ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

MANDARIN SPEECH RECOGNITION FOR NONNATIVE SPEAKERS BASED ON PRONUNCIATION DICTIONARY ADAPTATION

Jian Yang, Peishan Wu, Dan Xu

School of Information Science and Engineering, Yunnan University, 650091
nxryang@126.com, patient_woo@sina.com, danxu@vip.sina.com

ABSTRACT

Various techniques, such as acoustic model adaptation and pronunciation adaptation, have been reported to improve the recognition of nonnative or accented speech. In this paper, we propose to analyze the regular pairs of the pronunciation variation of the nonnative Mandarin speech spoken by *Dai*, *Lisu* and *Naxi* speakers from Yunnan. According typical pronunciation variations of these 3 accents, the more than one pronunciation for a part of words (i.e. tonal syllables or characters) have been inserted in the standard Mandarin pronunciation dictionary. The experiments show that an improvement is reached with the new dictionary and a simple 2-gram language model for all kinds of nonnative speakers.

Index Terms— speech recognition, accented Mandarin, nonnative speaker, pronunciation variations, pronunciation dictionary adaptation

1. INTRODUCTION

Over the past decade, there have been tremendous efforts on large vocabulary continuous speech recognition for Chinese. Among the multifarious Chinese dialects, Mandarin (or Putonghua, PTH) has received the most research and commercial interests, given its huge speaker population and the unique role as the official standard of spoken Chinese. Nevertheless, there has been an obvious and ever increasing demand for speech recognition technology that can deal with Chinese dialects and nonnative Mandarin, spoken by foreigner or the speakers from the minority areas in China. Generally, there is about 20-50% reduction when using a robust PTH recognizer to recognize accented Mandarin speech. For example, in NIST 1997 Broadcast News evaluation task [1], compared with the 21.38% character error rate (CER) of standard PTH, the recognizing results of read style and spontaneous speech achieve 61.89% and 72.17% when recognizing Shanghai's PTH (or Wu dialectal Chinese). The reason is the general baseline recognizer is based on the standard PTH, but the accented PTH has obvious pronunciation difference compared with Mandarin.

Yunnan is a big nation culture province which has twenty five minorities. Most minorities have their own native language. And, their settles are almost everywhere in Yunnan. Because these special reasons of geography, culture and environment, the Mandarin spoken by the minority speakers is obviously different from the standard Mandarin. So, studying the speech recognition of this accented Mandarin is very meaningful.

Various techniques, such as acoustic model adaptation and pronunciation adaptation, have been reported to improve the recognition of nonnative or accented speech [9]. Aiming at this problem, in this paper we us a technique of modeling accent-specific pronunciation variations through pronunciation dictionary adaptation (PDA). The aim is to get regular pairs of the pronunciation variation by combining the rule-based data-driven (DD) method with the experts' knowledge. Then, we can construct the Mandarin multi-pronunciation dictionary for different accents of nonnative speaker. Taking the accented Mandarin of minority language in Yunnan province as an example, firstly, baseline hidden Markov models (HMM) were trained by using Chinese 863 project standard Mandarin corpus. Secondly, the nonnative speech data spoken by *Dai*, *Lisu* and *Naxi* speakers from Yunnan was transcribed with the baseline HMMs. In addition, the transcribed result was aligned with the reference transcription through dynamic programming (DP). After calculating the confusion matrix of recognition results, we analyze the error pairs due to substitute error at the level of base syllables, initials and finals, respectively. According special rules, some typical pronunciation variations of minorities' accents in Yunnan have been inserted in the standard Mandarin dictionary. These new dictionaries were integrated into the speech recognition framework to get better performance. The experiment results show that an improvement is reached with the new dictionary and a simple 2-gram language model for all kinds of nonnative speakers

This paper is organized as follows. The principle of the pronunciation dictionary adaptation is presented in section 2. In section 3, we describe the speech corpus. The process to construct the multi-pronunciation dictionary is given in section 4. The baseline system and experiment results are

978-1-4244-2942-4/08/$25.00 (c) 2008 IEEE

given in section 5. Section 6 concludes with summary of our work.

