Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 2aSC: Linking Perception and Production (Poster Session)

2aSC12. Inter-rater agreement on Mandarin tone categorization: Contributing factors and implications

Puisan Wong*, Lingzhi Li and Xin Yu

***Corresponding author's address: Otolaryngology--Head and Neck Surgery, The Ohio State University, 915 Olentangy River Road, Columbus, Ohio 43212, pswResearch@gmail.com**

Factors that may/may not influence inter-rater reliability in assessing the accuracy of monosyllabic Mandarin tones produced by children and adults were examined in three experiments. Experiment 1 investigated inter-judge reliability in 2 groups of Mandarin-speaking adults--one group from China and the other from Taiwan-- on their categorization of filtered tones produced by adults and children. The results showed that the magnitude of inter-rater agreement was associated with the production accuracy of the speakers; the judges attained lower agreement in categorizing children\'s tones than adults\' tones. All judges who indicated that Mandarin was their strongest language and that they had learned and used Mandarin since birth performed similarly in their tone categorization regardless of their place of birth or country of residence. Similar results were found in Experiment 2, in which one group of the judges in Experiment 1 categorized tones produced by a new and larger group of adults and children, and in Experiment 3, in which a different group of adults categorized another new set of tones produced by different speakers. Implications of the findings in research design are discussed. [Work supported by NIH-NIDCD (1 F31 DC008479-01A1) and NSF (OISE-0611641)].

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4799488]
Received 21 Jan 2013; published 2 Jun 2013

INTRODUCTION

Recent studies that examine speech production accuracy in special populations, such as children, bilingual speakers and individuals with communication disorders, often employ a group of native speakers to determine the accuracy of the sounds (e.g., Flege & Hillenbrand, 1984; Wong, 2013; Wong, Schwartz, & Jenkins, 2005). One assumption is that the raters will agree on their judgments and if the inter-judge agreement is low, the ratings may not be useful. However, factors that may influence inter-judge reliability in these tasks have not been systematically examined. This study investigates factors that may contribute to the variations in inter-judge reliability in the judgment of monosyllabic Mandarin lexical tones produced by native Mandarin-speaking children and adults.

Mandarin is a tonal language that uses distinctive pitch/fundamental frequency (f0) contours to mark lexical contrasts. The f0 contours for the four Mandarin tones are high and level for Tone 1 (T1), low rising for Tone 2 (T2), dipping or low falling followed by rising for Tone 3 (T3), and high falling for Tone 4 (T4). Thus, the same syllable produced with the four f0 contours are lexically different in Mandarin. For example, /di/ means 低 (low), 敌 (enemy), 底 (bottom), and 地 (floor) when produced with the four tones, respectively.

The three experiments in this study sought to answer the following research questions: (1) how does the phonetic accuracy of the speaker influence inter-rater reliability? and (2) how does the linguistic background of the judges affect inter-judge reliability of Mandarin tones?

EXPERIMENT 1

Methods

Experiment 1 in this study is a modification of Wong (2012).

Speakers

Four native Mandarin-speaking adults and 13 three-year-old (mean age: 3;0; age range: 2;10-3;4) preschool children learning Mandarin as their first language in the U.S. participated in the study.

Task for Speakers

The adult- and child-speakers labeled 24 pictures of familiar objects or actions (4 tones x 6 monosyllabic words).

Stimuli for Tone Judgment

Altogether 92 monosyllabic words in the four tones (24, 24, 24, and 20 for T1, T2, T3, and T4, respectively) produced in isolation were collected from the 4 adults and 198 monosyllabic productions (55, 54, 47, and 42, for the four tones, respectively) were collected from the 13 children. To control for lexical expectation and biases in tone judgment, adults' and children's productions were low-pass filtered at 400 Hz and 500 Hz, respectively to eliminate lexical information and retain f0 information in the productions.

Judges

Two groups of native-Mandarin speakers listened to the filtered stimuli of the adults' and children's productions. One group of judges were five undergraduate and graduate students (3F, 2M) in Chung Cheng University in Taiwan (TJ). The other group of judges were 10 graduate students (6F, 4M) recruited in New York (UJ). Five of the UJ came from Taiwan and five came from China. All judges in the two groups indicated that Mandarin was their strongest language and all of them, except TJ05 and UJ08, reported learning Mandarin at birth. TJ05 learned Taiwanese at birth and Mandarin at the age of seven years. UJ08 learned Shanghainess, a Chinese dialect, at birth and Mandarin at the age of three years.

