

THE EFFECT OF NON-NATIVE DIALECT ON SPEECH RECOGNITION THRESHOLD
FOR NATIVE MANDARIN SPEAKERS

by

Nathan Richardson

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Communication Disorders

Brigham Young University

April 2008

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Nathan Richardson

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date

Richard W. Harris, Chair

Date

Shawn Nissen

Date

Ron W. Channell

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Nathan Richardson in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Richard W. Harris
Chair, Graduate Committee

Accepted for the department

Ron W. Channell
Graduate Coordinator

Accepted for the college

K. Richard Young
Dean, David O. McKay School of Education

ABSTRACT

THE EFFECT OF NON-NATIVE DIALECT ON SPEECH RECOGNITION THRESHOLD FOR NATIVE MANDARIN SPEAKERS

Nathan Richardson

Department of Communication Disorders

Master of Science

Speech recognition thresholds are used for several clinical purposes, so it is important that they be accurate reflections of hearing ability. Variations in the acoustic signal may artificially decrease threshold scores, and such variations can result from being tested in a second dialect. Thirty-two native Mandarin-speaking subjects (sixteen from mainland China and sixteen from Taiwan) participated in speech recognition threshold testing in both dialects to see whether using non-native dialect test materials resulted in a significantly lower score. In addition, tests were scored by two interpreters, one from each dialect, to see whether the scorer's dialect resulted in a significantly different score. Talker dialect was found to be statistically significant, while scorer dialect was not. Factors explaining these findings, as well as clinical implications, are discussed.

ACKNOWLEDGMENTS

Naming all the people I am indebted to for my accomplishments thus far would be an endless task, but that is all the more reason to start now, by thanking those who primarily helped me with this project. I thank my committee members for their willingness to meet with me as many times as it took to hammer out the details and for their good-natured realism and support when I didn't know how to complete each next step. I thank my thesis partner, Jamie Garlick, for taking those blind steps with me and laughing with me when they occasionally led to a brick wall. I thank our Mandarin interpreters, Shu-Ling Ko and Hong Wu, for being so disarmingly competent inside the lab and reminding me that real life went on outside the lab. I thank my subjects, many of whom recruited other subjects out of kindness and perhaps pity toward a busy graduate student.

Even outside the direct participants of this project, however, I was helped in innumerable ways. I thank authors like C. S. Lewis, Dallin H. Oaks, and Orson Scott Card, who taught me how to convey clearly to others what is already clear to me. I thank my parents, Darwin and Nina Richardson, for quietly assuming other tasks that were occupying my life, and for their encouraging phone calls reminding me that the benefits of completing my degree eclipsed the many menial moments, even though it

was hard for me to see at times. I thank my siblings, who through our banter over the years were unwittingly teaching me how to recognize good reasoning and to form an opinion only after considering all the information. I thank my entire family for continually reminding me what matters most in life and how individual endeavors like this one serve our higher goals. I thank my fiancée, Jelaire Lemmon, for becoming one of those higher goals by taking a chance on this lump of unrefined silver. Because of her, I have discovered better reasons for completing grad school than I had for starting it.

Finally, I thank my Heavenly Parents and my Savior, Jesus Christ, for giving me life and new life, and for putting me in a family so wonderful that I sometimes thought my joys would climax early in life, only to discover that, inconceivably, it actually does keep getting better and better.

Table of Contents

Introduction	1
Literature Review.....	3
Speech Recognition Threshold.....	3
Second Language	4
Second Dialect	5
Mandarin Dialects.....	7
Method	9
Participants	9
Materials.....	12
Procedure	12
Results.....	16
Test-retest	16
Talker Dialect.....	19
Figure 1: Mean SRT Scores for Test and Retest for Mandarin Speakers	
Hearing Talkers from Mainland China and from Taiwan.....	20
Scorer Dialect.....	22
Discussion	23
Talker Dialect.....	23
Scorer Dialect.....	25
Limitations	26
Future Research.....	26

References	28
Appendix: Informed Consent.....	30

List of Tables

Table	Page
1. Subjects Information and Audiograms	10
2. Scorers Information and Audiograms.....	11
3. SRT Scores in dB HL for Mandarin Speakers Hearing Talkers from Mainland China and from Taiwan.....	17
4. Mean SRT Scores in dB HL for Mandarin Speakers Hearing Talkers from Mainland China and from Taiwan.....	18

Introduction

The purpose of this study is to determine whether administering a speech recognition threshold (SRT) test to a Mandarin-speaking client using materials or testers of a Mandarin dialect different from that of the speaker will result in a significantly different score. This question is applicable to the administration of clinical SRT tests and to the production and selection of SRT test materials, as well as to more theoretical questions related to dialect.

SRT scores are important for several reasons. When a person has a hearing impairment severe enough to merit a hearing aid, one of the pieces of information used in fitting the hearing aid and calibrating it is the client's speech recognition threshold (SRT). The SRT is also used as a base for word-recognition testing, as well as for checking the reliability of pure-tone responses (Schneider, 1992). Therefore, it is important that a client's SRT be accurate and truly reflect the person's hearing status and receptive speech abilities in the acoustic realm.

Since SRT testing is done at the lowest possible intensity levels, it tests the ability of the brain to process the auditory signal and language content. It follows that any modification to the signal that changes it from optimal conditions could artificially decrease the SRT and inaccurately portray a person's hearing abilities. One such change is dialect. This paper examines the effect of using a non-native dialect on the SRT score. Participants were tested using word lists in their native Mandarin dialect and in a non-native Mandarin dialect to see whether the SRT scores' accuracy was significantly affected by the speech recordings used. They were also judged by two scorers, one from

both dialects, to see whether the SRT scores' accuracy was significantly affected by the dialect of the tester.