2. PRONUNCIATION DICTIONARY ADAPTATION

When using the standard dictionary to recognize, the general situation is as follows [3]:

Generally, a system of large vocabulary independent speaker continuous Mandarin speech recognition based on HMM uses n-gram syntax and lexicon construction searching tree. Recognition is an action to find paths with the highest score in all possible paths of searching tree by using Viterbi algorithm. Each path's score is calculated by accumulated possibility of syntax and acoustic possibility in searching progress like equation (2-1).

$$Score = w_{LM} \log \mathrm{P}_{LM} + w_{AM} \log \mathrm{P}_{AM} \qquad (2\text{-}1)$$

Where, P_{LM} is syntax possibility, and P_{AM} is acoustic possibility, w_{LM} and w_{AM} is weighted coefficients.

For example, using 2-gram syntax recognizes the word: "欢迎". In standard dictionary, the terms of "欢" and "迎" are:

欢：huan1　　　迎：ying2

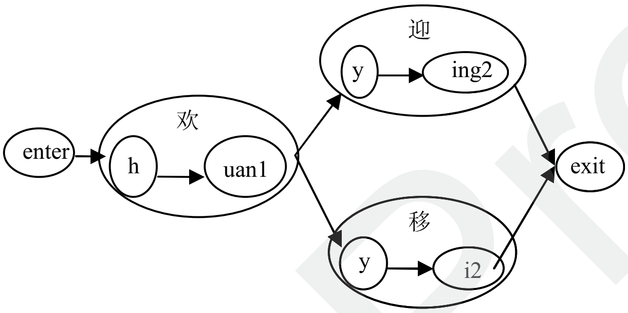

Figure 1: two possible recognition paths of word "欢迎"

In Figure 1, we can observe two possible recognition paths about "欢迎". In both paths, the P_{LM}(迎|欢) is very large. If the pronunciation is standard, the acoustic possibility of "移" also is very large, and the score is highest to get the correct recognizing result. But if the speaker is one with *Naxi* accent of who usually speak /ying/ as /yi/, such as "移", the two paths are too hard to distinguish. Though the path "欢→移" has a quite lower syntax possibility (because it is not a Chinese word), its acoustic possibility is very high. So, scores of the two paths is similar, moreover, we couldn't tell apart them by adjusting w_{LM} and w_{AM}. We maybe get the wrong result.

To resolve above problems, according to the pronunciation variations' regularity of nonnative speakers, we must construct new dictionary with multiple pronunciation entries as following steps [3][4]:

- Taking a standard PTH recognizer as the baseline system.
- Preparing nonnative speech data, and transcribing the Chinese sentences with standard pronunciation syllable (i.e. pinyin), and that is called base-form.
- Using baseline system to recognize the nonnative speech data, the recognition results are syllable sequence, named as surface-form.
- The surface-form syllable sequence was forced aligning with the base-form syllable sequence through dynamic programming (DP).
- Computing the confusion matrix between base-form and surface-form.
- Analyzing these errors distribution of syllable, then find some typical syllable pairs to be extended in the standard dictionary.

So, the new dictionary is one that aims at nonnative speech data, and each word in it may has more than one pronunciation, including standard and nonnative accent syllables.

Taking "欢迎" as an example, in *Naxi* dictionary, the term of "迎" is:

ying2：y　ing2 (standard pronunciation entry)
ying2：y　i2　(*Naxi* accent pronunciation entry)

The syllable /ying2/ is added a pronunciation entry, and the standard pronunciation is as usual.

3. SPEECH CORPUS

In this study, two native speech corpora and three nonnative speech corpora shown in Table 1 are used. The native Mandarin speech waveforms which use to train HMM models are extracted from the Mandarin Dictation Corpora supported by China National Hi-Tech Project 863. We used the sentences collected from 87 speakers (38 males and 49 females) to train the baseline system. Otherwise, as a part of the Microsoft Research Mandarin Speech Toolkit [5], the sentences collected from 100 male speakers are used to test the baseline system.

Corpus	Accent	Speakers	Sentences
Project 863	Native	87	39800
MS Toolkit	Native	100	19690
Naxi	Nonnative	19	13051
Dai	Nonnative	17	8861
Lisu	Nonnative	12	7235

Table 1: Speech Corpus Overview

The nonnative Mandarin speech waveforms are extracted from the Linguistic Minorities Accents Mandarin Speech Corpus (LMAMSC), which collected by our lab. In the LMAMSC, the Chinese sentence prompts were the same sentences as the Project 863' Mandarin Dictation Corpora. Recordings were made with a high-quality head-mounted microphone in a quiet laboratory environment. The data was digitized at 16bit per sample and a sampling rate of 16kHz. The all speakers are from national minority areas in Yunnan

978-1-4244-2942-4/08/$25.00 (c) 2008 IEEE

and their native languages are not Chinese. The nonnative accents are obvious when they speak Mandarin.

