Phonetic Research on Accented Chinese in Three Dialectal Regions: Shanghai, Wuhan and Xiamen

Aijun LI, Qiang FANG, Ziyu XIONG

Institute of Linguistics, Chinese Academy of Social Sciences
{ liaj,fangqiang,xiongzy }@cass.org.cn

10.21437/Interspeech.2006-245

Abstract

There are 10 major dialects in China. Most people in dialectal regions are bilingual speakers, i.e. native dialect and Mandarin. Although lots of people can speak Mandarin, they speak it with different accents (called regional accented Chinese in this paper) depending on how well they grasp the language. In this study, we categorize the regional accented Chinese into 3 levels of accents according to phonetic annotation and subjective evaluation on a regional accented speech corpus of three regions: Shanghai, Wuhan and Xiamen. Three accent evaluation methods, namely segmental annotation, clustering on phonetic annotation and subjective evaluation, are compared based on phonetic error rates. The results show that objective evaluation score based on segmental pronunciation is higher than subjective evaluation for the same speaker. This implies that supra-segmental features play an important role in rating accent degree and segmental features alone are not enough for objective evaluation. In accent level criterion, the errors from prosodic and segmental aspects are not equal and the percentage of these two parts are various for different regional speakers. The result is helpful for machine evaluation, L2 teaching and acquisition.

Index Terms: Chinese regional accent, accent level evaluation

1. Introduction

Nowadays Standard Chinese (SC, also called Putonghua or Mandarin) is widely used all over China on almost every activity from broadcast news to commercial trades. People from different dialectal areas might not be able to communicate with each other simply because the differences among the dialects are so significant. Popularizing Standard Chinese in dialectal regions, as well as the evaluating the learners' spoken skills or accent levels are a long term policy being carried on by China's Ministry of Education.

The accented Chinese can be regarded as the inter-language in L2 acquisition theory. In the area of accented Chinese study, most researchers have focused on qualitative and phonological description on dialects in popularizing the Standard Chinese. Although some contrastive studies between two dialects or between a dialect and the SC have been published, few investigations have been carried out from the perspective of phonetics. In recent years, we have been doing phonetic analysis on accented Chinese from the view of language teaching and objective evaluation on accent levels. We wish to help provide an auxiliary evaluation tool to reduce the labor waste in evaluation tasks. In paper [2], we investigated the difference between Shanghai accented Chinese and SC, giving some good advice on language teaching and objective evaluation.

The spoken skill test called PSC (Putonghua Shuiping Ceshi) is an accent level evaluation on SC held by China's Ministry of Education. The accent levels are categorized into 3 major levels - L1: light, L2: mid, and L3: heavy - within each level there are 2 sub-levels A and B. So there are altogether 6 levels. Test materials for each speaker are randomly selected from a test bank provided by the Ministry of Education, including 5 parts: monosyllables, multi-syllabic words, sentences, a narrative, and a piece of spontaneous speech on a specific topic.

The professional tester rates the speakers' accent levels according to the expressiveness of their tones, intonations, and rhythm together with the correctness rate of the initials and finals of syllables.

For each accent category, a simple rating criterion is described by PSC as follows:

L1-A: the best, with an error rate on pronunciation, lexical usage and fluency lower than 3%, i.e. the speaker on this level is qualified for a national radio announcer, displaying good reading and speaking skills in beautiful intonation and rhythm with standard pronunciation and very high lexical and grammatical accuracy.

L1-B: the second best, with an error rate less than 8%, i.e. fluent expression with seldom tonal or segmental errors.

L2-A: error rate less than 13%; segments and tones acceptable with fluent intonation; few lexical and grammatical errors.

L2-B: error rate less than 20%; some tonal errors and more segmental errors; occasional use of dialectal words or grammar.

L3-A: error rate less than 30%. Numerous tonal and segmental errors; usually pronouncing with native dialect.