The following questions will be examined in this study:

1. Does the dialect of the talker on the speech audiometry recordings matter when determining SRT for Mandarin speakers? That is, if the subjects are unfamiliar with the dialect spoken on the recordings, will their SRT scores be artificially lower?
2. Does the dialect of the test scorer matter when determining SRT for Mandarin speakers? That is, if the person scoring the responses is unfamiliar with either the dialect spoken on the recordings or the dialect spoken by the subject, will his or her scores be significantly different from those of an examiner more familiar with the dialect being spoken?

Literature Review

The purpose of audiometry is to assess and quantify a person's ability to detect and process sound. One form is pure-tone audiometry, which is the method preferred by audiologists for evaluating hearing loss because of its high reliability and validity. Pure-tone audiometry, however, does not measure a person's ability to understand speech. Since hearing is primarily used to process speech signals (Hagerman, 1993), another form of audiometry is frequently used to assess speech comprehension: speech audiometry.

Several types of speech audiometry are used to evaluate a person's ability to hear and process speech, such as word recognition scores (WRS) or SRTs. Speech audiometry is useful in evaluating the practical impact of hearing loss on social interactions and opportunities. It is also used for calibrating and evaluating hearing aids, diagnosing auditory disorders (Hood & Poole, 1977; Jerger, Speaks, & Trammell, 1968; Van Dijk, Duijndam, & Graamans, 2000), and measuring progress in auditory therapy. This study focuses on the use of SRT.

Speech Recognition Threshold

SRT is defined as the lowest intensity level at which a person correctly understands words with 50% accuracy. For English, the words presented are typically spondaic, two syllables with equal stress (ASHA, 1988; Hudgins, Hawkins, Karlin, & Stevens, 1947; Ramkissoon, 2001). This is because spondees can be heard more easily at low intensity levels (Bell & Wilson, 2001; Dennis & Neely, 1991).

The pure-tone hearing thresholds at 500, 1000, and 2000 Hz are often combined into the pure-tone average (PTA) as an estimate for the SRT because they have been found to correlate (Dennis & Neely, 1991). The two data are frequently used to verify each other (ASHA, 1988; Egan, 1979; Young et al., 1982).

Several factors affect the reliability and validity of SRT scores, including the number of words in the list (Grubb, 1963; Resnick, 1962), the words selected for the list (Cambron, Wilson, & Shanks, 1991; Luce, 1986; Wilson & Carter, 2001), the dialect or accent of the speaker (Weisleder & Hodgson, 1989), the method and intensity level at which the words are presented (Beattie, Svihovec, & Edgerton, 1975; Hood & Poole, 1980; Pisoni, 1985), and the type of recording made (Kamm, Carterette, Morgan, & Dirks, 1980; Ridgway, 1986). Two major factors to consider are the listener's familiarity with the language the words come from and with the dialect in which they are spoken.

Second Language

The choice of language used in the speech audiometry recordings has been shown to have an effect on the outcomes of SRT testing. For example, one study reports that when tests are given to native Spanish speakers, "Spanish tests of auditory function consistently yield better performance than English tests for native speakers" (von Hapsburg & Pena, 2002, p. 202).

This finding was confirmed in a study by Ramkissoon et al (2002). SRTs were obtained from 24 subjects, only half of whom were native English speakers, using 2 sets of English speech recordings. The first consisted of 36 familiar spondees (e.g., *horseshoe*, *drawbridge*); the second consisted of digit pairs (e.g., *two-five*, *eight-three*). In theory, the

digit pairs would be more familiar to a person just beginning to learn English, while many of the spondees might be unique to American English or less common and thus not learned early on during second language acquisition. For each subject, the two SRTs were compared to the subject's PTA. The researchers found that the recordings of digit pairs produced SRTs that closely correlated to the PTA, while the recordings of spondees produced SRTs that were significantly poorer than the PTA would predict. This implies that a lack of familiarity with the speech materials can adversely affect the SRT, and that non-native English speakers are less likely to produce accurate SRTs when tested with materials that require a certain level of familiarity with the language.

Clients who are less familiar with the language in which the word list is presented will not do as well as native speakers of that language, even though their actual hearing abilities and true SRTs may be equal to, or even better than, the SRTs of native speakers. Thus, professionals must accommodate the native languages of their clients in order to obtain accurate test results and best meet their clients' needs.

Second Dialect

In addition to the effect of non-native languages, some researchers have explored whether different dialects within the same language could also have an effect on SRT testing. In theory, since SRT testing requires that the client make very fine perceptual judgments of slight featural differences at the phonological and morphological levels, and at very low intensity levels, even the slightest deviation from the phonetic patterns to which clients are accustomed may cause them to perform more poorly than they would have if they were being presented with their native dialect. Dialects differ far less

than languages, but the slight allophonic variations at low intensity levels may be enough to make a difference in the SRT and thus in hearing aid recommendation and calibration, estimations of PTA, pure-tone threshold reliability checking, and word-recognition checking.

Some researchers have already begun to study the effect of dialect on SRT. Schneider (Schneider, 1992) administered word lists in three Spanish dialects (Caribbean, Castilian, and Mexican) to 12 children of Puerto Rican background who spoke a Caribbean dialect of Spanish, ranging in age from 6 to 7 years old. The resulting SRTs did not differ from each other in a statistically significant way. That is, the outcomes of an SRT test were similar whether the children were being tested with recordings in their native Spanish dialect (Caribbean) or in an unfamiliar dialect (Castilian or Mexican). However, the mean SRTs did differ from the mean PTA in a statistically significant way; the mean SRTs for the three dialects ranged from 1.9 to 2.4 dB lower than the PTA. This difference was considered negligible and not clinically significant for practical purposes.

