(2) \mid yi1) &= w_{LM} * P_{LM}(shi2(2) \mid yi1) \\ &+ w_{AM} * P_{AM}(shi2(2)) + w_{PDA} * P_{PDA}(shi2(2) \mid shi2) \\ &= w_{LM} * P_{LM}(shi2 \mid yi1) + w_{AM} * P_{AM}(si2) \\ &+ w_{PDA} * P_{PDA}(si2 \mid shi2) \end{aligned} \tag{2}$$

$$Score(si2 \mid yi1) = w_{LM} * P_{LM}(si2 \mid yi1) + w_{AM} * P_{AM}(si2) + w_{PDA} * P_{PDA}(si2 \mid si2) \tag{3}$$

Where $P_{LM}, P_{AM}$ and $P_{PDA}$ stand for the logarithmic score of Language Model (LM), Acoustic Model (AM) and Pronunciation variation respectively. $w_{LM}, w_{AM}$ and $w_{PDA}$ are the corresponding weight coefficients and adjusted according to experience.

Obviously, the partial path yi1→shi2 (2) has adopted the factual pronunciation (as / $si2$ /) while keeping the ought-to-be LM, e.g. bigram of ( $shi2 \mid yi1$ ), at the same time, prior information about pronunciation transformation was incorporated. Theoretically, it should outscore the other two paths. As a result, the recognizer successfully recovers from user's pronunciation error using PDA.

# 3. EXPERIMENTS AND RESULT

## 3.1 System and Corpus

Our baseline system is an extension of the Microsoft Whisper speech recognition system [3] that focuses on Mandarin characteristics, e.g. pitch and tone have been successfully incorporated [4]. The acoustic model was trained on a database of 100,000 sentences collected from 500 speakers (train_set, male and female half each, here we only use 250 male speakers) coming from Beijing area. The baseline dictionary is based on an official published dictionary that is consistent with the base recognizer. The language model is tonal syllable trigram with perplexity of 98 on the test corpus. Other data sets are as follows:

- Dictionary Adaptation Set (pda_set): 24 male speakers from Shanghai area, at most 250 sentences or phrases from each speaker;
- Test Set (Test_set) 10 male speakers, 20 utterances from each speaker;
- MLLR adaptation sets (mllr_set): Same speaker set as test sets, at most another 180 sentences from each speaker;
- Accent specific SH model (SH_set): 480 speakers from Shanghai area, at most 250 sentences or phrase from each speaker. (Only 290 male speakers used)

## 3.2 Analysis

2000 sentences from pda_set were transcribed with the benchmark recognizer in term of standard sets and syllable loop grammar. Dynamic programming was applied to these results and many interesting linguistic phenomena were observed.

### 3.2.1 Front nasal and back nasal

Final ING and IN are often exchangeable, while ENG are often uttered into EN and not vice versa. This is shown in Table 1.

| Canonical Pron. | Observed Pron. | Prob. (%) | Canonical Pron. | Observed Pron. | Prob. (%) |
|---|---|---|---|---|---|
| QIN | QING | 47.37 | QING | QIN | 19.80 |
| LIN | LING | 41.67 | LING | LIN | 18.40 |
| MIN | MING | 36.00 | MING | MIN | 42.22 |
| YIN | YING | 35.23 | YING | YIN | 39.77 |
| XIN | XING | 33.73 | XING | XIN | 33.54 |
| JIN | JING | 32.86 | JING | JIN | 39.39 |
| PIN | PING | 32.20 | PING | PIN | 33.33 |
| (IN) | (ING) | 37.0 | (ING) | (IN) | 32.4 |
| RENG | REN | 55.56 | SHENG | SHEN | 40.49 |
| GENG | GEN | 51.72 | CHENG | CHEN | 25.49 |
| ZHENG | ZHEN | 46.27 | NENG | NEN | 24.56 |
| MENG | MEN | 40.74 | (ENG) | (EN) | 40.7 |