Tasks for Judges

The judges listened to 2 training blocks and 17 experimental blocks of stimuli. The first training block included 12 naturally produced (unfiltered) monosyllabic Mandarin syllables and the second training block had 12 naturally produced non-sense syllables. All judges attained over 90% accuracy in the training blocks. The 17 experimental blocks contained the filtered stimuli produced by the 17 adults and children.

All filtered stimuli were presented to the judges one at a time via headphones at a comfortable hearing level. The judges listened to the stimuli as many times as needed and clicked on the corresponding label (Tone 1, Tone 2, Tone 3, or Tone 4) on the computer screen to indicate their judgment of the tones.

Data Analysis and Results

Fleiss Kappa statistics were used to assess the reliability in the tone ratings among the 5 TJ, 10 UJ and all 15 judges (5 TJ + 10 UJ) as a group. As shown in Table 1, all 3 groups of judges showed substantial to almost perfect agreement on adults' tones (Range of Fleiss Kappa: 0.76-0.97). Inter-judge reliability among the TJs (Range of Kappas: 0.76-0.93) appeared to be slightly lower than among the UJs (Range of Kappas: 0.83-0.97). All three groups of judges had lower inter-judge reliability in their categorization of children's tone productions, with Fleiss Kappa ranged from fair to substantial agreement (Range of Fleiss Kappa coefficient: 0.37-0.79). Again, the reliability among the TJs were slightly lower than the OJs.

TABLE 1. Overall inter-judge reliability on the three groups of judges on their ratings on the adult- and child-produced tones. Fleiss Kappas of 0-.20, .21-.40, .41-.60, .61-.80, and .81-1.00 represent slight, fair, moderate, substantial, and almost perfect agreement, respectively.

Judges	Speakers	Tone 1	Tone 2	Tone 3	Tone 4	All Tones
5 TJ	4 Adults	0.90	0.78	0.76	0.93	0.84
10 UJ	4 Adults	0.94	0.83	0.83	0.97	0.89
5 TJ + 10 UJ	4 Adults	0.93	0.81	0.80	0.96	0.87
5 TJ	13 Children	0.64	0.59	0.40	0.69	0.58
10 UJ	13 Children	0.79	0.61	0.37	0.70	0.62
5 TJ + 10 UJ	13 Children	0.73	0.59	0.38	0.70	0.60

Cohen Kappas were used to perform pair-wise comparisons among the 15 judges on their reliability in the judgment of children's and adults tone productions. The results showed that all judges attained almost perfect reliability with all the other judges on adults' productions (Range of Cohen Kappas: 0.80-0.97), except for TJ05 who reached almost perfect agreement with UJ02 (Cohen Kappas: 0.82) but substantial agreement with all the other 13 judges (Range of Cohen Kappas: 0.72-0.78). TJ05 also showed lower agreement with the other judges on children's tone productions (Cohen Kappas ranged from 0.44-0.58). The other fourteen judges reached moderate to substantial agreement among each other on children's tone productions (Cohen Kappas ranged from 0.48 to 0.77), lower than their agreement on adults' productions.

Given the lower agreement between TJ05 and other judges. TJ05 was excluded. Fleiss Kappas with the exclusion of TJ05 showed improved inter-judge judge reliability. Inter-judge reliability among the 4 TJs, the 10 UJs and the 14 TJs and UJs all reached almost perfect agreement on adults' tone productions (Range of Fleiss Kappas: 0.82-0.97) and mostly substantial agreement in children's tone productions (Ranged of Fleiss Kappas: 0.59-0.79) except for T3, which yielded moderate to fair agreements (Fleiss Kappas ranged from 0.37-0.44).

Perceptual accuracy of the judges' rating of the children's and adults' tone productions were examined. Table 2a and Table 2b show the results. As shown, the two groups of judges rated adults' tones with higher accuracy rates than children's tones (Chi-square test: p-values ranged from .000-.002).