These results imply that, while second language has been found to significantly impact speech audiometry testing, second dialect has not been clearly shown to do the same, at least not for these three dialects of Spanish, for subjects within this age range. Further study is needed since this study examined only one language and the sample size was small. Similar research involving more subjects and other languages, whether corroborating or contradicting existing findings, will serve to clarify the question of language's effect on speech audiometry.

While most researchers attribute the observed differences created by second-language SRT materials to the subjects' lack of familiarity with the language used, others have proposed that "the Spanish language might somehow be more perceptually salient than English" (von Hapsburg & Pena, 2002, p. 202). If one language might theoretically have more inherently discriminable features at low intensity levels than another, it follows that one dialect might have inherently more discriminable features than another at low intensity levels, based on allophonic variation. The purpose of the present study is to determine whether the dialect in which words are presented significantly affects SRT scores. This study could lay the groundwork for further research into that question. It could also improve client services through more reliable speech recognition threshold scores. The outcome of this study could ultimately help answer the question of whether a client may be tested with materials recorded in a dialect not native to the client, and whether a tester who speaks one dialect can administer an SRT test to a client who speaks another dialect, without significantly affecting the score.

The scope of applicability will of course be limited to the language used in the study. The language used in examining this question is Mandarin Chinese, and the two dialects are Pǔtōnghuà Mandarin and Taiwan Mandarin. These were chosen because speech audiometry materials were available in those dialects.

Mandarin Dialects

Mandarin has a relatively simple syllable structure. It is distinct from Western languages in that it uses tones to convey semantic distinctions (Zhou & Marslen-Wilson,

1995). The four tones, not including the neutral tone, are high, rising, dipping, and falling (with some tones changing in certain combinations). Mandarin lacks inflectional morphology. Instead, grammatical categories are conveyed through syntactic markers and logical order (Campbell, 1991). Each word consists of a single syllable (and corresponding character) that may have multiple meanings. Thus, they are often combined with other words to avoid ambiguity (Campbell, 1991).

This study utilizes the Pǔtōnghuà dialect, spoken by many people in the northern People's Republic of China, and the Taiwan dialect of Mandarin. While they are mutually intelligible, they differ orthographically, syntactically, and semantically (Li, 1984). Orthographically, the Taiwanese use the traditional character set, while mainland Chinese use a simplified character set that uses fewer strokes. Thus, printed familiarization lists were given in both orthographic systems. Semantically, the dialects share about 70% of lexical items (Cheng, 1985).

In mainland China, Pǔtōnghuà is a common dialect, particularly in the north. In Taiwan, most people learn Taiwan Mandarin as a first language (Lee, 1981). Thus, all subjects in this study were required to have learned Mandarin as their first language.

Method

Participants

For this study, 32 native Mandarin speakers were recruited, 16 from the People's Republic of China (mainland China) and 16 from the island of Taiwan (see Table 1). The Mainland subjects ranged in age from 19 to 50 years ($M = 27.7$ years); the Taiwanese subjects ranged in age from 18 to 49 years ($M = 26.3$ years). All participants learned Mandarin as their first language and have used it regularly in daily communication since coming to the United States.

Each participant had a pure-tone air-conduction threshold of ≤ 15 dB HL at both octave and mid-octave frequencies from 125 Hz to 8000 Hz (see Table 1). Each participant had static acoustic admittance between 0.3 and 1.4 mmhos with peak pressure between -100 and $+50$ daPa in the test ear (ASHA, 1990; Roup, Wiley, Safady, & Stoppenbach, 1998). Each participant had an acoustic reflex present at levels of ≤ 95 dB HL at 1000 Hz in the test ear. Each participant read and signed an informed consent form before participating in the study.

Both scorers were female native Mandarin speakers. They learned Mandarin as their first language and have used it regularly in daily communication since coming to the United States. The Mainland scorer was 28 years old and the Taiwanese scorer was 34 years old. Both met the same minimum hearing ability specifications required of the subjects (see Table 2).

