Comparison of speaking fundamental frequency in English and Mandarin

Patricia Keating[a] and Grace Kuo
Department of Linguistics University of California, Los Angeles, California 90095-1543

(Received 5 August 2010; revised 23 April 2012; accepted 31 May 2012)

To determine if the speaking fundamental frequency (F0) profiles of English and Mandarin differ, a variety of voice samples from male and female speakers were compared. The two languages' F0 profiles were sometimes found to differ, but these differences depended on the particular speech samples being compared. Most notably, the physiological F0 ranges of the speakers, determined from tone sweeps, hardly differed between the two languages, indicating that the English and Mandarin speakers' voices are comparable. Their use of F0 in single-word utterances was, however, quite different, with the Mandarin speakers having higher maximums and means, and larger ranges, even when only the Mandarin high falling tone was compared with English. In contrast, for a prose passage, the two languages were more similar, differing only in the mean F0, Mandarin again being higher. The study thus contributes to the growing literature showing that languages can differ in their F0 profile, but highlights the fact that the choice of speech materials to compare can be critical.
 [http://dx.doi.org/10.1121/1.4730893]

PACS number(s): 43.70.Kv, 43.70.Gr [AL] Pages: 1050–1060

I. INTRODUCTION

Several studies over the past decades have begun to establish that speakers of different languages or dialects may use characteristically different ranges and typical values of speaking fundamental frequency, or F0 (see Dolson, 1994, for a review). It is known that different social groups within a single language may use F0 differently (Crystal, 1969; Loveday, 1981; Graddol and Swann, 1983; Henton, 1989; Podesva, 2007). Dialects of a language can also differ: Deutsch *et al.* (2009) compared two villages of Standard Mandarin speakers and found that the "overall pitch levels" differed by about 30 Hz; Torgerson (2005) compared Taiwan and Beijing Mandarin speakers and found that the maximum and median F0 (though not the minimum F0 or F0 range) were lower for the Taiwan speakers; similarly Huang and Fon (2011). Thus it is not surprising that languages may differ as well. Some studies (e.g., Yamazawa and Hollien, 1992; Ohara, 1992; Todaka, 1993; Xue *et al.*, 2002) have even compared bilingual speakers and found they differ when speaking their two languages, thus demonstrating that such differences need not be due to physiological differences between speakers of different languages. Speaking F0 is to some extent an arbitrary aspect of speech, and a particular F0 range may be part of the phonetic structure of a language, such that in the limit, a speaker would sound non-native (have a foreign accent) using a different F0 range.

Perhaps the earliest experimental test of language differences in speaking F0 was a pair of studies begun at the University of California, Los Angeles (UCLA) in the 1960s, Hanley *et al.* (1966) and Hanley and Snidecor (1967). These studies compared the medians and standard deviations of the speaking F0s (in semitones) in readings of the Rainbow Passage (from Fairbanks, 1940) by male native speakers of English, Spanish, or Japanese and female native speakers of English, Spanish, Japanese, or Tagalog. Results were mixed, with the only clear result that the English males had the lowest median F0. However, these English values were unusually low in comparison with other studies of English summarized in Baken and Orlikoff (2000).

Later studies have compared Japanese versus English (Loveday, 1981; Yamazawa and Hollien, 1992; Ohara, 1992; Todaka, 1993); British English versus German (Mennen *et al.*, 2012); Polish versus English (Majewski *et al.*, 1972); Mandarin versus Min (Taiwanese) (Chen, 2005). Chen compared speaking F0 range (95th percentile), mean, and standard deviation in semitones from prose passage reading by Taiwan Mandarin versus Min speakers, but found no F0 differences between these two groups of speakers. She then suggested, however, that both Mandarin and Min showed wider F0 ranges than seen in most previous studies of English (though not in all); in addition, inspection of her tables on p. 3228 indicates that the Mandarin and Min mean F0s are at the low end of published values for English.