**Table 1:** front nasal and back nasal mapping pairs of accent speaker in term of standard phone set.

### 3.2.2 ZH (SH, CH) VS. Z (S, C)

| Canonical Pron. | Observed Pron. | Prob. (%) | Canonical Pron. | Observed Pron. | Prob. (%) |
|---|---|---|---|---|---|
| ZHI | ZI | 17.26 | CHAO | CAO | 37.50 |
| SHI | SI | 16.72 | ZHAO | ZAO | 29.79 |
| CHI | CI | 15.38 | ZHONG | ZONG | 24.71 |
| ZHU | ZU | 29.27 | SHAN | SAN | 19.23 |
| SHU | SU | 16.04 | CHAN | CAN | 17.95 |
| CHU | CU | 20.28 | ZHANG | ZANG | 17.82 |

**Table 2:** Syllable mapping pairs of accented speakers in term of standard phone set.

Because of phonemic diversity, it is hard for Shanghai speakers to utter initial phoneme like /zh/, /ch/ and /sh/. As a result, syllables that include such phones are uttered into syllables initialized with /z/, /s/ and /c/, as shown in Table 2. It reveals a strong correlation with phonological observations.

## 3.3 Result

In this subsection, we report our result with PDA only, MLLR only and the combination of PDA and MLLR sequentially. To measure the impact of different baseline system on the PDA and MLLR, the performance of accent-dependent SI model and mixed accent groups SI model are also present in both syllable accuracy and character accuracy for LVCSR.

### 3.3.1 PDA Only

Starting with many kinds of mapping pairs, we first remove pairs with fewer observation and poor variation probability, and encode the remaining pairs into dictionary. Table 3 shows the result when we use 37 transformation pairs, mainly consisting of pairs shown in Table 1 and Table 2.

| Dictionary | Syllable Error Rate (%) |
|---|---|
| Baseline | 23.18 |
| + PDA (w/o Prob.) | 20.48 (+11.6%) |
| +PDA (with Prob.) | 19.96 (+13.9%) |

**Table 3:** Performance of PDA (37 transformation pairs used in PDA)

### 3.3.2 MLLR

To evaluate the acoustic model adaptation performance, we carry out the MLLR experiments. All phones (totally 187) were classified into 65 regression classes. Both diagonal matrix and bias offset were used in the MLLR transformation matrix. Adaptation set size ranging from 10 to 180 utterances for each speaker was tried. Results are shown in the Table 4. It is shown that when the number of adaptation utterances reaches 20, relative error reduction is more than 22%.

| # Adaptation Sentences | 0 | 10 | 20 | 30 | 45 | 90 | 180 |
|---|---|---|---|---|---|---|---|
| MLLR | 23.18 | 21.48 | 17.93 | 17.59 | 16.38 | 15.89 | 15.50 |
| Error reduction (Based on SI) | -- | 7.33 | 22.65 | 24.12 | 29.34 | 31.45 | 33.13 |

**Table 4:** Performance of MLLR with different adaptation sentences

### 3.3.3 Combined PDA and MLLR

Based on the assumption that PDA and MLLR can be complementary adaptation technologies from the pronunciation variation and acoustic characteristics respectively, experiment combining MLLR and PDA were carried out. Compared with performance without adaptation at all, 28.4% was achieved (only 30 utterances used for each person). Compared with MLLR alone, a further 5.7% was improved.

| # Adaptation Sentences | 0 | 10 | 20 | 30 | 45 | 90 | 180 |
|---|---|---|---|---|---|---|---|
| + MLLR<br>+ PDA | 19.96 | 21.12 | 17.5 | 16.59 | 15.77 | 15.22 | 14.83 |
| Error reduction (Based on SI) | 13.9 | 8.9 | 24.5 | 28.4 | 32.0 | 34.3 | 36.0 |
| Error reduction (Based on MLLR) | - | 1.7 | 2.4 | 5.7 | 3.7 | 4.2 | 4.3 |

**Table 5:** Performance Combined MLLR with PDA

### 3.3.4 Comparison of Different Models

The following table shows the results of different baseline models or different adaptation techniques on recognition tasks across accent regions. It shows that accent-specific model still outperforms any other combination.