4. MULTI-PRONUNCIATION DICTIONARY

How to get the pronunciation variations is the most important problem. We can study it in two ways: knowledge-based [6] and data-driven [7].

4.1 Confusion Matrix

Firstly, let's see the recognition result of one sentence as:

```
LAB: wo men da duo shu ling dao     bu   gong
REC: gu wei ba dou     ling dao zhe fou  gong
```

In the result, LAB is base-form of the sentence, and REC is the surface-form of it. The result has been forced aligning by DP.

Just thinking about the meaningful substitute error, the mapping pairs are:

Syllable level:
/wo/→/gu/, /men/→/wei/, /da/→/ba/,
/duo/→/dou/, /bu/→/fou/.
Initial level:
/w/→/g/, /m/→/w/, /d/→/b/, /b/→/f/.
Final level:
/o/→/u/, /en/→/ei/, /uo/→/ou/, /u/→/ou/.

After analyzing all initials, finals or syllables' error pairs, we can get the confusion matrixes of initials, finals or syllables for one accent. Table 2 is a part of a final confusion matrix for Naxi accent, which can prove that /ing/ is always pronounced /i/ by Naxi speakers.

		REC			
		a	o	i	ing
LAB	a	59%	1%	3%	0%
	o	4%	32%	6%	0%
	i	1%	0%	81%	1%
	ing	3%	0%	41%	14%

Table 2: Final confusion matrix for Naxi accent (part)

4.2 Expert's Knowledge

We use the DD method which integrated expert's knowledge to study the regularity of the pronunciation variation. And, the expert's knowledge is from literature [2] in which the experts concluded a series of regular errors by testing minority students' PTH. There are 3 accents' typical errors in PTH test as following:

Dai accent: /ü/ pronounced as /i/, like "鱼" read as "一", "女" as "你"； /eng/, /ing/ pronounced as /en/, /in/， like "声" as "身"， "晴" as "秦" and so on.

Naxi accent: /z/, /c/, /s/ pronounced as /zh/, /ch/, /sh/; /n/ and /l/ couldn't be told apart when speaking; /in/ and /ing/ pronounced as /i/; /ie/ pronounces as /i/; /ai/ pronounced as /a/ and so on.

Lisu accent: /shi/ pronounced as /si/; /na/ pronounced as /la/; /neng/ pronounced as /leng/; /ang/ and /an/ pronounced as /a/; /iong/ pronounced as /io/ or /iu/; /ian/ and /iang/ pronounced as /i/ or /in/; /ong/ pronounced as /o/ and so on.

4.3 Dictionaries' Constructing

After getting syllable's error pairs, we can put them into the standard dictionary. In [8], it was proved that when the new dictionary's scale is 1.3 times of the standard one, the effect is the best. So, we must set some rules to decide which term can get more entries in the standard dictionary. The aim is not only present the regular pronunciation variation but also avoid more confusion.

According the appearing frequency of syllables which have mapping error in adaptive speech corpus, we selected syllables that have higher frequency to become terms with an accent pronunciation entry. At last we got 3 accents multiple pronunciation dictionaries: Dai dictionary is probably 1.26 times of standard one. Naxi and Lisu dictionaries are about 1.25 times.

5. EXPERIMENT RESULTS

All recognition experiments described in this paper use the HTK Toolkit. The acoustic models of the baseline system are trained on the native Mandarin corpora data. The basic acoustic units of the baseline system are composed of 27 initials and 157 tonal finals. The feature used is a 39order feature vector, consisting of 12 MFCCs (Mel Frequency Cepstral Coefficient), energy, and their first and second order differences. The feature vector is calculated using a window size of 25ms and a moving size of 10ms. After the monophone models are trained, all possible triphone expansions based on the full syllable dictionary are performed. Final, we use the decision tree based clustering capability of the HTK toolkit to tie similar states of triphones to each other. The word correct rate using the baseline system on the native test set is list in the Table 3.

base	lm_base	tonal	lm_tonal
82.4%	93.3	63.8	91.3

Table 3: Recognition results on the native test set

Based on the above baseline system, we tested the pronunciation dictionary adaptation for 3 accents with or without a simple 2-gram language model. Table 4~6 list the experiments' results with the word correct rate (WCR). Where, 'tonal' means tonal syllable and 'base' means the base syllable after ignoring tone. The prefix 'lm' means language model.

