Apparent-talker height is influenced by Mandarin lexical tone

Santiago Barreda[a] **and Zoey Y. Liu**
Department of Linguistics, University of California, Davis, Davis, California 95616, USA
sbarreda@ucdavis.edu, yiliu@ucdavis.edu

Abstract: Apparent-talker height is determined by a talker's fundamental frequency (*f*0) and spectral information, typically indexed using formant frequencies (FFs). Barreda [(2017b). J. Acoust. Soc. Am. **141**, 4781–4792] reports that the apparent height of a talker can be influenced by vowel-specific variation in the *f*0 or FFs of a sound. In this experiment, native speakers of Mandarin were presented with a series of syllables produced by talkers of different apparent heights. Results indicate that there is substantial variability in the estimated height of a single talker based on lexical tone, as well as the inherent *f*0 and FFs of vowel phonemes.

[SKL]
Date Received: September 27, 2017 **Date Accepted:** January 10, 2018

1. Introduction

Listeners use the acoustic information in speech sounds to estimate the height of the apparent talker. In general, lower average fundamental frequencies (*f*0s) and formant frequencies (FFs) are associated with taller talkers (Van Dommelen and Moxness, 1995; Rendall *et al.*, 2007). Most previous research on the perception of apparent height has assumed that listeners base apparent-height estimates on information related to the aggregate spectral or source characteristics of voices (e.g., mean *f*0, mean FF). However, any given speech sound represents a conflation of talker-specific information (e.g., mean *f*0, mean FF) and phoneme-specific information (e.g., vowel-intrinsic *f*0 and formant-pattern). As a result, listeners would have to control for the linguistic content of an utterance in order to estimate aggregate, phoneme-independent acoustic characteristics for a talker from a limited amount of speech. In contrast, if listeners were simply using acoustic information directly with no controls, we might see apparent-talker size vary greatly between different linguistic items produced by a single talker.

In a series of experiments, Barreda (2016, 2017a,b) reports that vowel-specific acoustic information affects apparent-talker height so that vowels with lower formants are associated with taller talkers. In addition, because of the inverse relationship between *F*1 and *f*0 in naturally-produced vowels (Whalen and Levitt, 1995), low vowels in Barreda (2017b) were associated with taller talkers because of their lower *f*0, despite having higher *F*1s. Although phoneme-specific effects on talker height are large enough to meaningfully affect apparent-talker height, they are not as large as they might be given that between-phoneme acoustic differences are usually much larger than between-talker differences. As a result, the findings reported in Barreda (2016, 2017a,b) suggest that listeners do control for linguistic content to some extent, though not completely, when estimating apparent height.

Phoneme-dependent variation in apparent-talker height may seem counterintuitive; however, the patterns of phoneme-effects reported in Barreda (2016, 2017a,b) are in line with patterns of sound-size symbolism frequently reported in the literature. Specifically, it has been noted that low and back vowels are associated with lexical items denoting large sizes, and high and front vowels are associated with small sizes, more often than one would expect by chance alone across languages (Ohala, 1997). Barreda (2017b) suggests that between-phoneme effects on apparent-height may help explain these cross-linguistic patterns of sound symbolism, and that perhaps both phenomena arise because of the same associations between sounds and sizes.

Shinohara and Kawahara (2010) report an experiment where participants were presented with disyllabic nonce adjectives from an "unknown" language. These adjectives were all VCVC words where the first and second vowel always matched,

[a]Author to whom correspondence should be addressed.

and could be one of /a e i o u/. Participants were asked to guess the size of the objects described by each word. Mandarin speakers consistently provided a ranking of /i/ < /e/ < /u/ < /o/ < /a/ with respect to associated sizes, a similar pattern to that reported for English speakers in Barreda (2017b). This suggests that Mandarin listeners may exhibit similar phonemic effects on their apparent-height judgments as English listeners. However, it is not clear what effect, if any, lexical tone may have on the perception of apparent-talker height.

In addition to phoneme-specific acoustic variability, tonal languages such as Mandarin have specific *f*0 contours associated with specific lexical items, independent of phonemic content. For example, the syllable /ma/ can take on four different meanings based on its *f*0 contour: (1) high level (mā, 妈, "mom"), (2) high rising (má, 麻, "numb"), (3) low dipping (mǎ, 马, "horse"), (4) high falling (mà, 骂, "scold"). The current experiment investigates whether this suprasegmental, within-talker variation in speech acoustics affects apparent height independent of phoneme-specific variation in *f*0 and FFs between vowels.

