The pitch levels of female speech in two Chinese villages

Diana Deutsch[a)]
Department of Psychology, University of California, San Diego, La Jolla, California 92093
ddeutsch@ucsd.edu

Jinghong Le
School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China
jhle@psy.ecnu.edu.cn

Jing Shen and Trevor Henthorn
Department of Psychology, University of California, San Diego, La Jolla, California 92093
jshen@psy.ucsd.edu, trevor@music.ucsd.edu

Abstract: The pitch levels of female speech in two villages situated in a relatively remote area of China were compared. The dialects spoken in the two villages are similar to Standard Mandarin, and all subjects had learned to read and speak Standard Mandarin at school. Subjects read out a passage of roughly 3.25 min in Standard Mandarin, and pitch values were obtained at 5-ms intervals. The overall pitch levels in the two villages differed significantly, supporting the conjecture that pitch levels of speech are influenced by a mental representation acquired through long-term exposure to the speech of others.

PACS numbers: 43.71.Bp, 43.71.Es, 43.71.An [JH]
Date Received: January 24, 2009 **Date Accepted:** March 12, 2009

1. Introduction

While a substantial literature exists concerning the features of particular languages and dialects, overall pitch level as a feature has so far received little attention. This is due in part to the assumption frequently made that the pitch level of speech is physiologically determined and that it serves as a reflection of body size (Kunzel, 1989; Van Dommelen and Moxness, 1995). However, taking male and female speech separately, a convincing absence of correlate has been obtained between overall pitch level and the speaker's body dimensions such as height, weight, size of larynx, and so on (Hollien and Jackson, 1973; Kunzel, 1989; Van Dommelen and Moxness, 1995; Collins, 2000; Gonzales, 2004; Lass and Brown, 1978; Majewski *et al.*, 1972). In contrast, various studies have found that the pitch level of speech varies with the speaker's language (Hollien and Jackson, 1973; Majewski *et al.*, 1972; Hanley *et al.*, 1966; Yamazawa and Hollien, 1992), indicating that it is subject to a cultural influence (Dolson, 1994; Honorof and Whalen, 2005; Xue *et al.*, 2002), though the precise nature of this influence has not received much consideration.

Deutsch and co-workers (cf. Deutsch, 1992) proposed that the pitch level of an individual's speaking voice is strongly influenced by the pitch levels of speech in his or her linguistic community. More specifically, it was hypothesized that through long-term exposure to the speech of others, the individual acquires a mental representation of the expected pitch range and pitch level of speech (for male and female speech taken separately), and that this representation includes a delimitation of the octave band in which the largest proportion of pitch values occurs. It should be noted that the pitch range of speech has frequently been determined to be roughly

[a)]Author to whom correspondence should be addressed.

an octave, for both male and female speakers, and across a diversity of languages and dialects (Dolson, 1994; Hudson and Holbrook, 1982; Kunzel, 1989; Majewski *et al.*, 1972; Xue *et al.*, 2002; Yamazawa and Hollien, 1992; Hanley *et al.*, 1966; Hollien and Jackson, 1973). A detailed account and appraisal of the proposed model can be found in Dolson (1994). It leads to the further assumption that, taking two communities, each of which is linguistically homogeneous, the overall pitch levels of speech should cluster within each community, but might differ across communities. It is further hypothesized that when acquiring a second language or dialect, the individual imports the mental representation of the pitch levels that he or she had originally acquired (Deutsch *et al.*, 2004).

The present study was designed to test the above hypothesis by comparing the overall pitch levels of female speech in two communities. The subject populations were located in two villages situated in a relatively remote area of China, which is considered to be stable and homogeneous in terms of ethnicity, culture, and lifestyle (Blunden and Elvin, 1998). The villages are less than 40 miles apart, though travel time between them by automobile takes several hours. The dialects spoken in these villages are quite similar, being in the same general family as Standard Mandarin, and communication between residents of the two villages using their native dialects occurs without difficulty. Further, the speech of residents of both villages can be readily understood by speakers of Standard Mandarin. All subjects in the study had learned to speak and read Standard Mandarin in school.

Each subject was given a passage of roughly 3.25 min in duration to read out in Standard Mandarin, and from this reading, pitch values were obtained at 5-ms intervals. Two types of analysis were then performed. First, for each subject an average fundamental frequency (F0) was obtained, and statistical comparison was made between the average F0s obtained from subjects in the two villages. Second, for each subject the F0s were allocated to semitone bins, a histogram was created of the percentage occurrence of F0 values in each bin, and from this histogram the octave band containing the largest number of F0 values was derived. Statistical comparison was then made between the positions of the octave bands derived from subjects in the two villages.