L3-B: error rate less than 40%; mostly dialectal.

But this criterion on phonetic part is not an objective one based on phonetic analysis. The error rate is just an abstract number without any particular meaning to tell how much percent for segmental error and how much for prosodic error.

In this paper, evaluation methods are tried on phonetic annotation and compared with the subjective evaluation approach, whose results tell us that both prosodic and segmental elements must be taken into account in objective or machine evaluation.

2. The regional accented speech corpus and its phonetic annotation

2.1 Speech corpus and subjective evaluation

There are 10 major dialect groups in China, namely the groups of Mandarin, Jin, Wu, Hui, Xiang (or Hunanese), Gan, Yue (or Cantonese), Min, Ping and Hakka.

1 Funded by the Ministry of Education Project ZDI105-B02 and Project2004AA114011-1 on China's National 863 Project

We collected an accented Chinese corpus produced by the speakers or testees taking part in PSC in three big cities: Shanghai (SH), Wuhan (WH) and Xiamen (XM). The dialect of Xiamen, located in the southeast of China, is the representative of the Min. And so are the Wuhan dialect of the Southwest Mandarin and the Shanghai dialect of the Wu.

2.2 Phonetic annotation

We selected speech samples from this corpus, including mono-syllables, multi-syllabic words, sentences and monologues. For each speaker, two experienced phoneticians made a subjective evaluation on his or her accent level according to the PSC criterion. The results of the subjective evaluation are shown in Table 1.

In this paper, phonetic transcriptions include time aligned segmentations and prosodic annotations. The segmental annotation is automatically done by computer first, and then examined by 4 professional transcribers. There are four tiers annotated: 1) Chinese character tier; 2) Pinyin tier, labeled with orthographic pinyin and tone transcription based on SC; 3) Syllable tier, labeled with syllable boundary, orthographic pinyin and tone transcription; and 4) Initial-final tier, labeled with the initial and final boundary of each syllable and their real pronunciations with sound variations. The sound variations are manually annotated which include sound addition, sound deletion, centralization, nasalization and phoneme variation. Besides, *mispronunciations including phoneme and tone variation caused by accent, i.e. sound variations deviated from ideal pronunciation of Standard Chinese, are manually annotated with the symbol '#'* in the tiers of Chinese character, syllable and initial-final, while those mispronunciations caused in other ways are annotated with the symbol '*'. The prosodic information is labeled based on C-ToBI system.[3]

Table 1: *Speaker distribution in accent levels based on subjective evaluation*

Spk	L1	L2	L2	L3	L3	Heavy	Total
	B	A	B	A	B		
SH	1	6	12	18	3	0	40
WH	4	20	50	38	45	0	157
XM	2	4	6	6	9	2	29

3. Accent level based on segmental annotation

Theoretically, if we look the Standard Chinese as the terminal language, the dialect as the source language of learners, the more pronunciation deviated from that of SC (i.e. higher pronunciation error rate of '#'), the low accent levels the learners have.

We made statistical analysis on pronunciation errors for each speaker just based on segmental annotation. Table 2 (at the final page) gives part of the statistical results of the initials and finals for one speaker from Shanghai. Column 1 shows the initials and finals, column 2 shows their occurrence times, and column 3 shows the occurrence times of the consistency annotations with the citation forms. '*' stands for mispronunciation, and '#' stands for mispronunciation caused by accent.

We applied the error rates in PSC criterion to categorize the speakers' accent levels. Here we only took segmental errors into account without considering the prosodic aspect. For the speaker in Table 2, as the average error rate is 7.26% (3% <7.26% <8%), the speaker's accent level is assigned to L1-B.

For each speaker from Shanghai or Xiamen, about 600 annotated syllables are statistically calculated, including mono-syllables, words and sentences. For each Wuhan speaker, on the other hand, only 200 annotated syllables are included for mono-syllables and words. Table 3(at the final page) shows the evaluation results for Xiamen speakers by this PSC method (hereafter PSC method).