2. Methods

2.1 Subjects

Listeners were 20 undergraduate students (19 females, 1 male) at the University of California, Davis. Listeners participated in the experiment for partial course credit. All listeners were native speakers of Mandarin who had moved to the United States after 15 yrs of age, and 16 of 20 listeners had lived in the United States for 2 yrs or less.

2.2 Stimuli

Stimuli were resynthesized natural productions of 12 Mandarin morphemes produced by an adult male native-speaker of Mandarin. These natural productions were manipulated to make four synthetic "talkers" meant to vary in apparent height. Stimulus morphemes differed in vowel phoneme and/or lexical tone (Table 1). Stimuli were manipulated and synthesized using STRAIGHT (Kawahara *et al.*, 2008), which decomposes speech sounds into source information, and information regarding the filter (resonator) the source was passed through. Using STRAIGHT, the *f*0 and formant pattern of speech sounds can be independently manipulated and resynthesized.

Differences in apparent vocal-tract length were simulated by linearly scaling the spectral envelope of stimuli up or down by a given multiplicative scale-factor. Differences in average *f*0 were also implemented by increasing or decreasing the *f*0 of the naturally produced stimuli up or down by given multiplicative factors. In both cases, these manipulations result in translations of *f*0 contours and formant patterns along a logarithmic frequency axis (Fig. 1), but no other differences in the temporal or source characteristics of the stimuli.

The four synthetic talkers differed in their apparent *f*0 and spectral scaling in four equal, correlated steps, resulting in 48 unique stimuli.[1] The lowest size level, meant to be perceived as a tall talker, was made by decreasing the spectral envelope of the natural productions by a factor of 0.95 and the *f*0 by a factor of 0.86. After this, each size level was made by increasing the previous *f*0 level by 1.31 and the spectral-scale factor by 1.094, so that each successive size level would be perceived as being relatively shorter. Scale factor magnitudes were selected so that *f*0 and formant ranges would span from those appropriate for an adult male to those appropriate for a prepubescent boy. The end result was the highest size level having FFs 1.31 times higher than the lowest, and the highest *f*0 being 2.25 times higher than the lowest.

2.3 Procedure

Listeners were presented with stimuli over headphones in a sound-attenuated booth. Each stimulus was presented 3 times, blocked by repetition but randomized along all other stimulus dimensions, for a total of 144 responses per listener. Listeners were told they would be hearing the voices of a series of male talkers varying in age from

Table 1. Information regarding stimulus morphemes used in the experiment.

Character	妈	麻	马	骂	眯	谜	米	秘	摸	魔	抹	默
Tone	1	2	3	4	1	2	3	4	1	2	3	4
Pinyin	mā	má	mǎ	mà	mī	mí	mǐ	mì	mō	mó	mǒ	mò
Meaning	mom	numb	horse	scold	squint	puzzle	rice	secret	touch	monster	wipe	silent

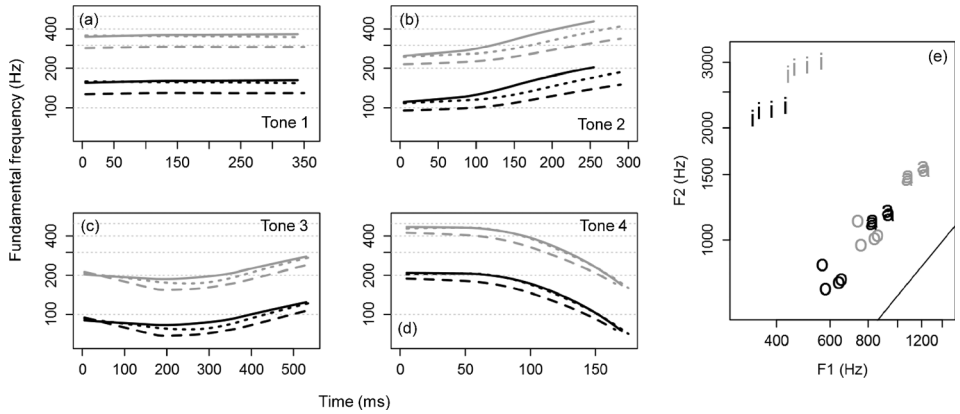

Fig. 1. (a)–(d) Tone contours for each lexical tone and vowel, for the voice at the lowest (black line) and highest (gray line) size levels. Tone contours for different vowels are indicated by line type: /i/ (solid), /o/ (dotted), /a/ (dashed). (e) *F*1 and *F*2 at vowel midpoint for each stimulus morpheme for the voice at the lowest (black line) and highest (gray line) size levels.