Other studies have directly compared Mandarin and English. Chen (1972) compared the mean, standard deviation, and range of F0 of 4 English and 4 Mandarin speakers (2 males and 2 females each, i.e., a very small sample) reading words and sentences. The Mandarin speakers, especially the women, had wider F0 ranges and larger standard deviations; the Mandarin women's means were lower, while the men's were the same as the English. These results are broadly in line with S. Chen's. In contrast, Eady (1982) compared several measures of F0 from passages read by male Taiwan Mandarin and English speakers. The mean F0 and measures of F0 fluctuation (dynamic movement) were all greater in Mandarin, but the standard deviation (taken as the measure of F0 range) was the same in the two languages. Xue *et al.* (2002) compared the F0 mean, standard deviation,

[a]Author to whom correspondence should be addressed. Electronic mail: keating@humnet.ucla.edu

 0001-4966/2012/132(2)/1050/11/$30.00

minimum, maximum, and range of younger and older bilingual speakers (both male and female, but analyzed together), and found that while the older bilinguals had no differences between their Mandarin and English, the younger bilinguals had lower minimum F0 and larger F0 range in their Mandarin (with no differences in maximum F0, mean F0, or standard deviation). Finally, Mang (2001) compared the longitudinal means for speaking and singing for 8 pre-school girls who were either monolingual English, or bilingual (English-Mandarin or English-Cantonese). The speaking F0 decreased over time for both language groups, but was lower in Chinese than in English from ages 2 to 5. However, when the girls were 5–6 years old, the English-speaking F0 dropped below the Chinese. In sum, there is some evidence that F0 range is greater in Mandarin than in English, but results about mean F0 are mixed.

A possible source of language differences is phonological inventory differences. For example, if one language has more (or more frequent) voiceless obstruents or high vowels than another, and if voiceless obstruents and high vowels have a raising effect on F0 (Lehiste, 1970), then those two languages might well have characteristic F0 differences, though they would likely be small. Yet another proposal (e.g., Chen, 2005) is that tone languages will have larger F0 ranges and/or higher average F0s than non-tone languages, on the assumption that lexical tones require a greater extent of F0 than intonation alone does [though see Xu and Xu (2005) for a contrary assumption]. Chen's discussion suggests that the locus of such a tone-language effect could lie specifically with the occurrence of a high level tone, as in Mandarin and Min. The typical F0 of a high level lexical tone might be systematically higher than that of high intonational tones in a language like English, or there might simply be more lexical high tones in running speech in a tone language than there are intonational high tones in running speech in a non-tone language. Liu (2002) looked at the F0 range in Mandarin by tone, and found that from Tone 1 through Tone 4, the tones have progressively larger F0 ranges. Thus an alternative scenario is that the more Tone 4s in Mandarin speech, the more likely a larger F0 range.

However, the hypothesis that tone languages as such have an *overall larger F0 range* is not supported by Eady (1982), who found no difference between English and Mandarin F0 standard deviations (the measure of range in that study). Similarly, although Chen (2005) interpreted Japanese's lexical pitch-accent as making Japanese somewhat like a tone language, there was no difference in the standard deviations between Japanese and the other languages in Hanley *et al*. (1966) or Hanley and Snidecor (1967). The hypothesis that tone languages have an *overall higher average F0* likewise receives only limited support from the studies reviewed above. Eady (1982) and (for one age group) Mang (2001) found that the mean F0 was higher in Mandarin than in English, but other studies comparing Mandarin and English have given different results.

The present study examines the idea that Mandarin and English have different characteristic F0 properties, and that these differences are related to the tonal nature of Mandarin. In addition, the type of speech sample is varied, so that any language differences can be understood in a broader context. The research questions of the present study were the following: (1) Do English and Mandarin F0 profiles differ? (2) If so, how does that difference relate to the fact that Mandarin is a lexical tone language? In addition, there was a methodological question: (3) How does the type of voice sample elicited affect the F0 characterization? We compare a variety of voice samples from male and female speakers of English and Mandarin.

II. METHODS

A. Speech samples

Previous literature (see Baken and Orlikoff, 2000, for a review) distinguishes between a speaker's physiological F0 range—the maximum F0 range the speaker's voice is capable of producing—and speaking F0, or the range of F0s a speaker habitually produces in normal speech. Obviously the former is much larger than the latter. We followed previous practice in recording speech samples for estimating both of these kinds of F0 range. For estimating speaking F0, we included more than one kind of speech sample: isolated words and connected read speech, both in a neutral style and in a livelier style. Baken and Orlikoff (2000) review the literature comparing read with spontaneous speech and conclude that because the mean F0 is only slightly higher in reading (about 0.5 to 2 ST), with no large differences in variability, F0 measures from read speech are representative of more natural speech. More recently, Torgerson (2005) found no differences between read sentences and spontaneous interviews/spoken descriptions in Mandarin in terms of median F0 (in ERB) and various measures of F0 range, reinforcing Baken and Orlikoff's conclusion.