Table 1

Subjects Information and Audiograms

Ss ^a	Gender	Age	Hometown	Time in US ^b	Better ear	Frequency in Hz										
						125	250	500	750	1k	1.5k	2k	3k	4k	6k	8k
M1	f	37	Beijing	5:0	L	5	5	0	0	0	5	0	0	-5	5	5
M2	f	25	Shanghai	2:0	L	10	10	10	10	5	0	5	-5	0	0	15
M3	m	50	Huhhot	0:11	L	0	0	10	5	5	10	5	0	0	10	5
M4	m	26	Guangzhou	3:0	R	10	5	5	0	0	5	0	5	0	10	10
M5	f	20	Changsha	5:0	R	5	5	5	0	0	15	15	10	0	-5	-5
M6	m	22	Guangzhou	1:11	L	5	0	5	10	5	5	5	0	5	-5	-5
M7	m	23	Shanghai	4:10	L	5	5	5	5	5	10	15	10	0	5	5
M8	f	27	Huangshan	0:3	L	15	10	5	10	5	5	5	5	10	-5	5
M9	m	27	Nanjing	3:6	R	5	5	5	5	5	5	10	5	5	5	10
M10	f	30	Zhuji	1:0	R	10	10	10	5	5	5	10	5	0	0	5
M11	m	19	Shanghai	3:0	R	10	5	0	5	10	5	5	-5	-5	-5	5
M12	f	35	Tianjin	2:10	R	5	0	0	0	5	10	10	5	5	0	0
M13	f	32	Yantai	3:2	R	10	5	5	5	5	5	10	0	0	0	-5
M14	f	27	Zi Bo	3:0	L	0	0	0	-5	5	0	0	0	-5	0	0
M15	m	23	Guiyang	0:6	L	15	10	5	10	15	0	5	0	0	10	10
M16	f	28	Xi'an	3:0	L	5	0	0	10	10	10	5	10	10	0	0
<i>M</i>		28.2		2:8		7.2	4.7	4.4	4.7	5.3	5.9	6.6	2.8	1.3	1.6	3.8
Min		19		0:3		0	0	0	-5	0	0	0	-5	-5	-5	-5
Max		50		5:0		15	10	10	10	15	15	15	10	10	10	15
<i>SD</i>		7.7		1:6		4.5	3.9	3.6	4.6	3.9	4.2	4.7	4.8	4.7	5.4	5.9
T1	f	29	Keelung	5:0	L	-5	5	5	5	0	5	5	-5	-5	0	0
T2	f	27	Taitung	0:10	L	5	5	5	0	5	15	10	10	10	10	10
T3	f	49	Taipei	18:0	L	15	15	10	5	-5	0	0	5	15	10	15
T4	f	21	Pingtung	2:0	L	-5	-5	0	0	0	5	5	0	0	-10	-5
T5	m	19	Nantou	1:0	R	10	5	0	5	-5	5	5	0	0	5	0
T6	f	29	Kaohsiung	1:10	L	5	5	5	5	5	10	5	15	0	5	0
T7	f	20	Kaohsiung	1:6	L	15	5	10	10	10	10	15	10	10	10	10
T8	m	32	Taipei	3:0	R	5	0	0	5	10	0	0	5	0	-5	0
T9	f	27	Kaohsiung	1:0	L	15	10	10	10	10	10	10	5	5	15	5
T10	f	18	Taipei	0:11	R	5	5	0	0	5	5	-5	10	10	10	10
T11	f	35	Taipei	10:0	L	10	10	15	15	15	10	0	10	15	10	0
T12	m	31	Taipei	6:0	R	5	5	5	0	5	10	10	10	10	0	5
T13	f	19	Taichung	0:6	L	5	0	5	5	10	0	5	5	0	0	-5
T14	f	18	Yingko	0:6	L	5	5	5	5	5	5	5	-5	0	-10	-5
T15	m	27	Changhua	1:2	L	5	5	5	5	0	0	0	5	0	-5	0
T16	f	19	Taoyuan	0:11	L	10	5	0	0	0	5	0	5	5	0	-5
<i>M</i>		26.3		3:5		6.6	5.0	5.0	4.7	4.4	5.9	4.4	5.3	4.7	2.8	2.2
Min		18.0		0:6		-5	-5	0	0	-5	0	-5	-5	-5	-10	-5
Max		49.0		18:0		15	15	15	15	15	15	15	15	15	15	15
<i>SD</i>		8.3		4:8		6.0	4.5	4.5	4.3	5.7	4.6	5.1	5.6	6.2	7.7	6.3

^aSubject numbers that begin with *M* are Mainland subjects; those that begin with *T* are Taiwanese.

^bTime lengths are in Years:Months notation.

Table 2

Scorers Information and Audiograms

Scorer ^a	Gender	Age	Hometown	Time	Ear	Frequency in Hz										
				in US ^b		125	250	500	750	1k	1.5k	2k	3k	4k	6k	8k
M	f	28	Beijing	6:0	R	0	0	0	5	5	10	0	-10	0	5	10
					L	5	10	10	10	10	5	0	5	5	5	15
T	f	34	Taipei	3:4	R	10	10	5	5	5	5	5	0	0	0	5
					L	10	5	5	0	0	5	5	0	-5	0	-5

^aM is the Mainland scorer; T is the Taiwanese scorer.

^bTime lengths are in Years:Months notation.

Materials

Word lists for the two Mandarin dialects were obtained from Brigham Young University (Jennings, 2005; Slade, 2006). They consisted of one list of 24 different trisyllabic words for Pǔtōnghuà Mandarin and one list of 28 trisyllabic words for Taiwan Mandarin, spoken by native male talkers of the respective dialects. “Words that had the same pronunciation but different meanings, represented by different characters, were avoided. Additionally, words were eliminated from the original lists if they were considered, by native Mandarin Chinese judges, to be culturally insensitive, unfamiliar, and/or representative of inappropriate content” (Jennings, 2005, p. 12).

All threshold tests and SRT tests were performed in an double-walled sound booth with a Grason-Stadler model 1761 audiometer with TDH-50 Telephonics supra-aural earphones, meeting ANSI S3.1 standards for maximum permissible ambient noise levels for the ears not covered condition using one-third octave-bands (American National Standards, 1999). The audiometer and participant’s headphone were calibrated regularly to within 1 dB of the standard for each frequency. The two scorers each listened with Sennheiser HD 650 headphones maintained at the same comfortable listening level for all sessions.

Procedure

Each participant went through two identical sessions. The sessions were always at least one week apart, usually on the same day of the week, and usually at the same hour of the day. In both sessions, the participant received two SRT tests, one in the native dialect and one in the foreign dialect. Tests were administered in the better ear

and all were done with a single Telephonics TDH-50P headphone. The talkback microphone was 24 inches from the speaker's mouth.