children to adults. They were presented with a single morpheme at a time and, for each trial, they were asked to indicate: (1) The morpheme they had just heard, and (2) the height of the apparent talker. Morpheme identifications were entered by clicking on one of four buttons, each one displaying the character for a /mV/ morpheme with the same vowel as the stimulus morpheme, but with one of the four Mandarin tones. Morpheme was correctly identified in 98% of trials. Apparent height was indicated using a ruler presented on a computer monitor. The ruler spanned from 90 to 200 cm with indications at 115, 145, and 175 cm. When listeners clicked on the ruler, the selected height was displayed rounded to the nearest centimeter. Listeners could provide responses in any order. When listeners were satisfied with their responses, they clicked on a button marked "submit" and the next stimulus played after a 1 s pause. Listeners were allowed to replay stimuli up to 3 times per stimulus presentation.

3. Results and discussion

Results were analyzed using a random-coefficients regression analysis (Gumpertz and Pantula, 1989), which fits linear models to the data from each listener and then analyzes the distribution of relevant coefficients across listeners. For each model, apparent height was regressed on: size (a 4-level factor), tone (a 4-level factor), and vowel (a 3-level factor). Effect coding was used for all predictors so that estimated effects represent deviations from the overall mean response provided by each listener.

There were significant main-effects for tone [$F(3,17) = 12.7$, $p < 0.001$], size [$F(3,17) = 63.7$, $p < 0.001$], and vowel [$F(2,18) = 17.0$, $p < 0.001$] on apparent height. Information about individual effects is presented in Table 2 and Fig. 2(a). Significant effects for vowels and tones indicate consistent within-talker variation in apparent height on the basis of linguistic content, relative to the overall mean response. To investigate the relative importance of the different predictors, a regression model was fit to responses across all listeners, standardized within-listener, with size, vowel, and tone as predictors. This model indicates that these three predictors explain 68.3% of

Table 2. Model results across all listeners. Means reflect the mean height response (in cm) for each predictor level across all differences, effects represent these means as deviations (in cm) from the overall grand mean of 166.9 cm. The significance tests for each coefficient indicate when the predictor is significantly different from the overall grand mean.

	Tone				Size				Vowel		
	1	2	3	4	1	2	3	4	/a/	/i/	/u/
Mean	166.7	165.3	168.3	167.2	177.1	170.1	164.9	155.3	169.0	164.5	167.1
Effect	−0.16	−1.59	1.47	0.29	10.25	3.27	−1.93	−11.59	2.15	−2.41	0.26
S.D.	1.55	1.38	1.36	1.79	3.47	3.31	2.55	5.68	1.73	1.89	1.24
$t(19)$	−0.47	−5.16	4.83	0.72	13.2	4.41	−3.38	−9.13	5.56	−5.69	0.92
p	0.64	<0.001	<0.001	0.48	<0.001	<0.001	0.003	<0.001	<0.001	<0.001	0.37