1. Maximum F0 ranges: Tone sweeps

a. Unprompted sweeps. The first productions collected were for maximum (physiological) F0 ranges. Two sets of tone sweeps were produced by the speakers. First, speakers produced sweeps following the general procedure described by Honoroff and Whalen (2005, p. 2194), though with slight differences. In our experiment, English-speaking participants were instructed as follows (and similarly for the Mandarin speakers, but translated into Mandarin):

(1) You will record a series of "ah" sounds in which you let your voice sweep over a wide range of pitches (tones). The goal is to see how wide a range of pitches your voice can *comfortably* cover, without straining.
(2) Start by taking a big breath.
(3) Say "ah" at a comfortable, normal pitch.
(4) Then move gradually (but quickly) higher until you feel your voice break. Going into "falsetto" is fine.

They then listened to a demo recording of a rising sweep by the first author. After some practice, three rising sweeps were recorded. Falling sweeps were instructed, practiced, and recorded in the same way. Specifically, speakers were instructed to "move gradually (but quickly) lower until you feel your voice break or give out." In the recorded demo that

subjects heard, there was no creaky voice, with the lowest pitches instead being very breathy. In the literature, F0 elicitations and/or measurements generally exclude low-pitched glottal creak or creaky voice. However, fluent speech often includes creak. Therefore, we elicited a third, new, kind of F0 sweep, which explicitly asked speakers to produce low F0s in creaky voice. Speakers were asked to "move gradually (but quickly) lower, letting your voice "creak," until you can't go any lower." The recorded demo was very creaky, so this presumably helped them to understand what was meant, and later listening to the recordings showed that speakers almost always did creak. This third type will be called creaky sweeps. We refer to all three of these types of sweeps as unprompted, because each speaker determined how to perform his or her own sweeps.

b. Prompted sweeps. Based on the discussion in Baken and Orlikoff (2000), the fast-glissando procedure described in Reich *et al.* (1990) was also used to elicit speakers' F0 ranges. Reich *et al.* (1990) found that this method resulted in larger ranges than other methods that they compared. In our experiment participants heard, and simultaneously imitated, two rapid tone glides. The glides were 4 s sawtooth waves whose F0s varied linearly as in Table I, with different 2-octave ranges for men and women per the specifications of Reich *et al.* (1990). Speakers were instructed to imitate the tone glides as follows: "You are asked to imitate those sweeps as much as possible. It's not expected that any single voice can cover the whole range, but do the best you can. You should imitate the tones *while listening to them.*" They then clicked on an icon for a tone glide, and practiced imitating it. They recorded three rising sweeps followed by three falling sweeps. We refer to these two types of sweeps as prompted, meaning that each speaker heard and imitated a tone glide prompt for each production.

2. Speaking F0 ranges

Three sets of materials were used to determine the speaking F0 range and other properties of an F0 profile in real speech, specifically in reading.

a. Isolated words. A monosyllabic word was produced in isolation. For English, there was one word, sure (ʃɝ). It was produced in 4 different ways, chosen to produce different pitch contours and ranges (and yet be understandable to naive subjects), and repeated 3 times each way. Two of these will be analyzed here. They were described, and the prompts were written, as follows

1. Normal pitch: In a regular way, at a normal comfortable pitch: —Sure
2. Excited exclamation: An exclamation with more and more excitement, and higher and higher pitched voice: —Sure! —Sure!! —SURE!!!

TABLE I. Tone sweep prompt F0 ranges (beginning and end values, in Hz), taken from Reich *et al.* (1990). Each sweep prompt covers two octaves.