Before beginning a session, subjects were familiarized with the words used. They received a written copy of the two wordlists, in both character sets. Subjects were read the following instructions:

Before the test begins, you will now hear a list of words we will use in this part of the research study. These words will be presented at a comfortable listening level. As you hear them, please read the list of words silently to yourself to make sure you are familiar with the words.

Do you have any questions?

Subjects then heard a recording of all the words used in the study. They heard both the Mainland list (24 different trisyllabic words from the Pütōnghuà dialect) and the Taiwan list (28 trisyllabic words from the Taiwan dialect). Subjects then heard the following instructions.

For this test, you will now hear Mandarin words at a number of different loudness levels. Each word is three syllables in length. At the very soft loudness levels, it may be difficult for you to hear the words.

For each word, listen carefully to the word, then repeat what you think the word was. If you are not sure, you may guess. If you have no guess, simply say, "I don't know," in English or wait silently for the next word.

Do you have any questions?

Which list was presented first (Taiwan then Mainland, or Mainland then Taiwan) was determined by a randomized block, so that each list was presented first 8 times for both groups. This order of the two lists remained the same for the second session. The order of the words in each list was randomized for each participant for every threshold test.

For the SRT test, the ASHA method was used (ASHA, 1988), which is described in the rest of this section. The test consists of three phases: Finding an initial starting level, qualifying a starting level, and finding the SRT.

To find an initial starting level, the first word was played at 30 dB HL. If the subject responded correctly (repeated the word that was played), the next word was played 10 dB softer. If the subject responded incorrectly, the next word was played at the same intensity a second time. When the subject responded incorrectly twice at the same intensity, the initial starting level was set at 10 dB above that intensity, and phase two began.

To qualify a starting level, six words were played at three intensity levels. Two words were played at the initial starting level, then two more at 2 dB below the initial starting level, then two more at 2 dB below that level. If the subject responded correctly to 5 out of those 6 words, the starting level qualified, and phase three began. If the subject responded incorrectly to 2 or more of those words, an adjusted starting level was established at 6 dB higher than the initial starting level. This phase then began again, beginning at the adjusted starting level. When using a new starting level, the subject was retested at the levels already tested.

To find the SRT, two words continued to be played at each intensity level, beginning at the qualified starting level (established in the second phase) and decreasing in 2 dB increments. The test ended when the subject responded incorrectly to 5 out of 6 consecutive words.

The SRT score was determined by subtracting the number of correct responses (obtained at the qualified starting level and lower) from the qualified starting level and adding one (a correction factor; ASHA, 1988). The accuracy of each response was rated by two native scorers (one whose first dialect was Pǔtōnghuà and one whose first dialect was Taiwan Mandarin), both of whom were present for every test. In every instance, the scorers heard both the word presented and the subject's response at a comfortable listening level through their headphones.

Results

Two effects are being analyzed: SRT score and scorer agreement. For the SRT score analysis, four SRT scores were collected from each of the 32 participants. Each participant produced two native dialect SRT scores (test and retest) and two non-native dialect SRT scores (test and retest), producing 128 data points for analysis. The dependent variable was SRT score. The independent variables were talker dialect and subject dialect. In addition, test-retest was examined. A mixed-model ANOVA was used with two within-subject factors (talker dialect and test-retest) and one between-subject factor (subject dialect).

For scorer agreement, the two scorers' scores for individual words were compared for percentage agreement by subtracting the number of disagreements from the total number of scores, dividing the difference by the total number of scores, and multiplying by one hundred. A high percentage agreement would mean that there was no significant difference between the scoring results of the two native scorers, and would have implications for clinical testing, as will be discussed.

The average SRT was 3.20 dB HL, with a range from -6 dB to +10 dB HL. The greatest improvement in score upon retesting was 10 dB (occurring for 2 subjects), and the greatest decline in score upon retesting was 2 dB (occurring for 5 subjects; see Tables 3 and 4).

Test-retest

The average SRT score improved upon retesting, as expected. However, 10 subjects had one score that worsened upon retesting, and 2 subjects had both scores

Table 3

SRT Scores in dB HL for Mandarin Speakers Hearing Talkers from Mainland China and from Taiwan

Subject ^a	Mainland talker		Taiwan talker		PTA ^b
	Test	Retest	Test	Retest	
M1	0	-5	2	-1	0.0
M2	7	-1	6	2	6.7
M3	1	2	4	4	6.7
M4	7	4	5	4	1.7
M5	6	4	6	4	6.7
M6	5	3	9	4	5.0
M7	2	1	3	3	8.3
M8	2	2	5	6	5.0
M9	1	-2	3	1	6.7
M10	8	0	10	10	8.3
M11	5	1	4	3	5.0
M12	1	-1	1	2	5.0
M13	5	-1	7	6	6.7
M14	-2	-2	1	0	1.7
M15	1	2	4	1	8.3
M16	3	5	4	5	5.0
T1	10	8	7	8	3.3
T2	5	2	9	4	6.7
T3	4	4	6	4	1.7
T4	-1	-3	1	3	1.7
T5	3	1	4	5	0.0
T6	5	5	6	8	5.0
T7	7	4	8	5	12.0
T8	0	1	5	1	3.3
T9	5	6	6	8	10.0
T10	1	0	2	2	0.0
T11	5	5	7	5	10.0
T12	5	-2	4	-1	6.7
T13	5	-1	4	1	6.7
T14	0	0	2	4	5.0
T15	4	-6	6	-4	1.7
T16	3	0	3	-1	0.0

^aSubjects that begin with *M* are Mainland subjects; subjects that begin with *T* are Taiwanese