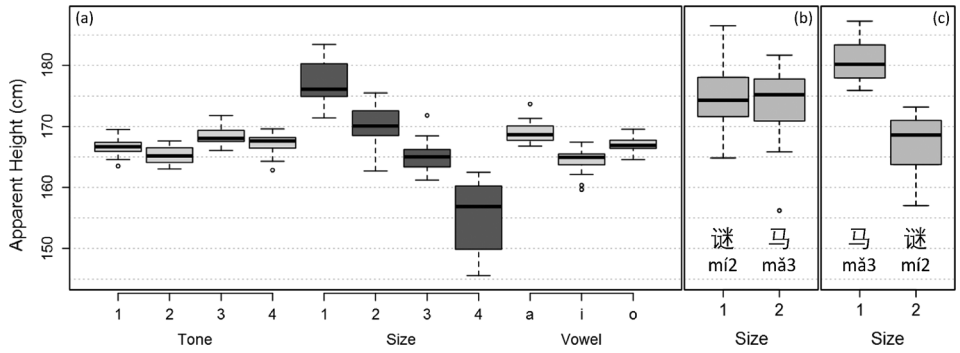

Fig. 2. (a) Distribution of estimated effects across all subjects. (b) and (c) Average apparent-height within-listener for sizes 1 and 2 when presented with specific vowel and tone combinations meant to (b) minimize and (c) maximize the perceived size difference between the two talker sizes.

the variance in apparent-height responses, with size accounting for 92.3% of this variance, tone accounting for 2.3%, and vowel accounting for 5.4% of this total.

The size effects represent average apparent-heights for each size level across all tokens, thus, the size predictors represent "between-talker" variation in apparent height. We may consider apparent-height responses across the size levels presented in Table 2 and Fig. 2 relative to the distribution of urban Chinese male's heights reported in Zong and Li (2013). The average 18 yr-old Chinese male is 172.7 cm tall, and 178.7 cm is one standard deviation above average. The average height of 14 yr-old males is 165 cm, and 11 and 13 yr-old boys are 145.3 and 159.5 cm tall, respectively. These ranges indicate that the size manipulations were successful, resulting in apparent-heights spanning from a tall adult male to an 11 yr-old boy.

Although size was the most important predictor of apparent height, there was also significant variation on the basis of tone and vowel, representing "within-talker" variation in apparent height. The pattern of vowel effects is similar to the pattern reported previously for English (Barreda, 2017b) and Mandarin listeners (Shinohara and Kawahara, 2010), with /a/ being associated with the tallest talkers, /i/ associated with the shortest talkers, and /o/ falling in between. As described in Barreda (2017b), phoneme effects on apparent height can be largely understood in terms of the inherent spectral and *f*0 characteristics of different phonemes. It also appears likely that the effects of tone on apparent height can be understood in terms of the acoustic characteristics of the different tones. For example, tone 3 was associated with the lowest average *f*0 (102 Hz) and the tallest talkers.

However, the use of *f*0 in apparent height estimation also appears to involve consideration of specific aspects of the *f*0 contour, rather than simply relying on average *f*0. Tone 2 was associated with the smallest talkers, despite having a lower mean *f*0 (152 Hz) than tone 1 (172 Hz) or tone 4 (186 Hz). One possible explanation is that tone 2 has a rising intonation and, as noted by Ohala (1997), rising intonation is cross-linguistically used to form questions, and to show politeness, more often than would be expected by chance alone. These relationships are hypothesized to have arisen out of a subconscious association between rising *f*0 contours and smaller sizes. In any case, the association of specific *f*0 contours with taller or shorter talkers would mean a more complicated use of *f*0 in apparent-height estimation than is typically considered, and suggests that prosodic information may also affect apparent height in non-tonal languages such as English.

Variation in apparent height between different tones and vowels was large enough to meaningfully affect the apparent height of talkers of different sizes. An example of this is given in Fig. 2(b). The largest difference in apparent-height between vowel categories was 4.6 cm between /a/ and /i/, while the largest difference in apparent-height between tones was 3 cm between tone 2 and 3. As a result, although size level 1 was judged to be about 7 cm taller on average than size level 2, when size 1 was paired with the tone and vowel categories associated with the shortest talkers (Tone 2, /i/, 谜), and size 2 was paired with the tone and vowel categories associated with the tallest talkers (Tone 3, /a/, 马), the difference between the two sizes decreases to 0.9 cm. Conversely, when size level 1 was presented with 马 and size level 2 was presented with 谜, the difference between the levels nearly doubles to 13.9 cm [Fig. 2(c)].

Although the vowel and phoneme effects on apparent height were large enough to noticeably affect height estimates, they are not as large as they might be

given the substantial within-talker acoustic differences between tones and phonemes (see Fig. 1). For example, average *f*0 for tone 1 was 69% higher than average *f*0 for tone 3, and *F*2 for /i/ is nearly 300% higher than *F*2 for /o/. These differences are much larger than the 30% increase in *f*0 and 9% increase in FFs between adjacent size levels. Despite this, the tone and phoneme effects on apparent height are not larger than those of the size levels. As discussed in Barreda (2017b), this suggests that listeners do correct for linguistic information when estimating apparent-talker height, to some extent.