Table 2a. Judgment Accuracy of Adult's Tone Productions

	T1	T2	T3	T4	T1	T2	T3	T4
	Judged by 4 TJ (Exclude TJ05) (%)				Judged by UJ (%)			

Target Tone	T1	94	6	0	0	96	4	0	0
	T2	0	98	2	0	1	96	3	0
	T3	0	14	86	0	0	17	83	0
	T4	0	0	5	95	0	0	2	98

Table 2b. Judgment Accuracy of Children's Tone Productions

		T1	T2	T3	T4		T1	T2	T3	T4
		Judged by 4 TJ (Exclude TJ05) (%)					Judged by UJ (%)			
Target Tone	T1	68	16	5	10	78	11	2	9	68
	T2	5	64	28	3	5	70	19	6	5
	T3	1	37	51	11	2	40	44	14	1
	T4	2	2	15	80	4	2	18	76	2

Correlation between the Fleiss Kappa coefficients and the accuracy rates of the two groups of judges on their categorizations of the four tones produced by children and adults was examined. Pearson correlations of 0.95 and 0.97 for the TJs and UJs indicated very strong association between speakers' production accuracy and degree of agreement among the judges (Figure 1).

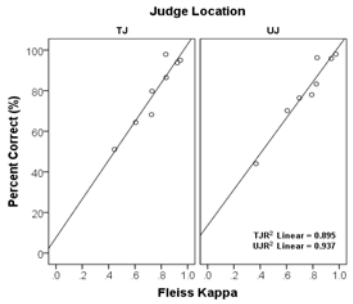

FIGURE 1. Correlations between Fleiss Kappas and accuracy of the judges on categorizing the four tones produced by the 13 children and 4 adults.

EXPERIMENT 2

Methods

The experimental procedures in Experiment 2 were the same as in Experiment 1. The differences between the 2 experiments were that only 4 TJs participated in this study (TJ05 was excluded due to lower inter-judge reliability with other judges) and that the 4 TJs categorized the tones produced by a new group of adults who produced Taiwanese influenced Mandarin and the tones produced by a new group of Mandarin learning children growing up in Taiwan.

Speakers

Seven Mandarin-speaking adults with Taiwanese exposure and 35 three to five-year-old children residing in Taiwan labeled the same 24 pictures as in Experiment 1. None of the speakers participated in Experiment 1 or 3.

Stimuli for Tone Judgment

The seven Taiwan adults and the 35 children produced, 290 and 792 monosyllabic words, respectively. Following the procedures in Experiment 1, adults' tones were low-pass filtered at 400 Hz while children's tone productions were low-passed at 500 Hz.

Judges

The four TJs, with the exclusion of TJ05, in Experiment 1 took part in Experiment 2. They listened to the 1082 filtered stimuli and determined the target tones.

Data Analysis and Results

Table 3 shows the results of Fleiss Kappa analysis of the inter-judge agreement of the 4 TJ as a group. Consistent with the findings in Experiment 1, the judges reached a higher inter-judge reliability on adults' productions than on children's productions. However, the Fleiss Kappa coefficients on adults' production were lower in this experiment than in Experiment 1 (Compare Table 1 and 3).

TABLE 3. Fleiss Kappas of the four TJs (excluded TJ05) on their ratings of the tones produced by the 7 Taiwan adults and 35 Taiwan children. Fleiss Kappas of 0-.20, .21-.40, .41-.60, .61-.80, and .81-1.00 represent slight, fair, moderate, substantial, and almost perfect agreement, respectively.

Judges	Speakers	Tone 1	Tone 2	Tone 3	Tone 4	All Tones
4 TJ	7 Taiwan Adults	0.90	0.78	0.76	0.93	0.84
4 TJ	35 Taiwan Children	0.64	0.59	0.40	0.69	0.58

Analysis of judgment accuracy showed that the TJs were less accurate in categorizing the tones produced by the 7 adults in this experiment than by the 4 adults in Experiment 1 (Compare Tables 2a and 4).

TABLE 4. Judgment accuracy of the 4 TJs on the tones produced by the 7 adults and 35 children in Taiwan

		T1	T2	T3	T4	T1	T2	T3	T4
		Productions by 7 Taiwan adults (%)				Productions by 35 children (%)			
Target Tone	T1	74	26	0	0	64	25	7	4
	T2	0	78	22	0	5	52	39	4
	T3	0	30	70	0	4	30	53	13
	T4	0	0	2	98	3	1	16	80

Pearson correlation between the Fleiss Kappa coefficients and the judges' accuracy on the adults' and children's four tones showed strong correlations (r = 0.85, see Figure 2, left penal).