4. Clustering Analysis on segmental annotation

In this section, we clustered the speakers according to their phonetic annotation results and mapping the clusters onto the accent levels.

4.1 Clustering on 157 Wuhan speakers

According to the results of subjective evaluation, we found that few out of 157 speakers fall into L1. So we clustered the speakers into four classes of L2-A, L2-B, L3-A and L3-B according to the segmental annotation results. The speakers who failed in the examination, e.g. wh2-01, are not taken into account. Tables 4 and 5 show the cluster results.

Table 4: *Final Cluster hubs for Wuhan speakers*

	Cluster			
	1(L3-A)	2(L3-B)	3(L2-A)	4(L2-B)
ERR	22.657	30.615	7.476	14.959
Num	33	8	65	50

Table 5: *Accent levels for Wuhan speakers (part) based on phonetic annotation and subjective evaluation*

SPK	By Cluster	PSC	Subjective
Wh1-01	L2-A	L2A	L2-B
Wh1-02	L2-B	L2B	L3-B
Wh1-03	L2-B	L2B	L3-B
Wh1-04	L2-B	L2B	L3-B
Wh1-05	L2-A	L1B	L2-B
Wh1-06	L2-B	L2B	L2-A
Wh1-07	L2-B	L2A	L1B
Wh1-08	L2-A	L1B	L2-A
Wh1-09	L2-A	L2A	L2-A
Wh1-10	L2-A	L1B	L1-B
Wh1-11	L2-A	L2A	L2-A
Wh1-12	L2-A	L1B	L2-A
Wh1-13	L2-A	L1B	L2-B
Wh1-14	L2-A	L1B	L2-B
Wh1-15	L2-A	L1B	L2-A
Wh1-16	L2-A	L1B	L2-A
Wh1-17	L2-B	L2B	L3-B
Wh1-18	L3-A	L3A	L2-A
Wh1-19	L2-B	L2B	L2-A
Wh1-20	L2-A	L1B	L3-A

4.2 Clustering on 40 Shanghai speakers

For the speakers from Shanghai, none of the 40 speakers were found to be on level L1. The clustering method applied is therefore the same to what we have used on Wuhan speakers. The results are shown in Tables 6-7.

Table 6: *Final Cluster hubs for Shanghai speakers*

	Cluster			
	1(L3-A)	2(L3-B)	3(L2-B)	4(L2A)
ERR	10.5637	15.9396	6.9524	4.2696
Num	6	1	19	14

Table 7: *Accent levels for Shanghai speakers based on phonetic annotation and subjective evaluation (part of the results)*

SPK	By Cluster	PSC	Subjective
SH1-1	L2-B	L1-B	L3-A
SH 1-2	L2-B	L1-B	L3-B
SH 1-3	L2-B	L1-B	L3-A
SH 1-4	L2-A	L1-B	L3-A
SH 1-5	L3-A	L2-A	L2-B
SH 1-6	L2-B	L2-A	L3-A
SH 1-7	L2-A	L1-B	L3-B
SH 1-8	L2-B	L1-B	L3-A
SH 1-9	L2-A	L1-B	L2-A
SH1-10	L3-A	L2-A	L2-B
SH1-11	L2-A	L1-A	L2-A
SH1-12	L2-A	L1-A	L2-B
SH1-13	L2-B	L1-B	L2-B
SH1-14	L2-B	L2-A	L2-B
SH1-15	L3-A	L2-A	L3-A
SH1-16	L3-A	L2-A	L3-A

4.3 Clustering on 27 Xiamen speakers

For Xiamen speakers, an additional cluster, L1-B, is added. Hence the 27 speakers are divided into 5 classes. Tables 8-9 show the results in detail.