^bPTA was averaged from the thresholds at 500, 1000, and 2000 Hz

Table 4

Mean SRT Scores in dB HL for Mandarin Speakers Hearing Talkers from Mainland China and from Taiwan

Talker Dialect	Subject Dialect	Mean	Min	Max	Range	SD
Mainland Talker Test	Mainland	3.25	-2	8	10	2.93
	Taiwan	3.81	-1	10	11	2.81
Mainland Talker Retest	Mainland	0.75	-5	5	10	2.67
	Taiwan	1.50	-6	8	14	3.67
Taiwan Talker Test	Mainland	4.63	1	10	9	2.55
	Taiwan	5.00	1	9	8	2.28
Taiwan Talker Retest	Mainland	3.38	-1	10	11	2.68
	Taiwan	3.25	-4	8	12	3.45

worsen upon retesting (see Table 3). Test-retest was significant, $F(1, 30) = 18.50, p < .001$. Subjects scored, on the average, 4.17 dB HL on first testing and 2.22 dB HL on retest, averaging 1.95 dB improvement in SRT score on retest. This improvement was observed for both groups of subjects (Mainland and Taiwan), and with both sets of speech audiometry recordings (Mainland and Taiwan; see Figure 1).

The two-way interaction of subject dialect and test-retest was not significant, $F(1, 30) = .030, p = .865$. That is, having established that average SRT scores for both Mainland and Taiwan subjects improved when retested, there was no statistically significant difference in the amount that each group improved. Likewise, the two-way interaction of talker dialect and test-retest was not significant, $F(1, 30) = 3.474, p = .072$. That is, having established that average SRT scores improved when obtained with both Mainland and Taiwan talker recordings, there was no statistically significant difference in the amount of improvement observed using either set of speech audiometry materials. The three-way interaction of subject dialect, talker dialect, and test-retest was not significant, $F(1, 30) = .500, p = .485$. That is, SRT scores improved overall when subjects were retested. This improvement was observed regardless of the subjects' native dialect or of the recordings used.

Talker Dialect

For Mainland subjects, listening to the recordings of Mainland talkers resulted in an average SRT of 2.00 dB HL (3.25 dB HL on first test and of 0.75 dB HL on retest). These thresholds, obtained with materials in these subjects' native dialect, were better than the thresholds obtained with materials in the Taiwan dialect; listening to the

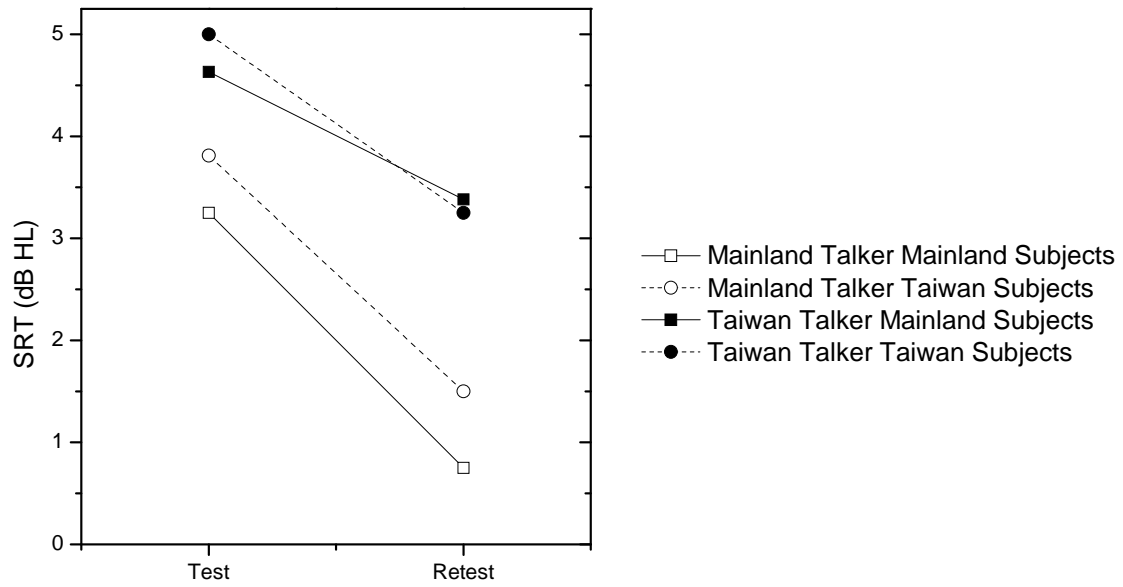


Figure 1. Mean SRT Scores for Test and Retest for Mandarin Speakers Hearing Talkers from Mainland China and from Taiwan

Taiwan dialect recordings resulted in an average SRT of 4.01 dB HL (4.63 dB HL on first test and of 3.38 dB HL on retest; see Figure 1).

For Taiwan subjects, listening to the recordings of Mainland talkers resulted in an average SRT of 2.66 dB HL (3.81 dB HL on first test, and of 1.50 dB HL on retest). Notably, these thresholds were better than those obtained with materials in the Taiwan dialect, which is these subjects' native dialect; listening to the Taiwan dialect recordings resulted in an average SRT of 4.13 dB HL (5.00 dB HL on first test, and of 3.25 dB HL on retest; see Figure 1).

Talker dialect was found to be significant, $F(1, 30) = 37.633, p < .001$. Mainland test materials resulted in significantly better (lower) SRT scores. Mainland test materials produced an average SRT score of 2.33 dB HL, while Taiwan test materials produced an average of 4.07 dB HL. On average, SRT scores obtained with Mainland talker recordings were 1.74 dB better than those obtained with Taiwan talker recordings: 1.29 dB lower on first testing, and 2.19 dB lower on retest.