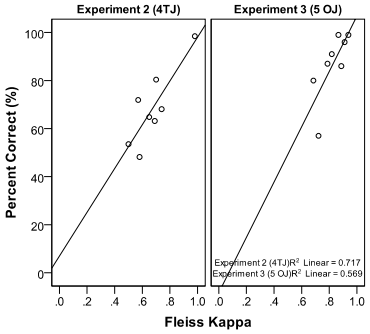

FIGURE 2. Correlation between Fleiss Kappa coefficients and accuracy of the judges on their identification of the four tones produced by children and adults. The left penal shows the results of Experiment 2 in which the 4 TJs judged the tones produced by 7 Taiwan adults and 35 Mandarin-speaking children in Taiwan. The right penal shows the results of Experiment 3 in which 5 OJs judged the tones produced by 35 Mandarin-speaking adults and 43 Mandarin-speaking children in the U.S.

Inter-judge reliability between each pair of the judges examined by Cohen Kappa analysis showed substantial to near perfect agreements on adults' productions (Range of Cohen Kappas: 0.70-0.87) and moderate to substantial agreements on children's productions (Range of Cohen Kappas: 0.53-0.77), consistently indicating that the judges were less reliable in their judgment of children's productions.

EXPERIMENT 3

Methods

This experiment followed the same experimental procedures as in the previous two experiments. The differences between this and the previous 2 experiments were that this experiment employed a new group of judges and a new group of speakers who did not participate in the previous two experiments and that the words produced by these speakers were mostly different from those in Experiments 1 and 2.

Speakers

Twenty-five native Mandarin-speaking adults and 43 two- to six-year-old Mandarin-speaking children residing in the U.S. labeled 12 monosyllabic words of the four tones in a picture naming task. None of the speakers participated in Experiment 1 or 2.

Stimuli for Tone Judgment

The 25 adults and 43 children produced 591 and 687 monosyllabic words, respectively. The adults' tones were low-pass filtered at 400 Hz while children's tone productions were low-passed at 500 Hz.

Judges

Five Mandarin speaking adults (OJ) were who did not participate in Experiment 1 or Experiment 2 listened to the 1287 filtered stimuli and determined the target tones. All these judges (3F, 2M, age range = 18;11-28;4) were undergraduate or graduate students coming from China and studying in the U.S. All of them indicated Mandarin as their strongest language and language at birth.

Data Analysis and Results

Table 4 shows the results of Fleiss Kappas of the inter-judge agreement of the judges. The results were consistent with the findings in the previous two studies. Inter-judge agreements among the judges were higher for adults' than children's tone productions.

TABLE 4. Fleiss Kappas of the OJ on their ratings of the tones produced by the 25 adults and 43 children. Fleiss Kappas of 0-.20, .21-.40, .41-.60, .61-.80, and .81-1.00 represent slight, fair, moderate, substantial, and almost perfect agreement, respectively.

Judges	Speakers	Tone 1	Tone 2	Tone 3	Tone 4	All Tones
5 OJ	25 Adults	0.91	0.87	0.89	0.94	0.90
5 OJ	43 Children	0.79	0.68	0.72	0.82	0.75

Table 5 shows the accuracy rates of the judges' identification of the tones produced by the 25 adults and 43 children.

TABLE 5. Judgment accuracy of the 5 OJs on the tones produced by the 25 adults and 43 children

		T1	T2	T3	T4	T1	T2	T3	T4
		Productions by 25 adults (%)				Productions by 43 children (%)			
Target Tone	T1	96	4	0	0	87	11	1	1
	T2	0	99	1	0	6	80	11	3
	T3	1	10	86	3	5	22	57	16
	T4	1	0	0	99	7	0	2	91

Again, strong correlation was found between the Fleiss Kappa Coefficients and the judges' accuracy on identifying the four tones produced by the 25 adults and 43 children ($r = 0.72$, Figure 2, right penal). Inter-judge reliability between each pair of judges indexed by Cohen Kappas showed near perfect agreements on adults' productions between all pairs of judges (Cohen Kappas ranged from 0.84-0.94) and substantial to near perfect agreements on children's tone productions (Range of Cohen Kappas: 0.65-0.83).