2. Method

2.1 Subjects

Thirty-three female subjects participated in the experiment. They were tested in two locations: 17 subjects in Taoyuan Village, near Guandu, Zhushan County, in Hubei Province, and 16 subjects in Jiuying Village, near Bailu, Wuxi County, in the municipality of Chongqing. Those tested in Taoyuan Village were of average age 33.7 years (18–48 years) and had received an average of 7.6 years (4–12 years) of school education where they had learned to speak and read Standard Mandarin. They had all been born in or near Guandu and had not lived outside Zhushan County for more than 5 years. Those tested in Jiuying Village were of average age 37.6 years (25–52 years) and had received an average of 7.1 years (3–12 years) of school education where they had learned to speak and read Standard Mandarin. They had all been born in or near Bailu and had not lived outside Wuxi County for more than 5 years. All subjects except two were married, and their husbands were all locally born. With two exceptions, the subjects' parents had been born in the same county as the subjects and had lived in the same county for most of their lives. All subjects reported that they had normal hearing and were free of respiratory illness at the time of testing.

2.2 Apparatus and procedure

The subjects in both locations were tested individually in a quiet environment. They were first interviewed to inquire into their state of health and hearing and to determine that they had an adequate level of competence in speaking Standard Mandarin. They were also administered a questionnaire that inquired into their linguistic background and life history. Then they were given a short, emotionally neutral article to read out in Standard Mandarin for practice. Following this, they were given the test article to read out in Standard Mandarin, and their speech was

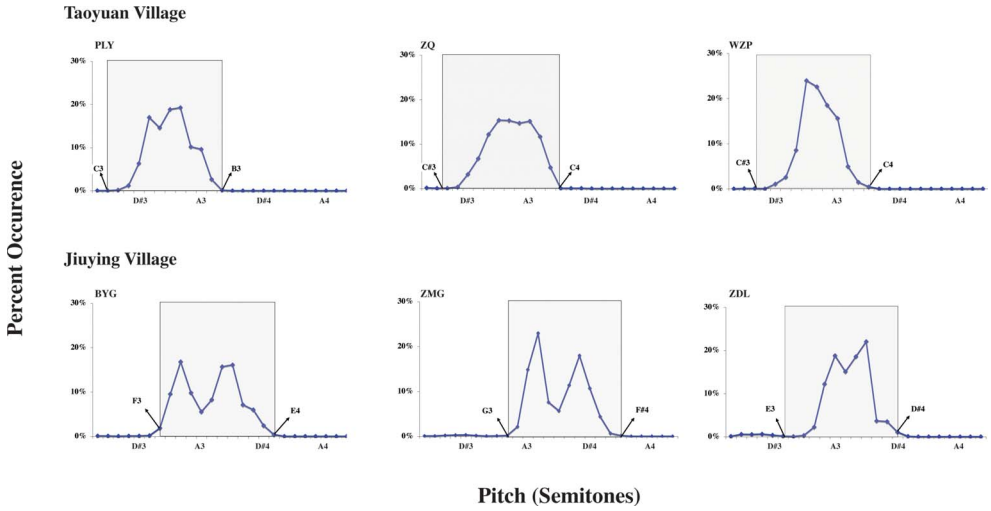

Fig. 1. (Color online) The percentage occurrence of F0 values in a 3-min segment of speech plotted in semitone bins. The data from three subjects in each village are displayed. The center of each bin is displayed on the abscissa: D#3=155.6 Hz; A3=220 Hz; D#4=331.1 Hz; A4=440 Hz. The gray area on each histogram shows the octave band in which the largest number of F0 values occurred.

recorded. The test article was also emotionally neutral, contained 480 Chinese characters, and took an average of roughly 3.25 min to read out. The subjects were paid for their services.

For each subjects' reading, the speech samples were recorded via a SONY ECM-CS10 lavalier microphone onto an Edirol R-1 digital recorder as 16-bit, 44.1-kHz WAV files. The sound files were transferred to an iMac running OSX 10.5, converted to a sampling rate of 11.025 kHz, and the first 20 s of each file was deleted. The soundfiles were then converted to NeXT format and transferred to a NeXT computer (NeXTstation Turbo Color).

The sound files were lowpass filtered with a cutoff frequency of 1300 Hz. F0 estimates were then obtained at 5-ms intervals, using a procedure derived from Rabiner and Schafer (1978). The low and high boundaries for the F0 estimates were set at 107 and 639 Hz, respectively. In addition, for each subject's recording, the time-varying energy level of the signal was obtained, and only those F0 estimates that were associated with levels no lower than 25 dB below the peak level were saved for further analysis. (This procedure was employed so as to eliminate spurious F0 estimates, such as obtained during pauses in the subject's speech.) Then for each subject's reading, the F0 estimates were averaged along the musical scale; that is, along a log frequency continuum, so producing an average F0 for each subject. Furthermore, as a separate procedure, the raw F0 estimates were allocated to semitone bins, with the center frequency of each bin determined by the equal-tempered scale (A=440 Hz). Histograms were then generated for each subject showing the percentage occurrence of F0 values in each semitone bin.