| Different Setup | Different Baseline (Syllable Error Rate (%)) | | |
|---|---|---|---|
| | Train_set | BES | SH_set |
| Baseline | 23.18 | 16.59 | 13.98 |
| + PDA | 19.96 | 15.56 | 13.76 |
| + MLLR (30 Utts.) | 17.59 | 14.40 | 13.49 |
| + MLLR + PDA | 16.59 | 14.31 | 13.52 |

**Table 6:** Syllable error rate with different baseline model or different adaptation technologies (BES means a larger training set including 1500 speakers from both Beijing and Shanghai)

### 3.3.5 PDA and MLLR in LVCSR

To investigate the impact of the above strategies on large vocabulary speech recognition, we designed a new series of experiments to be compared with results shown in Table 6. A canonical dictionary consisting of up to 50K items and language model of about 120M were used. The result is shown in Table 7. Character accuracy is not so significant as syllable accuracy shown in Table 6. It is mainly due to the following two simplifications: Firstly, because of the size limitation of dictionary, only twenty confusion pairs were encoded into pronunciation dictionary. Secondly, no probability is assigned to each pronunciation entry at present. However, we still can infer that PDA is a powerful accent modeling method and is complementary to MLLR.

| Different Setup | Different Baseline (Character Error Rate (%)) | | |
|---|---|---|---|
| | Train_set | BES | SH_set |
| Baseline | 26.01 | 21.30 | 18.26 |
| + PDA | 23.64 | 20.02 | 18.41 |
| + MLLR (30 Utts.) | 21.42 | 18.99 | 18.51 |
| + MLLR + PDA | 20.69 | 18.87 | 18.35 |
| + MLLR (180 Utts.) | 19.02 | 18.60 | 17.11 |

**Table 7:** Character error rate with different baseline model or different adaptation technologies (BES means a larger training set including 1500 speakers from both Beijing and Shanghai)

## 4. CONCLUSION

This paper presents in detail the accent modeling technology using PDA and its successful integration into one-pass search framework of Whisper. Acoustic-based adaptation, MLLR and its combination with PDA were also investigated. When PDA, MLLR (30 utterances used) and a combined method were used in the large vocabulary Mandarin speech recognition task, the relative error rate was reduced 13.9%, 24.1% and 28.4% respectively.

Future work will investigate the inclusion of PDA into the training process, especially when corpora from different accent regions were put together to train a general model. It should help strengthen the distinguishability of model.

## 5. ACKNOWLEDGEMENTS

We would like to thank Dr. Mei-Yuh Hwang, Microsoft Research in Redmond for providing great help on Whisper. We also wish to thank Dr. M. J. Li for supplying both syllable and word language model. We would like to acknowledge Z. Y. Qiao for implementing algorithms like extended dynamic programming and dictionary modification. We would like to express our gratitude to Ben Walters for revising the paper carefully.

## 6. REFERENCES

[1] J. J. Humphries and P.C. Woodland, “The Use of Accent-Specific Pronunciation Dictionaries in Acoustic Model Training”, page 317-320, Vol.1, *Proc. ICASSP 1998*, Seattle, USA.

[2] M. K. Liu, B. Xu, T. Y. Huang, Y. G. Deng, C. R. Li, “Mandarin Accent Adaptation Based on Context-Independent/Context-Dependent Pronunciation Modeling”, page 1025-1028, Vol.2, *Proc. ICASSP 2000*, Turkey.

[3] X. D. Huang, A. Acero, F. Alleva, M. Y. Hwang, L Jiang, M. Mahajan, “Microsoft Windows highly intelligent speech recognizer: Whisper.” page 93-96. Volume 1, *Proc. ICASSP 1995*.

[4] E. Chang, J. L. Zhou, C. Huang, S. Di, K. F. Lee, “Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones”, to appear in *Proc. ICSLP 2000*.

[5] M. D. Riley and A. Ljolje, “Automatic Generation of Detailed Pronunciation Lexicon”, *Automatic Speech and Speaker Recognition: Advanced Topics*. Kluwer. 1995.

[6] C. J. Leggetter, P. C. Woodland, “ Maximum likelyhood linear regression for speaker adaptation of continuous density hidden Markov models”, pp. 171-185, *Computer Speech and Language,* Vol. 9, No. 2, April, 1995.