From the results we can get the conclusions like that: combining multiple the pronunciation dictionaries with a language model can improve the nonnative speech recognition. The best improvement is 4.36% and the worst only is 0.2%. It is also proved that the multiple pronunciation entries without any language model could reduce the correct rate of accented speech recognition.

Speaker	Recognition Method	Standard Dictionary	New Dictionary
Male Mdy000	base	23.76%	20.99%
	lm_base	40.32%	41.41%
	tonal	10.31%	9.39%
	lm_tonal	35.05%	35.32%
Female Fdy005	base	20.98%	16.69%
	lm_base	23.82%	24.16%
	tonal	11.38%	9.15%
	lm_tonal	18.92%	19.93%

Table 4: Recognition results of Dai accent

Speaker	Recognition Method	Standard Dictionary	New Dictionary
Male Mls003	base	32.41%	30.43%
	lm_base	53.16%	53.75%
	tonal	22.53%	20.95%
	lm_tonal	50.00%	50.20%
Female Fls004	base	29.44%	26.69%
	lm_base	56.27%	57.72%
	tonal	13.49%	11.39%
	lm_tonal	50.25%	51.99%

Table 5: Recognition results of Lisu accent

Speaker	Recognition Method	Standard Dictionary	New Dictionary
Male Mnx000	base	42.86%	32.47%
	lm_base	57.61%	60.85%
	tonal	24.49%	17.25%
	lm_tonal	49.26%	53.62%
Female Fnx000	base	33.84%	27.09%
	lm_base	47.24%	49.24%
	tonal	17.98%	15.33%
	lm_tonal	41.27%	43.40%

Table 6: Recognition results of Naxi accent

6. CONCLUSION

Nonnative or accented PTH is an important problem in Mandarin speech recognition studying and applying. Due to this speech appears significantly more difficultly to recognize than native Mandarin, this paper's work is just a beginning. There are many problems that remain to be investigated, for examples, deep studying dictionary pruning rule, construct speaker-specific dictionary through recognizing by experts, using more perfect language model, etc.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge the supports from grants of The Applied Basic Research Programs of Yunnan Province (No.2007F169M) and of The Midlife and Young Backbone Teacher's Training Programs of Yunnan University (2006.1-2008.12).

8. REFERENCES

[1] NIST, "The 1997 Hub-4NE evaluation plan for recognition of Broadcast News, in Spanish and Mandarin", http://www.nist.gov/speech/tests/bnr/hub4ne_97/current_plan.htm, 1997.

[2] 云南少数民族双语教学研究课题组, "云南少数民族双语教学研究", 昆明：云南民族出版社, 1995.

[3] 潘复平, 赵庆卫, 颜永红, "一种用于方言口音语音识别的字典自适应技术 (Pronunciation Dictionary Adaptation Based Accent Modeling for Large Vocabulary Continuous Speech Recognition) ", *计算机工程与应用*, pp.5-6, 2005 年 23 期.

[4] 刘明宽, 徐波, 黄泰翼, 胡伟湘, "音节混淆字典以及在汉语口音自适应中的应用研究 (Study on syllable confusion dictionary and putonghua accent adaptation)". *声学学报*，pp. 53-58, 2002 年 01 期.

[5] E. Chang, Y. Shi, J. L. Zhou and C. Huang, "Speech Lab in a Box: A Mandarin Speech Toolbox to Jumpstart Speech Related Research", *Eurospeech 2001*, pp.2799-2802 Aalborg, Denmark, 2001.

[6] V. Hoste, W. Daelemans, S. Gillis, "Using rule-induction techniques to model pronunciation variation in Dutch", *Computer Speech and Language*, Vol. 18, Issue 1, pp.1-23, Jan. 2004.

[7] M. Wester, "Pronunciation modeling for ASR - knowledge-based and data-derived methods", *Computer Speech and Language*, Vol. 17, Issue 1, pp. 69-85, Jan. 2003.

[8] 刘林泉, "基于小数据量的方言背景普通话语音识别声学建模研究 (Research on A Small Data Set Based Acoustic Modeling for Dialectal Chinese Speech Recognition) ", *博士论文*, 北京：清华大学计算机科学与技术系，2007.

[9] R. Sproat, F. Zheng, Gu L, et al., "Dialectal Chinese Speech Recognition: Final Report", CLSP Summer Workshop, http://www.clsp.jhu.edu/ ws2004/, Nov 15, 2004.