	Rising	Falling
men	277-1109	277-69
women	392-1568	392-98

For Mandarin, 4 different words were used, one for each of the 4 lexical tones, which involve different pitch contours (as indicated by the tone icons in the phonemic transcriptions). They were (where /ʂ/ represents a voiceless retroflex fricative):

(1) High level 師 (/ʂi ˥/ *teacher*)
(2) Mid rising 十 (/ʂi ˧˥/ *ten*)
(3) Low falling 使 (/ʂi ˨˩/ *to make (someone do something)*)
(4) High falling 示 (/ʂi ˥˩/ *to show*)

These will be referred to as "shi" words for orthographic simplicity. They were produced in three different ways, two of which will be analyzed here.

(1) Normal pitch: same as for English
(2) Exclamation: With a higher tone, as if there is an exclamation mark after the word: 師! 十! 使 ! 示!

Note that, because the instruction for the Exclamation condition in English contained multiple exclamation marks, and prompted the subjects to speak with increasing excitement, while the Mandarin instructions did not, only the first token from each exclamation will be analyzed here, to ensure greatest comparability between the languages.

b. Rainbow passage. Following many previous studies, a reading of the Rainbow passage (from Fairbanks, 1960) was also obtained from each speaker. The original English text was translated into Mandarin for the Chinese speakers. The English version includes 330 words in 19 sentences. The Mandarin version includes 444 characters in 19 sentences. Of these, 65 are Tone 1 and 131 are Tone 4 (44% of all tones). Speakers first read this passage silently to familiarize themselves with it before reading it aloud for recording, which took about 2 min.

c. Little Red Riding Hood story. Third, a livelier prose passage was recorded. The story "Little Red Riding Hood" was chosen because it contains dialog from a variety of characters, and is familiar to both English and Mandarin speakers from childhood. A shortened version of the story, which preserved a wide variety of character dialog but takes only about 4–5 min to read, was constructed by consulting the Little Red Riding Hood Project website (available at http://www.usm.edu/english/fairytales/lrrh/lrrhhome.htm). Our English version of the story has 857 words in 54 sentences. The Mandarin version has 1086 characters in 68 sentences. Speakers read this story twice, the first time in a neutral voice without acting out the characters (this served as familiarization), and the second time more dramatically, "as you would read it to a small child, in a story-teller voice, acting out the dialogs with different voices for the different characters." The dialog was printed in a different color for each character. A subset of the dialog from second reading of the story (mostly in higher-pitched voices) was analyzed for the current study. See supplementary material for the English text.[1] In the Mandarin version, this selection contained 218

characters, of which 29 were Tone 1 and 57 were Tone 4 (39% of all tones).

B. Speakers

With UCLA IRB approval, 23 American English and 23 Mandarin speakers (11 men and 12 women in each language) were recorded, with the goal of having 20 of each language after recording problems, reading errors, etc. Most were UCLA students in their late teens or early twenties, though a few were older staff or visitors. No information was collected about smoking or singing experience. The Mandarin speakers were self-described as native speakers, and their recorded speech was later verified as sounding native to the second author. All but four of the Mandarin speakers were from Taiwan; of the four mainland speakers, two were men and two were women. Speakers responded to a solicitation to "Read a few words and a short text; read the story of Little Red Riding Hood using different voices for the different characters; make high and low voice pitches," and they were paid for their time. A session lasted about 45 min plus breaks, and did not cause any noticeable voice fatigue.

C. Recording procedure

In a soundbooth in the UCLA Phonetics Laboratory, participants were seated individually in front of a laptop computer which presented the instructions for the recording session as a Powerpoint slideshow. They wore a Shure head-mounted microphone, and for part of the experiment, an earbud in one ear. The microphone signal was recorded direct-to-disk on another computer located outside the soundbooth, at a 44.1 kHz sampling rate and a 32 bit quantization rate, using an AudioBox and PCQuirerX. An assistant outside the booth operated the recording computer during the session. Participants paged through the Powerpoint slideshow, practicing each type of production as needed, and indicated to the assistant when they were ready to record each one. The assistant then saved each recording as a separate file. The assistant did not provide feedback or correction during the recording session.

The order of materials in the recording session was the same for every speaker, viz., unprompted rising sweeps, unprompted falling sweeps, unprompted creaky sweeps; prompted rising sweeps, prompted falling sweeps; isolated words; Rainbow Passage; Little Red Riding Hood in neutral voice, Little Red Riding Hood in story-teller voice.

