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

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

AN ACOUSTIC-PHONETIC ANALYSIS OF LARGE VOCABULARY CONTINUOUS MANDARIN SPEECH RECOGNITION FOR NON-NATIVE SPEAKERS

Jian Yang, Yuanyuan Pu, Hong Wei, Zhengpeng Zhao

Institute of Information Science, Yunnan University, Kunming, 650091
jianyang@ynu.edu.cn, oldsan@21cn.com

ABSTRACT

This paper addresses non-native accent issues in large vocabulary continuous speech recognition. We propose to analyze the transformation rules of non-native Mandarin speech spoken by native speakers of *Naxi* and *Dai* in *Yunnan* at the level of initials and finals. Firstly, baseline HMM models are trained using the project 863' standard Mandarin corpus to test their performance on non-native speech recognition. Secondly, the non-native speech data is transcribed based on the baseline HMM models. In more detail, we analyze the error recognition rates of all initials and all finals, and their typical substitute error. The results obtained from our experiments might be useful for adapting a native speaker ASR system to model non-native accented data.

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*) 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 non-native Mandarin, spoken by foreigner or the speakers from the minority areas in China. The reasons are at least two-fold. Firstly, more and more foreigners learn Chinese and speak Mandarin with foreign accent. Secondly, most of the national minorities in China, such as *Naxi*, *Dai*, *Zang* etc., have their languages, so they speak Mandarin with their native language accents. Non-native specific investigation is not only justifiable but also necessary for the advancement of Chinese speech recognition technology.

Despite the large progress in fields like large vocabulary continuous speech recognition or noise robustness, recognition accuracy has been observed to be drastically lower for non-native speakers of the target language than for the native ones [1]. One reason is because the non-native speakers' pronunciation differs from those native speakers' pronunciation observed during system training. A number of methods for handling non-native speech in speech recognition have been proposed. The most straightforward approach is to use the non-native speech from the target language spoken by the group of non-native speakers for recognizer training [2]. However the problem of this method is that the non-native speech data is only rarely available. Another approach is to apply accent modeling technology such as PDA (pronunciation dictionary adaptation) [3].

Conventional acoustic model adaptation technologies assume that speakers pronounce words according to a predefined and unified manner, which is not always valid for non-native accented speech. For example, a *Naxi* accent speaker probably utters syllable /yin/ as /yi/ in the canonical dictionary. Therefore, a recognizer trained on the pronunciation criterion of standard Mandarin cannot accurately recognize speech from a non-native speaker. Fortunately, pronunciation variation between accent groups usually presents certain clear and fixed tendency. There exist some distinct transformation pairs at the level of phones or syllables.

In this paper, we propose to analyze the transformation rules of non-native Mandarin speech spoken by native speakers of *Naxi* and *Dai* in *Yunnan* at the level of initials and finals. Firstly we train baseline HMM models using standard Mandarin speech to test their performance on non-native speech recognition. Secondly, the non-native speech data is transcribed based on the baseline HMM models. In more detail, we analyze the error recognition rates of all initials and all finals, and their typical substitute error.

This paper is organized as follows. The speech corpus is presented in section 2. In section 3, we describe the baseline system of our experiments. Detailed experiments and results on non-native speech recognition are given in section 4. Section 5 concludes with summary of our work.

2. SPEECH CORPUS

In this study, two native speech corpora and two non-native speech corpora shown in Table 1 are used.

Corpus	Accent	Speakers	Sentences
Project 863	Native	87	39800
MS Toolkit	Native	100	19690
Naxi	Non-Native	19	13051
Dai	Non-Native	17	8861

Table 1. Speech Corpus Overview

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* [4], the sentences collected from 100 male speakers are used to test the baseline system.

The non-native 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 16bits per sample and a sampling rate of 16kHz. The all speakers are from minority areas in *Yunnan* and their native languages are not Chinese. The non-native accents are obvious when they speak Mandarin.

3. BASELINE SYSTEM

3.1. Baseline Acoustic Models

All recognition experiments described in this paper use the HTK Toolkit [5].

The acoustic models of the baseline system for native Mandarin speech are trained on the native corpora data. The whole training procedure closely follows the one outlined in the *Microsoft Mandarin Speech Toolbox* [4]. The feature used is a 39order feature vector, consisting of 12 cepstral coefficients, energy, and their first and second order differences. The feature vector is calculated using a window size of 25ms and a step size of 10ms. The whole training procedure should be divided into two stages: monophone and triphone. In each stage, there are always two steps, which are repeated iteratively: estimation and realignment. The process begins with the training of the monophone models, followed by training of the triphone models. For predicting unseen triphone in recognition, the parameter of tied-state triphone should be estimated.

In this study, we train the acoustic models based on syllables. The basic acoustic units used for recognition are shown in Table 2. The baseline acoustic model was designed to be tonal since tone is an important feature of the Chinese language.