The poorest average SRT (combining test and retest) was that obtained by Taiwan subjects listening to Taiwan talker materials (4.13 dB HL), followed by Mainland subjects listening to Taiwan talker materials (4.01 dB HL). The best average SRT (combining test and retest) was that obtained by Mainland subjects listening to Mainland talker materials (2.00 dB HL), followed by Taiwan subjects listening to Mainland talker materials (2.66 dB HL).

The two-way interaction of subject dialect and talker dialect was not significant, $F(1, 30) = .883, p = .355$. That is, Mainland talker recordings produced better SRT scores

overall regardless of the subjects' native dialect. This is further reflected in Figure 1: note that the two data points for Taiwan talker materials are above the two data points for Mainland talker materials, for both test and retest.

Scorer Dialect

The difference in native dialect of the two scorers used in scoring the tests was not found to be a factor that greatly affected the outcome of the SRT tests. Combined, the 128 test administrations produced 2933 individual word scores. With only 8 disagreements, the interscorer percentage of agreement was 99.5%.

Discussion

There were several main questions in this study. First, does the dialect of the talker on the speech audiometry recordings matter when determining SRT for Mandarin speakers? That is, if the subjects are unfamiliar with the dialect spoken on the recordings, will their SRT scores be artificially lower? Second, does the dialect of the test scorer matter when determining SRT for Mandarin speakers? That is, if the person scoring the responses is unfamiliar with either the dialect spoken on the recordings or the dialect spoken by the subject, will his or her scores be significantly different from those of an examiner more familiar with the dialect being spoken?

Talker Dialect

Statistically, the talker's dialect matters. There was a statistically significant difference between the SRT scores produced using speech audiometry materials made with Mainland talkers and those produced with Taiwan talkers. Mainland talkers resulted in SRT scores that were better by an average of 1.74 dB. So in theory, clinicians with clients of either the mainland Pǔtōnghuà or the Taiwanese dialects would want to obtain and use speech audiometry materials in the mainland Mandarin dialect of Pǔtōnghuà.

However, there are two caveats to consider before applying the results. First, the statistical significance may possibly be attributed to one or more confounding variables. For example, the idiolect of the talkers who made the recordings may be a variable unaccounted for. The particular acoustic variations unique to each individual's manner of speech may have made enough of a difference as to affect the

listeners' scores. Some subjects reported that the Mainland talker spoke more slowly and deliberately, which may have made him more intelligible than the Taiwan talker. This would explain why talker dialect was found to be significant, but not the interaction between talker dialect and subject dialect (that is, the Mainland recordings produced better scores for all subjects, while using the recording that matched the subject's dialect did not improve scores). Thus, the statistical results may be more of a comment on the way the two sets of recordings were made rather than on the nature of the two dialects themselves.

Other factors could be considered to explain the results, too, besides a difference in idiolect. The two Mandarin dialects themselves may be so similar that they produce no difference in score. Alternately, the type of recordings made may have affected the outcomes, or the method used when balancing the word lists to be psychometrically equivalent (Jennings, 2005; Slade, 2006).

The second caveat is that, even if the talker's dialect matters statistically in this case, it does not matter clinically, at least not for the two dialects investigated in this study. Other studies of SRT have deemed such a small difference in SRT not clinically significant. For example, Schneider (Schneider, 1992) found that for the 12 subjects of her study, the differences in the mean SRTs relative to the subjects' PTAs for each of the three dialects tested ranged from 1.9 to 2.4 dB (a greater difference than the 1.74 dB difference found between the two dialects of this study). This difference was not deemed clinically significant.

For clinical purposes, a difference of 2 dB in an SRT will not significantly impact clinical decisions. So in practice, for example, a clinician would be able to use either set of speech audiometry materials safely and without negatively affecting the client's treatment. This conclusion is consistent with other studies of dialect and SRT, such as Schneider (1992), and may help clinicians and institutions with limited time and funds.

Scorer Dialect

Based on the analysis conducted in this study, the scorer's dialect does not matter, either. The percentage of agreement between the two native scorers, whose native dialects differed from each other, was so high that any discrepancies between scores would have no impact. In practice, a Mandarin-speaking scorer could be used to assess clients of either dialect safely and without negatively affecting the clients' treatment. This conclusion may also be good news to those of limited means.

However, caution must be employed when putting into practice the conclusions of this study. While the attempt was made to eliminate lurking variables, the possibility remains that other factors may account for the high percentage agreement between the two scorers. Considering that fact that this portion of the study is based on the comparison of only two participants, personal background could have greatly affected the outcome. One or both of the scorer's pre-existing familiarity with the non-native dialect may have given them an advantage that would not be held by another clinician who is trying to put the results of this study into practice by scoring an SRT test that involves a talker and/or a subject of another Mandarin dialect.

Limitations

The scope of this project includes native Mandarin speakers who speak the Pǔtōnghuà and Taiwanese dialects. Therefore, conclusions are necessarily limited to those parameters, and no definitive statements can be made regarding other dialects or other languages. English is too different from Mandarin to apply the findings of this study. Two dialects of another language (e.g., Russian, German) may have differences of greater magnitude than those between Pǔtōnghuà and Taiwan Mandarin. Therefore, it is incorrect to conclude from this study that dialectical differences do not matter in any language when obtaining SRTs.