D. F0 analysis

1. Pitchtracking

Fundamental frequencies of the tone sweeps and isolated words were measured using the STRAIGHT algorithm (Kawahara *et al.*, 1998) incorporated into VoiceSauce (Shue, 2010; Shue *et al.*, 2011). Other than setting the maximum expected F0 for a folder of utterances to a reasonable value, no parameter adjustments were made and the program ran entirely automatically; however, some pre- and/or post-processing was required. Specifically, first the "To PointProcess (periodic, cc)" function in Praat (Boersma, 2001) was used in a script to identify target voiced portions of files (tone sweeps, vowels in isolated words, all voiced intervals in read passages) and segment them in a Praat TextGrid file. Another Praat script then displayed each utterance and its TextGrid for manual checking. At this stage, utterances with recording artifacts, and errors in Mandarin tone production, were removed from further analysis, and in some cases the TextGrid segmentations were corrected. The audio and TextGrid files were then input to VoiceSauce for acoustic analysis of the segmented voiced portions. VoiceSauce computes many measures but here only the STRAIGHT F0 will be reported.

STRAIGHT pitchtracks were output either as text, or in the format for an Emu database (Harrington, 2010). Emu databases were made for the sweep and word corpora, with F0 values at 1 ms intervals. For each Emu database, the audio files were displayed along with their labels and the F0 track, and examined for gross errors of pitchtracking, which were corrected either by adjusting the interval boundaries to exclude them, or by removing the interval from the database. Then, values for the first and last 2% of each target interval were discarded to avoid F0 artifacts at segment edges. For the connected speech corpora, Emu was not used. Instead, F0 values at 10 ms intervals were output from VoiceSauce directly as text files and analyzed in Excel.

Creaky voice is pervasive in our recordings of both languages. While many studies of F0, whether of F0 range or of intonation, ignore creaky intervals, it is important to understand the extent to which they affect the measures extracted from the recordings. Though not mentioned in previous cross-language F0 studies, creaky voice can cause particular problems for pitchtrackers. These difficulties could result in misleading estimates of speakers' minimum F0s, and thus affect estimates of F0 means and ranges. Therefore we also include a comparison of manual versus automated calculation of the very lowest F0s in creaky voice. In this way, the effect of relying on automated pitchtracking can be better understood.

2. F0 measures

Baken and Orlikoff (2000) review the basic and most common measures of F0, which include measures of the average F0 (usually the mean), of the F0 variability (usually the standard deviation, or "pitch sigma"), and of the overall F0 range (either the MaximumF0-MinimumF0, or some subset which removes outliers). To these measures, we add here measures of the most extreme F0 values produced by each speaker.

All the measures used in this study are summarized in Table II. The minimum (Min) and maximum (Max) F0 value was found for each sweep/vowel/passage (depending on the corpus). Where there were multiple utterances from a speaker for one type of utterance in a given corpus, these values were averaged within speakers, giving a mean minimum F0 and a mean maximum F0 for each speaker, which will be referred to as MeanMin and MeanMax. Also from multiple Min and Max values of a single speaker, F0 extremes were identified:

TABLE II. F0 measures calculated. Mean measures here are within speakers; all the measures can be further averaged across speakers.

Measure	Abbreviation	Unit	Definition
Minimum F0	Min	Hz	Lowest F0 value in a token
Maximum F0	Max	Hz	Highest F0 value in a token
Mean Minimum F0	MeanMin	Hz	Average Min across tokens
Extreme Minimum F0	XMin	Hz	Lowest Min across tokens
Mean Maximum F0	MeanMax	Hz	Average Max across tokens
Extreme Maximum F0	XMax	Hz	Highest Max across tokens
Mean F0 Range	MeanRange	Hz	MeanMax - MeanMin
Mean F0 Range in semitones	MeanRangeST	Semitone	39.863 * log(MeanMax/MeanMin)
Extreme F0 Range	XRange	Hz	XMax - XMin
Extreme F0 Range in semitones	XRangeST	Semitone	39.863 * log(XMax/XMin)
Mean F0	Mean	Hz	Average of F0 values in a token
Mean of Mean F0	MeanMean	Hz	Average Mean across tokens
Standard deviation of F0	SD	Hz	SD of F0 values in a token
Mean of Standard deviations of F0	MeanSD	Hz	Average SD across tokens