Initial	b, c, ch, d, f, g, ga, ge, ger, go, h, j, k, l, m, n, p, q, r, s, sh, t, w, x, y, z, zh
Tonal Final	a(1-5), ai(1-4), an(1-4), ang(1-5), ao(1-4), e(1-5), ei(1-4), en(1-5), eng(1-4), er(2-4), i(1-5), ia(1-4), ib(1-4), ian(1-5), iang(1-4), iao(1-4), ie(1-4), if(1-4), in(1-4), ing(1-4), iong(1-3), iu(1-5), o(1-5), ong(1-4), ou(1-5), u(1-5), ua(1-4), uai(1-4), uan(1-4), uang(1-4), ui(1-4), un(1-4), uo(1-5), v(1-4), van(1-4), ve(1-4), vn(1-4)

Table 2. Initial and tonal final units [4]

3.2. Training

After the monophone models are trained, all possible triphone expansions based on the full syllable dictionary are performed. This results in a total of 270,998 triphones. Out of these triphones, 24,127 triphones actually occur in the training corpus.

After performing several iterations of embedded reestimation, we use the decision tree based clustering capability of the HTK toolkit to tie similar states of triphones to each other. After clustering, the number of unique Gaussian mixtures is reduced to 16,112. We then use the HTK toolkit's Gaussian splitting capability to incrementally increase the number of Gaussians mixture to 8.

4. EXPERIMENTS

After the acoustic models of the baseline system for native Mandarin speech are trained, we perform a set of recognition experiments and compare the recognition results at the level of syllable, initials and finals.

4.1. Correct Rates of Syllables Recognition

The baseline syllable recognition results on the test set of 19,690 sentences collected from 100 male native speakers in the MS Toolkit is shown in Table 3 and the results on two test sets of non-native accented speech, 13051 sentences from 19 *Naxi* accent speakers and 8861 sentences from 17 *Dai* accent speakers, is shown in Table 4. To reflect true pronunciation deviation, no language model is used here. As expected, the recognition accuracy of baseline models is drastically low for the non-native speakers of the Chinese than for the native ones.

Base Syllable Correct	Tonal Syllable Correct
84.48%	65.50%

Table 3. Recognition Results on the Native Test Set

Accent	Base Syllable Correct	Tonal Syllable Correct
Naxi	40.42%	24.39%
Dai	43.89%	27.63%

Table 4. Recognition Results on the No-Native Test Sets

4.2. Correct Rates of Initials Recognition

With the HTK performance analysis tool *HResult*, the obtained transcriptions are aligned with the reference ones through dynamic programming. Then error pairs can be identified. Here only initial and final substitution errors are considered. For example, the following text is included in the output file of *HResult* with base syllable.

```
LAB: wo men da ling
REC: gu wei    li
```

Where, there are 2 initial substitution errors: "/w/->/g/", "/m/->/w/", and 3 final substitution errors: "/o/->/u/", "/en/->/ei/", "/ing/->/i/".

4.2.1. Naxi Accented Speech

The recognition correct rates and typical substitutions of the all initials on the *Naxi* accented test set are shown in Table 5, where "null" denotes the null-initial (*Ling Sheng Mu*). If the rate of one substitution error is higher than 10%, we call it typical substitution. Here the correct rates of some initials are very low, but none of the substitution error rates is more than its correct rates. For example, the rate of correctly recognizing /ch/ as /ch/ is 33.88%, and the rate of substitution: "/ch/->c" is 19%.

Initial	Corr (%)	Typical Substitutions Errors and Rates (%)
n	23.47	d (16), l (15), null (18)
ch	33.88	c (19)
c	34.84	ch (13)
m	39.94	b (15), l (11), null (11)
r	46.10	l (11)
zh	47.84	z (22)
z	50.09	zh (13)
p	50.76	t (11), null (10)
k	50.82	t (10)
f	51.68	q (10)
t	52.15	d (16)
q	53.31	
j	54.93	q (10)
b	57.85	
s	58.47	sh (11)
sh	61.86	s (14)
x	64.15	
d	64.91	
l	70.22	null (10)
g	70.48	
h	75.31	
null	79.26	

Table 5. Recognition Correct Rates of Initials on the Naxi Accented Test Set

4.2.2. Dai Accented Speech

The recognition correct rates and typical substitutions of the all initials on the *Dai* accented test set are shown in Table 6. In this table the correct rates of some initials are very low, and the rate of substitution: "/zh/->z" is more than recognition correct rate of initial /zh/.