4. Summary and conclusion

Mandarin listeners were presented with morphemes varying in phonemic content and average acoustic characteristics, and were asked to estimate the height of the talker that produced them. There were large effects for average *f*0 and FFs on apparent height, as well as significant effects for tone and vowel category. Barreda (2016, 2017a,b) has previously reported that phoneme-specific formant and *f*0 patterns can affect apparent height, and the results presented here indicate that suprasegmental information can affect apparent-height as well.

Given that most apparent-talker characteristics are strongly influenced by *f*0, and that *f*0 can vary dramatically between tones, the perception of other apparent talker characteristics such as age or sex may also be affected by lexical tone in tone languages. Despite this potentially-interesting relationship, previous studies have mainly focused on examining the role of apparent-talker characteristics on tone perception rather than vice versa (Moore and Jongman, 1997; Wong and Diehl, 2003). In light of the results reported here, the relationship between speech perception, apparent-talker characteristics, and speech acoustics in tone languages warrants further investigation.

Acknowledgments

We would like to thank Zhuang Qiu for producing the stimuli used in this experiment, and Georgia Zellou for her helpful comments.

References and links

[1]The different talkers were made using correlated logarithmic shifts in *f*0 and FFs in order to preserve the naturalness and the linguistic identity of the stimuli. Independent manipulation of individual formants or specific details of *f*0 contours were not carried out as these will potentially result in shifts in linguistic identity, which would complicate the interpretation of the height judgments made by listeners. A consequence of this is that the independent contributions of individual cues (e.g., *F*1, *F*2, initial *f*0, duration) cannot be effectively investigated because of high correlations intrinsic to the phonological system. As a result, the tone and vowel effects in the analysis should be thought of as representing holistic effects that comprise a set of important cues. For an investigation into the independent contributions of vowel-intrinsic *f*0 and formant patterns to apparent-talker height, please see Barreda (2017a,b).

Barreda, S. (**2016**). "Investigating the use of formant frequencies in listener judgments of speaker size," J. Phonetics **55**, 1–18.

Barreda, S. (**2017a**). "Listeners respond to phoneme-specific spectral information when assessing speaker size from speech," J. Phonetics **63**, 1–18.

Barreda, S. (**2017b**). "An investigation of the systematic use of spectral information in the determination of apparent-talker height," J. Acoust. Soc. Am. **141**(6), 4781–4792.

Gumpertz, M., and Pantula, S. G. (**1989**). "A simple approach to inference in random coefficient models," Am. Stat. **43**, 203–210.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (**2008**). "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3933–3936.

Moore, C. B., and Jongman, A. (**1997**). "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Am. **102**, 1864–1877.

Ohala, J. J. (**1997**). "Sound symbolism," in *Proceedings of the 4th Seoul International Conference on Linguistics [SICOL]* (August 11–15, 1997), pp. 98–103.

Rendall, D., Vokey, J. R., and Nemeth, C. (**2007**). "Lifting the curtain on the wizard of Oz: Biased voice-based impressions of speaker size," J. Exp. Psychol. **33**(5), 1208–1219.

Shinohara, K., and Kawahara, S. (**2010**). "A cross-linguistic study of sound symbolism: The images of size," in *Proceedings of the 36th Annual Meeting of the Berkeley Linguistics Society*, Berkeley, CA (February 6–7, 2010).

Van Dommelen, W. A., and Moxness, B. H. (**1995**). "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," Lang. Speech **38**, 267–287.

Whalen, D. H., and Levitt, A. G. (**1995**). "The universality of intrinsic F0 of vowels," J. Phon. **23**, 349–366 (1995).

Wong, P. C., and Diehl, R. L. (**2003**). "Perceptual normalization for inter-and intratalker variation in Cantonese level tones," J. Speech, Lang., Hear. Res. **46**(2), 413–421.

Zong, X. N., and Li, H. (**2013**). "Construction of a new growth references for China based on urban Chinese children: Comparison with the WHO growth standards," PLoS One **8**(3), e59569.