the lowest minimum (XMin), and the highest maximum (XMax) for each speaker. Finally, the F0 values over time in each pitchtrack were used to calculate the mean and standard deviation for that track. Again, where there were multiple utterances of a given type of utterance, these were averaged within speakers (MeanMean and MeanSD for each speaker). Then in Excel four range measures were calculated for each speaker from their minimum and maximum measures: (mean) range in Hz ((Mean)Max - (Mean)Min), the same in semitones, extreme range (XRange) in Hz (XMax- XMin), and the same in semitones. Below, results are reported as means across groups of speakers.

E. Statistical analysis

Analysis of variance (ANOVAs) were performed on the F0 descriptive statistics described above, using either SPSS v.17, or the R interface to Emu. Specific tests will be described in the sections below.

III. ANALYSIS AND RESULTS

Means for all corpora are given separately by speaker language and sex in the supplementary material.

A. Prompted sweeps

We begin with the simplest corpus, the prompted sweeps, which provide an established way of estimating a speaker's physiological F0 range. While Reich *et al.* (1990) reported that they eliminated F0 values for any low-energy portions perceived as vocal fry or falsetto, these have been included in other studies of maximum F0 range (see Baken and Orlikoff, 2000), and we included all measured F0 values. Thus here, creaky voice is sometimes included in analyzed segments (whenever the pitchtracker returned values), sometimes not (whenever the pitchtracker failed). Measures of minimum, maximum, and mean F0, plus F0 range, were made separately for falling and rising sweeps. Range across falling plus rising sweeps was also calculated (MeanMax from rising minus MeanMin from falling). In contrast, the extreme minimum XMin always comes from falling sweeps (since those sweeps reach the lowest F0s) while the extreme maximum XMax always comes from rising sweeps (since those sweeps reach the highest F0s). As long as a speaker had at least one usable token of each utterance type, that speaker was included in the dataset. With these criteria, 41 of the 46 speakers were included. Two speakers (one English male, one English female) were excluded because of recording artifacts, and three speakers (all Mandarin females) were excluded because of untrackable voice qualities.

F0 measures were compared in a series of 2-way and 3-way ANOVAs (using R or SPSS) in which between-subjects factors were Language (English, Mandarin) and Sex (Male, Female), while SweepType (Rising, Falling) was a within-subject factor in some analyses. Not surprisingly, speaker Sex almost always has a significant effect on all F0 measures in Hz. Also not surprisingly, falling versus rising sweeps generally have very different Min and Max values. To avoid cluttering the statistical results, therefore, details of these significant effects will not be reported here. Only if they interact with the Language variable will they be discussed. Full results separated by Sex and Sweeptype are however given in the Supplemental Table for interested readers.

The only (near-)significant effects of Language were seen in Sex × Language interactions for the two measures of low F0, MeanMin and XMin. A significant interaction for MeanMin of falling sweeps ($F[1,37] = 6.171$, $p = 0.018$) as well as a trend for XMin ($F[1,37] = 3.938$, $p = 0.055$) both reflected relatively high Min F0 values from the English women and relatively low values from the English men, such that the Sex difference was larger in English than in Mandarin. There were no significant Language differences within either sex for MeanMin or XMin, though there were trends to significance for MeanMin. The overall F0 profile (the extreme range (as well as the mean range) and the mean), which is thus similar across the languages, is shown by Language and Sex in Fig. 1.

For the measures compared by SweepType, there was a trend to a Language × Sweeptype interaction for MeanSD, but no significant main or interaction effects involving

Language. Across all speakers, the average XRange was 770.6 Hz and 38.8 ST, the average MeanRange was 718.1 Hz and 35.1 ST, and the MeanMean was 375.4 Hz.