DISCUSSION

The three experiments above examined inter-judge reliability of native Mandarin speakers with different linguistics backgrounds on their categorization of monosyllabic tones produced by adults and children. The three groups of judges came from different geographic locations (i.e., different parts of China and Taiwan) and resided in different language environments (i.e., Taiwan and the U.S.). The three groups of speakers in the three experiments also came from different linguistic backgrounds (i.e., native-adult Mandarin speakers residing in the U.S., Mandarin and Taiwanese speaking adults, and Mandarin-speaking children growing up in the U.S. and in Taiwan).

Despite the differences in the linguistic backgrounds, all three groups of judges showed near perfect agreement on their categorization of the tone produced by 2 groups of Mandarin-speaking adults residing in the U.S. in Experiments 1 and 3. Their agreement on the categorization of the tones produced by the Mandarin-speaking adults with Taiwanese influence was lower (Experiment 2). The lowest agreement were consistently found in the judges' categorization of children's productions (Experiments 1-3). Correlation analyses confirmed that the degree of agreement among the judges was positively and strongly correlated with the speakers' tone accuracy. The lower inter-judge reliability on children's T3 productions seems to be another evidence to support this claim, given that T3 had the lowest accuracy rate among the four tones in children's productions (Table 2b).

Judges' geographic location does not seem to be a factor that influences inter-judge reliability. Judges from China and Taiwan residing in Taiwan or the U.S. all performed similarly and their ratings were highly reliable with

judges from a different geographical location. The fact that TJ05 had similarly poorer agreement with all other judges regardless of their linguistic background (Experiment 1), that judges from Taiwan categorized tones produced by adult Mandarin speakers in Taiwan less reliability and less accurately (Experiment 2) than the tones produced by Mandarin-speaking adults residing in the U.S. (Experiment 1), and that the two groups of judges in Experiments 1 and 3 recruited at different time and different locations performed similarly in categorizing the tones produced by different groups of adults (Experiments 1 and 3) is further evidence that support the observation. In short, it appears that judges who learned Mandarin at birth and maintain Mandarin as their strongest language all performed similarly regardless of their place of birth or place of residence.

CONCLUSION AND IMPLICATIONS

The findings of this study suggest that native Mandarin-speakers from China and Taiwan who are residing in the original country or in the U.S. perform similarly in their judgments of Mandarin tones. Thus, in future studies, when selecting judges, it is less important to control for the country of origin or the place of residence of the judges than to control for the language at birth and language proficiency. The findings also allow comparison of results in different studies employing judges from different geographic locations. Moreover, given the high inter-judge reliability among the judges, fewer judges, likely 3-5 judges, would be sufficient to do monosyllabic tone categorization in filtered speech.

The findings of this study also suggest that the magnitude of inter-judge agreement is affected by the phonetic accuracy of the productions. Less accurate productions are more ambiguous and less likely to be categorized into the same category by different listeners. Thus, for future studies that examine speech sound accuracy in speaker groups that may have low accuracy rates (e.g., individuals with speech sound disorders), it would be helpful to build in a control group with high accuracy rates to determine the reliable the judges. When setting exclusion criteria for the judges, some adjustments have to be made depending on the production accuracy of the target population.

ACKNOWLEDGMENTS

This work was supported by NIH-NIDCD (1 F31 DC008479-01A1) and NSF (OISE-0611641) to the first author. The authors would like to thank the participants and the members in the Speech and Language Acquisition Lab in the Department of Otolaryngology at the Ohio State University for their hard work and supports.

REFERENCES

Flege, J. E., & Hillebrand, J. (1984). Limits on pronunciation accuracy in adult foreign language speech production. *Journal of the Acoustical Society of America*, 76, 708–721.

Wong, P. (2012). Monosyllabic Mandarin tone productions by three-year-old children growing up in Taiwan and the U.S.: Inter-judge reliability and perceptual results. *Journal of Speech, Language, and Hearing Research*, 55, 1423-1437.

Wong, P. (2013). Perceptual evidence for protracted development in monosyllabic Mandarin lexical tone production in preschool children in Taiwan. *Journal of Acoustical Society of America*, 133(1), 434-443.

Wong, P., Schwartz, R.G., & Jenkins, J.J. (2005). Perception and production of lexical tones by 3-year-old Mandarin-speaking children. *Journal of Speech, Language and Hearing Research*, 48, 1065-1079.