3. Results

Figure 1 presents, as examples, the histograms showing the percentage occurrence of F0 values in each semitone bin derived from the readings of six subjects taken individually—three from Jiuying Village and three from Taoyuan Village. Also indicated on each histogram are the semitone bins delimiting the octave band containing the largest number of F0 values in the subject's speech. (We note that some of the histograms are bimodal and hypothesize that this reflects the characteristics of the tones in the subjects' speech.) Taking all those from Taoyuan Village, the F0 values included in the octave bands comprised 98.91% of the total, and taking all those from

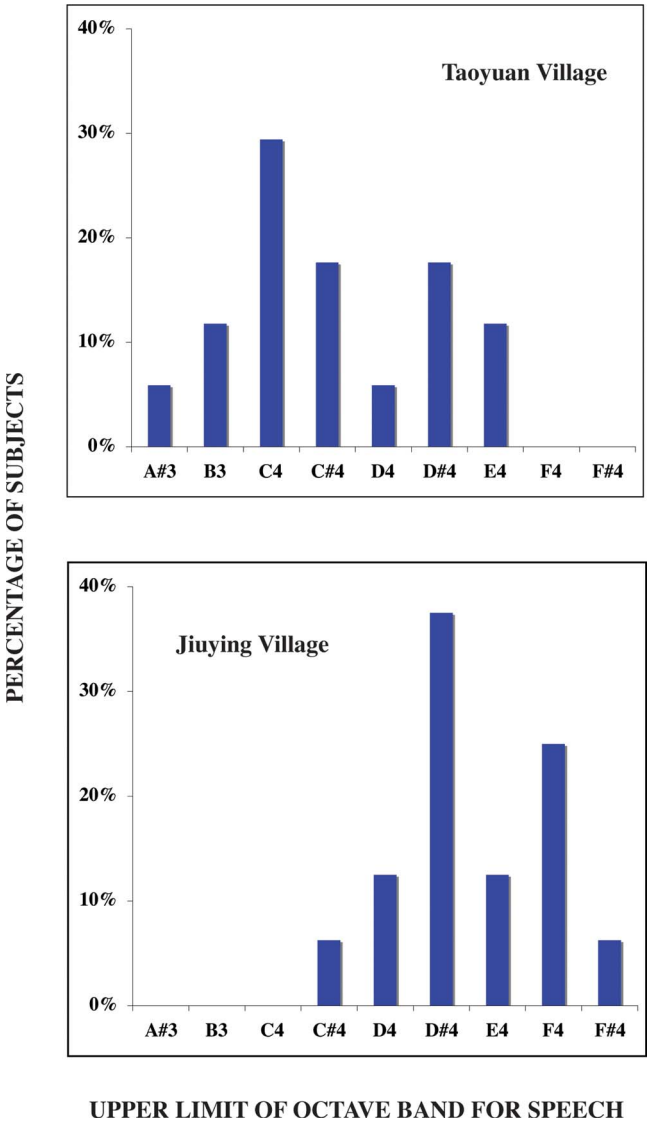

Fig. 2. (Color online) The upper limits of the octave band for speech in the two villages plotted in semitone bins. The center of each bin is displayed on the abscissa: A#3=233.1 Hz; B3=246.9 Hz; C4=261.6 Hz; C#4=277.2 Hz; D4=293.7 Hz; D#4=311.1 Hz; E4=329.6 Hz; F4=349.2 Hz; F#4=370 Hz.

Jiuying Village, these values comprised 97.24% of the total. Therefore, as had been found in the earlier studies referred to above, the F0 values in the speech of these subjects, taken individually, spanned close to an octave.

The F0s were found to be higher overall for subjects in Jiuying Village than those in Taoyuan Village; specifically, the F0s averaged over a log scale were 231.4 Hz for Jiuying Village and 200.6 Hz for Taoyuan Village. On a one-way analysis of variance (ANOVA), the difference in average F0s between the two villages was found to be highly significant [$F(1,31) = 19.106$; $p < 0.001$].

A further analysis was performed to test the hypothesis that the octave bands for speech would cluster within each village, but would differ significantly across villages. Figure 2 presents the percentages of subjects for whom the upper limit of the octave band fell in each

semitone bin, plotted for each village separately. As can be seen, the values indeed clustered within each village, but differed overall across villages by roughly 3 semitones. On a one-way ANOVA, this difference between the two villages was found to be highly significant [$F(1,31) = 19.803$; $p < 0.001$].