Table 8: *Final Cluster Centers*

	Cluster				
	1(L1_B)	2(L3_A)	3(L3_B)	4(L2_B)	5(L2_A)
ERR	2.390	13.820	19.223	9.120	6.158
Num	5	7	6	7	2

Table 9: *Accent levels for Xiamen speakers based on phonetic annotation and subjective evaluation (part of the results)*

SPK	By Cluster	PSC	Subjective
XMs-01	L3-B	L1-B	L3-A
XMs-02	L2-B	L1-B	L3-B
XMs-03	L2-B	L1-B	L3-A
XMs-04	L2-B	L1-B	L3-A
XMs-05	L3-A	L2-A	L2-B
XMs-06	L2-A	L2-A	L3-A
XMs-07	L3-A	L1-B	L3-B
XMs-08	L3-A	L1-B	L3-A
XMs-09	L2-A	L1-B	L2-A
XMs-10	L2-A	L2-A	L2-B
XMs-11	L3-B	L1-A	L2-A
XMs-12	L3-A	L1-A	L2-B
XMs-13	L2-B	L1-B	L2-B
XMx-01	L2-A	L2-A	L2-B

5. Observations from the evaluation results

The results of subjective evaluation, object evaluation (only considering error rate of the segments) and automatic clustering (only considering error rate of the segments) are shown in Tables 5, 7 and 9. It's obvious that the difference among the three results is significant for some speakers.

It shows that most accent levels of object evaluation are higher than those of subjective evaluation and the results of automatic clustering are closer to those of subjective evaluation within one level or just one sub-level discrepancy.

Figure 1 shows the clustering results for three cities based on only segmental annotation results. After comparing the error rates of all the accent levels, we found that WH speakers fit in with subjective results very well whereas SH and XM speakers show greater deviation. The possible reason is that WH speakers have better prosody than those from the other two regions, whereas, their segmental error accounts for most part of the pronunciation mistake.

One of the evidences from the lexical tones of the three dialects is shown in Table 10. There are no concave tones (T3: 214) in SH and XM dialects, so XM and SH speakers may have difficulty in learning this low dipping tone. Additionally, XM and SH speakers have problems in pronounce neutral tone syllables.

Table 10: *Phonological tones of SC and three dialects*

5-Letter tone	T1	T2	T3	T4	T5	T6	T7
SC	55	35	214	51			
XM	55	35	53	21	22	55	22
WH	55	213	42	35			
SH	53	34	13	5	23/12		

We know from the analysis above that if we take the results of subjective evaluation as the correct results, we find that: 1) the error rates in the accent level criterion have different percentages on prosodic and segmental parts for different regional speakers. In our experiment, XM has greater prosodic weight than SH and WH, WH may bear the lowest prosodic weight, which was caused by the phonology deviation between the dialects and the Standard Chinese. 2) Supra-segmental factors, such as tone, intonation, rhythm and stress, play an important role in evaluation, which implies that we should combine both segmental and supra-segmental factors to give a correct result when doing objective or machine evaluation. It's also true for the L2 learners.

6. Conclusion

The analysis of the accented Chinese in the three cities is still an on going endeavor. The present results indicate that the study of tone, intonation and rhythm is the cornerstone, and they are also the cornerstone for second language acquisition. Therefore, prosodic analysis on tone and discourse intonation [5] will be very useful. Furthermore, Standard Chinese corpus alone is not sufficient for evaluating the level of Standard Chinese. We need to build corpora of different accented speech and take both segmental and prosodic factors into account to get objective criteria. We are focusing on investigating the difference of tone realization between accented speakers and SC speakers.

Acknowledgements: Thanks for our students JIA Yuan, XU Ruiyuan and ZHU Lin to prepare and check the data.

7. References

[1] LI Rong,ZHOU Changji, edited, Xiamen Fanyan Cidian, Jiangsu Education Publishing House,1998.