Initial	Corr (%)	Typical Substitutions Errors and Rates (%)
zh	25.79	z (28)
ch	40.35	c (28)
b	42.01	f (13)
z	45.45	s (12), zh (10)
r	48.33	null (15)
c	50.07	ch (19)
d	52.23	
j	53.90	q (20)
s	54.75	sh (27)
n	55.57	l (13)
t	55.95	
g	56.48	k (11)
p	59.19	t (14)
l	59.56	
sh	61.46	s (21)
f	61.99	h (10)
k	65.72	
m	72.33	
q	72.76	x (11)
null	73.64	
x	75.02	
h	80.97	

Table 6. Recognition Correct Rates of Initials on the Dai Accented Test Set

4.3. Correct Rates of Finals Recognition

4.3.1. Naxi Accented Speech

The recognition correct rates of many finals on the Naxi accented test set are very low, as shown in table 7. And the correct rates of some finals are lower than their typical substitutions rates, such as /ueng/, /ing/, /en/ etc.

Final	Corr (%)	Typical Substitutions Errors and Rates (%)
ueng	3.88	u (16), ong (12)
ing	13.22	i (51), in (11)

eng	14.07	en (13)
vn	14.76	v (27), i (24)
in	15.00	i (50)
iang	18.74	ian (13), iao (10)
un	19.06	ui (32), i (11)
ang	19.89	a (22), an (15)
uang	24.14	uan (15), ua (14)
uai	25.75	ua (19), uan (14)
an	27.36	a (22), ang (10)
o	27.74	u (20), uo (26)
ie	27.85	i (38)
iong	31.12	iu (32)
en	32.08	ei (11)
ian	32.26	ie (17)
uan	33.57	ua (13)
ia	34.04	iao (31)
ai	36.75	a (17), an (15)
ve	37.11	i (21), v (15)
van	39.18	ve (16)
ou	41.63	e (15), u (10)
ong	43.73	u (19), ou (13)
ei	48.73	
a	59.27	ao (18)
uo	60.14	u (14)
iu	60.64	
ao	61.78	ou (11), a (10)
er	64.43	
iao	65.25	
v	65.67	i (17)
ui	66.37	
u	66.80	
ua	68.10	
e	68.12	
i	82.96	

Table 7. Recognition Correct Rates of All Finals on the Naxi Accented Test Set

4.3.2. Dai Accented Speech

The recognition correct rates and typical substitutions of a part of the all finals on the *Dai* accented test set are shown in Table 8.

Final	Corr (%)	Typical Substitutions Errors and Rates (%)
v	18.99	i (58)
ve	19.54	ie (22), i (20)
vn	20.58	in (31), ing (24)
er	22.48	e (27)
eng	24.57	en (35)
o	26.29	u (34), uo (13)
van	27.48	ian (40)
ueng	34.34	ou (10)
uo	38.19	u (23) , ou (11)
ie	38.70	i (25)
ou	40.38	e (16)
ing	45.74	in (32)
in	48.71	ing (27)
ang	49.75	an (17)

Table 8. Recognition Correct Rates of Finals on the Dai Accented Test Set

5. SUMMARY

In this paper, a new non-native accents speech corpus of dictation Mandarin for LVCSR has been described. Based on the corpus, we analyze the error recognition rates of all initials and all finals, and their typical substitute error. The results obtained from our experiments might be useful for adapting a native speaker ASR system to model non-native accented data. While this speech appears significantly more difficultly to recognize than native Mandarin, we expect performance on this task to benefit from progress in speaker adaptation in general with more non-native accents, such as *Bai, Yi, Zang, Lisu* etc. in *Yunnan*.

6. ACKNOWLEDGEMENTS

The authors wish to acknowledge Eric Chang for providing the Microsoft Research Mandarin Speech Toolkit and the set of CD-ROMs of speech corpus. Thanks Xuetao Li and Maopin Wen for testing the program which computes confusion matrix.

The research is supported by *National Natural Science Foundations* (No. 60265001) and *Research Foundations of Education Bureau of Yunnan Province* (No. 03Y176A).

7. REFERENCES

[1] L. Mayfield Tomokiyo, "Recognizing Non-Native speech: Characterizing and Adapting to Non-Native Usage in Speech Recognition", *Ph.D. Thesis*, Carnegie Mellon University, 2001.

[2] U. Uebler and M. Boros, "Recognition of Non-Native German Speech with Multilingual Recognizers", *Proc. Eurospeech*, Volume 2, pp.911-914, Budapest, 1999.

[3] Chao Huang, Eric Chang, Jianlai Zhou and Kai-Fu Lee, "Accent Modeling Based on Pronunciation Dictionary Adaptation for Large Vocabulary Mandarin Speech Recognition", *ICSLP 2000*, Beijing, Oct. 2000.

[4] Eric Chang, Yu Shi, Jianlai Zhou and Chao Huang "Speech Lab in a Box: A Mandarin Speech Toolbox to Jumpstart Speech Related Research", *Eurospeech 2001*, Aalborg, Denmark, 2001.

[5] http://htk.eng.cam.ac.uk.