Likewise, conclusions cannot safely be drawn about other dialects of Mandarin. Two other dialects beside the ones examined in this study may differ in ways more significant than Pǔtōnghuà and Taiwan Mandarin. It would thus be incorrect to conclude that dialectical differences do not matter in any dialect of Mandarin just because no clinical difference was found in this study. The two dialects in question may be very similar compared to other dialects of Mandarin. The nature and degree of such similarities, however, is also beyond the scope of this study.

Future Research

Several of the questions raised in this study could be grounds for further studies. Beside paving the way for other researchers who would like to shed light on related topics, such as Standard American English, studies could address the concerns specific to Mandarin dialects. For example, a similar study that used a larger sample of word lists spoken by more than one talker of each dialect would increase the robustness of the

results. Likewise, a study that used more than one scorer of each dialect, particularly if controlling for differences in backgrounds, would make the percentage agreement result more representative of and applicable to situations encountered in the field.

References

- American National Standards, I. (1999). *Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms*. ANSI S3.1-1999. New York: ANSI.
- ASHA. (1988). Guidelines for determining threshold level for speech. *American Speech-Language Hearing Association*, 30, 85–89.
- ASHA. (1990). Guidelines for screening for hearing impairments and middle-ear disorders. *ASHA*, 32, 17–24.
- Campbell, G. L. (1991). *Compendium of the World's Languages*. London & New York: Routledge.
- Cheng, R. L. (1985). A comparison of Taiwanese, Taiwan Mandarin and Peking Mandarin. *Language and Cognitive Processes*, 61, 352–377.
- Dennis, J. M., & Neely, J. G. (1991). Basic hearing tests. *Otolaryngologic Clinics of North America*, 24, 253–276.
- Jennings, L. J. (2005). *Psychometrically equivalent digital recordings for speech audiometry testing in Mandarin Chinese: Standard Mandarin dialect*. BYU, Provo.
- Lee, Y. L. (1981). A study on code-switching in Taiwan. *Stud Ling Sci*, 11, 121–136.
- Li, D. C. C. (1984). *The sociolinguistic context of Mandarin in Taiwan: Trends and developments*. Paper presented at the Fourteenth International Conference on Sino-Tibetan Languages and Linguistics.
- Lin, J., & Staecker, H. (2006). Nonorganic hearing loss. *Seminars in Neurology*, 26, 321–330.
- Roup, C. M., Wiley, T. L., Safady, S. H., & Stoppenbach, D. T. (1998). Tympanometric screening norms for adults. *American Journal of Audiology*, 7, 55–60.

- Schneider, B. S. (1992). Effect of Dialect on the Determination of Speech-Reception Thresholds in Spanish-Speaking Children. *Language, Speech, and Hearing Services in Schools, 23*, 159–162.
- Slade, K. B. (2006). *Speech reception threshold materials for Taiwan Mandarin*. BYU, Provo.
- von Hapsburg, D., & Pena, E. D. (2002). Understanding bilingualism and its impact on speech audiometry. *Journal of Speech, Language, and Hearing Research, 45*, 202–213.
- Weisleder, P., & Hodgson, W. R. (1989). Evaluation of four Spanish word-recognition-ability lists. *Ear and Hearing, 10*, 387–392.
- Zhou, X., & Marslen-Wilson, W. (1995). Morphological structure in the Chinese mental lexicon. *Language and Cognitive Processes, 10*, 545–600.

Appendix
Informed Consent

Research Participation Form

Participant: _____ Age: _____

You are asked to participate in a research study sponsored by the Department of Communication Disorders at Brigham Young University, Provo, Utah. The faculty directors of this research are Richard W. Harris, Ph.D., and Shawn L. Nissen, Ph.D. Students in the Communication Disorders program may assist in data collection.

This research project is designed to evaluate a word list recorded using improved digital techniques. You will be presented with this list of words at varying levels of intensity. Many will be very soft, but none will be uncomfortably loud to you. You may also be presented with this list of words in the presence of background noise. The level of this noise will be audible but never uncomfortably loud to you. This testing will require you to listen carefully and repeat what is heard through earphones or loudspeakers. Before listening to the word lists, you will be administered a routine hearing test to determine that your hearing is normal and that you are qualified for this study.

It will take approximately two to three hours to complete the test. Testing will be broken up into two or three one-hour blocks. Each subject will be required to be present for the entire time, unless prior arrangements are made with the tester. You are free to make inquiries at any time during testing and expect those inquiries to be answered.

As the testing will be carried out in standard clinical conditions, there are no known risks involved. Standard clinical test protocol will be followed to ensure that you will not be exposed to any unduly loud signals.

Names of all subjects will be kept confidential to the investigators involved in the study. Participation in the study is a voluntary service and no payment of monetary reward of any kind is possible or implied. You are free to withdraw from the study at any time without any penalty, including penalty to future care you may desire to receive from this clinic. If you complete your participation in this research project, you will be paid the amount of \$10/hour for your participation.

If you have any questions regarding this research project, you may contact Dr. Richard W. Harris, 131 TLRB, Brigham Young University, Provo, Utah 84602, phone (801) 422-6460; or Dr. Shawn L. Nissen, 138 TLRB, Brigham Young University, Provo, Utah 84602, phone (801) 422-5056. If you have any questions regarding your rights as a participant in a research project, you may contact Dr. Renea Beckstrand, Chair of the Institutional Review Board, 422 SWKT, Brigham Young University, Provo, Utah 84602, phone (801) 422-3873, email renea_beckstrand@byu.edu.

YES: I agree to participate in the Brigham Young University research study mentioned above. I confirm that I have read the preceding information and disclosure. I hereby give my informed consent for participation as described.

Signature of participant

Date

Signature of witness

Date