[2] LI Aijun, YU Jue, CHEN Juanwen and WANG Xia , A Contrastive Study of Standard Chinese and Shanghai-Accented Standard Chinese, in H. Fujisaki, G. Funt, J. Cao and Y. XU ed.' From Traditional Phonology to Modern Speech Processing'. Foreign language teaching and research press, 2004.

[3] Aijun Li, "Chinese prosody and prosodic labeling of spontaneous speech", Proceedings of speech prosody 2002.[4] Aijun LI, Qiang FANG, Ruiyuan XU Xia WANG, Yunzhong TANG, A Contrastive Study between Minnan-accented Chinese and Standard Chinese, Coming in O-COCOSDA2006, Indonesia.

[4] Dorothy M. Chun, Discourse intonation in L2- from theory to pratice, John Benjamin Publishing Company,2002.

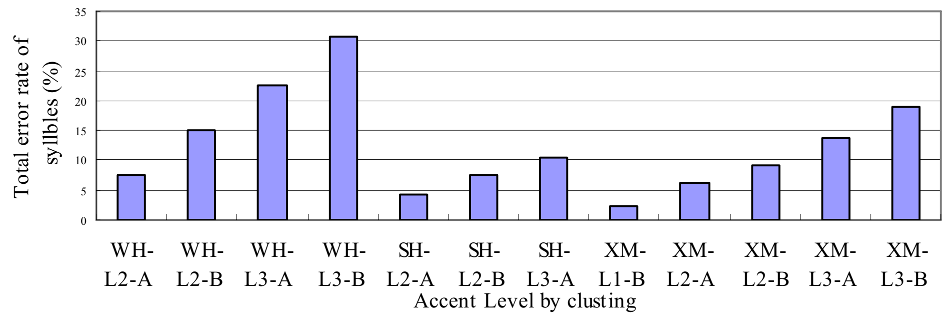

Fig.1 Error rate by clustering method for three cities

Table 2: *Example of annotation results of initials and finals for one SH speaker (partial result for the limited space)*

I/F	Total no.	Consistency no.	Inconsistency no.(including sound variations)	Correct rate (including sound variations)%	#	*	Mis-reading %	#%	Total error Rate	correct rate (excluding Sound variations)%
[p]	19	16	3	84.21	2	0	0	0.105	0.105	0.894
[ts^{h}]	9	9	0	100.00	0	0	0	0	0	1
[tʂ^{h}]	19	14	5	73.68	4	0	0	0.210	0.210	0.789
[t]	53	33	20	62.26	0	0	0	0	0	1
[f]	11	10	1	90.91	1	0	0	0.090	0.090	0.909
[k]	47	43	4	91.49	0	0	0	0	0	1
[x]	20	20	0	100.00	0	0	0	0	0	1
[tɕ]	45	40	5	88.89	6	0	0	0.133	0.133	0.866
[k^{h}]	14	13	1	92.86	1	0	0	0.071	0.071	0.928

Table 3: *Accent levels for XM speakers evaluated by PSC error rates (only partly showed)*

SPK.	Total syll.	# %	* %	total error rate %	Total correct rate%	Level
XMs-01	508	18.260	2.412	20.672	79.328	L3-A
XMs-02	601	6.156	2.329	8.486	91.514	L2-A
XMs-03	605	5.950	2.149	8.099	91.901	L2-A
XMs-04	600	7.5	3.333	10.833	89.167	L2-A
XMs-05	593	9.949	3.204	13.153	86.846	L2-B
XMs-06	167	4.5	2.5	7	93	L1-B
XMs-07	599	9.850	4.007	13.856	86.144	L2-B
XMs-08	608	14.145	1.316	15.460	84.539	L2-B
XMs-09	600	5.5	1.667	7.1667	92.833	L1-B
XMs-10	605	5.124	1.653	6.777	93.223	L1-B
XMs-11	602	15.116	2.658	17.774	82.226	L2-B
XMs-12	600	10.1667	1.667	11.833	88.167	L2-A