4. Discussion

The present findings are in accordance with the conjecture that the overall pitch level of an individual's speaking voice varies as a function of his or her linguistic community and so reflects an influence of long-term exposure to the speech of others (Deutsch, 1992). In most previous work on this issue, comparison was made between the pitch levels of speech derived from passages that were read out in different languages (see, for example, Hollien and Jackson, 1973; Majewski *et al.*, 1972; Hanley *et al.*, 1966; Xue *et al.*, 2002). Since readings of different words were compared, this procedure introduced possible confounds. As an exception, Yamazawa and Hollien (1992) tested two groups of female speakers—one speaking primarily Japanese and the other speaking primarily American English—with all subjects reading out passages in both Japanese and English. The authors found that the Japanese speakers exhibited higher F0s than did the English speakers, though the differences between the two groups were more pronounced for passages read out in the speakers' native languages. The authors concluded that these F0 differences could be due to a number of factors, including ethnicity, culture, and the substantial differences in the structural characteristics of the Japanese and English languages.

In the present study, the subjects in the two communities are considered to be homogeneous ethnically and culturally, and their dialects are quite similar. They also read out the identical passage in Standard Mandarin so that no confound could have been introduced by differences in the words that were spoken. Our present findings are therefore in accordance with the hypothesis that the overall pitch level of a speaker's voice is influenced by a mental representation that is acquired through exposure to the speech of others. Assuming a relatively homogeneous linguistic community, such a representation would be particularly useful for tone languages. Here individual words assume different lexical meanings depending on the tones in which they are enunciated. For example, the first tone in Mandarin is high in pitch, and the word "ma" when spoken in this tone means "mother." In contrast the overall pitch level of the third tone is low, and the word "ma" spoken in this tone means "horse." An agreed-upon pitch level (taking male and female speech separately) would therefore facilitate the identification of individual tones and so the comprehension of individual words. Such a pitch representation would also be useful for speakers of nontone languages, for example, in facilitating speaker identification and evaluating the emotional tone of the speaker's voice. However, it remains to be determined whether effects similar to those found here occur in speakers of nontone languages also.

Acknowledgments

Diana Deutsch and Jinghong Le contributed equally to this paper and should both be considered first authors. The authors are grateful to Yu Chen and Rong Zhou for research assistance. They are also grateful to James Hillenbrand and two anonymous reviewers for helpful comments on an earlier version.

References and links

Blunden, C., and Elvin, M. (**1998**). *Cultural Atlas of China*, 2nd ed. (Checkmark Books, New York).
Collins, S. A. (**2000**). "Men's voices and women's choices," Anim. Behav. **60**, 773–780.
Deutsch, D. (**1992**). "Some new pitch paradoxes and their implications," Philos. Trans. R. Soc. London, Ser. B **336**, 391–397.
Deutsch, D., Henthorn, T., and Dolson, M. (**2004**). "Speech patterns heard early in life influence later perception of the tritone paradox," Music Percept. **21**, 357–372.
Dolson, M. (**1994**). "The pitch of speech as a function of linguistic community," Music Percept. **11**, 321–331.
Gonzalez, J. (**2004**). "Formant frequencies and body size of speaker: A weak relationship in adult humans," J. Phonetics **32**, 277–287.
Hanley, T. D., Snidecor, J. C., and Ringel, R. L. (**1966**). "Some acoustic difference among languages," Phonetica **14**, 97–107.

Hollien, H., and Jackson, B. (**1973**). "Normative data on the speaking fundamental frequency characteristics of young adult males," J. Phonetics **1**, 117–120.
Honorof, D. N., and Whalen, D. H. (**2005**). "Perception of pitch location within a speaker's F0," J. Acoust. Soc. Am. **117**, 2193–2200.
Hudson, A. I., and Holbrook, A. (**1982**). "Fundamental frequency characteristics of young black adults: Spontaneous speaking and oral reading," J. Speech Hear. Res. **25**, 25–28.
Kunzel, H. J. (**1989**). "How well does average fundamental frequency correlate with speaker height and weight?," Phonetica **46**, 117–125.
Lass, N. J., and Brown, W. S. (**1978**). "Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies," J. Acoust. Soc. Am. **63**, 1218–1220.
Majewski, W., Hollien, H., and Zalewski, J. (**1972**). "Speaking fundamental frequency of Polish adult males," Phonetica **25**, 119–125.
Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ).
Van Dommelen, W. A., and Moxness, B. H. (**1995**). "Acoustic parameters in speaker height and weight identification: Sex-specific behavior," Lang Speech **38**, 267–287.
Xue, S. A., Hagstrom, F., and Hao, J. (**2002**). "Speaking fundamental frequency characteristics of young and elderly bilingual Chinese-English speakers: A functional system approach," Asia Pac. J. Speech, Lang. Hear. **7**, 55–62.
Yamazawa, H., and Hollien, H. (**1992**). "Speaking fundamental frequency pattern of Japanese women," Phonetica **49**, 128–140.