B. Unprompted sweeps

Next we consider the unprompted sweeps, for which a separate creaky condition was also elicited. The unprompted falling and rising sweeps were analyzed with STRAIGHT as above, and these results are presented in this section. The unprompted creaky sweeps will be presented in Sec. III C below. Intervals with pitchtracking problems were excluded from analysis as before. As long as a speaker had at least one usable token of each utterance type, that speaker was included in the dataset. With these criteria, measurements were obtained from 40 speakers (10 English women, 9 English men, 10 Mandarin women, 11 Mandarin men); 5 others (including two who were also excluded from the previous analysis) had recording artifacts and one speaker's rising sweeps could not be pitchtracked at all.

As above, the F0 measures for falling and rising sweeps were compared in a series of 2-way and 3-way ANOVAs (using SPSS) in which between-subjects factors were Language (English, Mandarin) and Sex (Male, Female), and SweepType (Rising, Falling) was a within-subject factor. Again, results related to the Sex and Sweeptype factors apart from Language will not be described, but are shown in the Supplemental Table.

There were no significant main or interaction effects with Language in any of the ANOVAs. The meanRange and XRange measures showed no Language differences. Grand means across all speakers from the ANOVAs were as follows: XMin was 110.6 Hz, XMax was 707.6 Hz, and therefore XRange was 597 Hz and 31.8 ST; the MeanMin of falling sweeps was 117.9 Hz, the MeanMax of rising sweeps 677.7 Hz, and therefore the MeanRange was 560 Hz and 29.8 ST; the MeanMean for falling and rising sweeps combined was 263.6 Hz and the MeanSD was 99.6 Hz.

Another comparison with respect to cross-language F0 that was made concerns the starting F0s self-selected by the speakers. That is, do Mandarin and English speakers choose to begin a rise or a fall at similar F0s? If not, that suggests that their normal comfortable F0s differ. Separate two-way ANOVAs (Language × Sex) were performed on the starting F0 (MeanMin) for the rising sweeps and on the starting F0 (MeanMax) for the falling sweeps. The starting F0s for the falling sweeps were the same across the two languages (263 Hz on average), but the starting F0s for the rising sweeps did differ between the languages ($F[1,36] = 5.0$, $p = 0.032$), with Mandarin speakers, both men and women, choosing higher starting F0s than the English speakers did (171 versus 145 Hz on average).

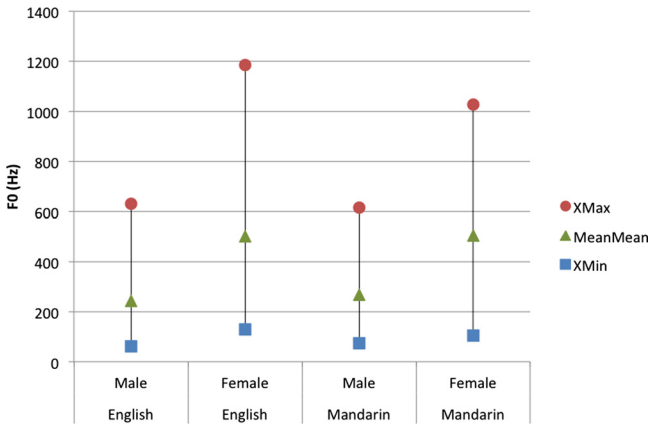

FIG. 1. (Color online) Three F0 measurements from the prompted sweeps (where MeanMean is average of rising mean and falling mean), separately by language and sex of the speakers. On each bar, the bottom number is the mean XMin (from the falling sweeps), the middle number is the mean Mean, and the top number is the mean XMax (from the rising sweeps).

C. Comparison of minimum F0 values

In the present study, for the prompted falling sweeps speakers were told nothing about how they might achieve the low F0s of the prompt. In contrast, as part of the unprompted sweeps corpus, speakers were asked to produce falling creaky sweeps, in which they allowed their voices to creak in order to reach the lowest possible pitch. Thus we can ask, in which of these two kinds of elicited sweeps did the speakers of the present study produce lower F0s? And, does the answer differ when the lowest F0 values are extracted manually, from the time-domain waveform?

XMin values were obtained manually by finding the longest single pitch pulse in the time-domain in waveforms displayed by Praat, and taking the frequency (1/T) of this pulse. These XMin values were compared to values from the STRAIGHT algorithm, as reported above. A 4-way repeated-measures ANOVA had Language and Sex as between-subjects variables and Sweeptype (unprompted creaky or prompted falling) and AnalysisMethod (hand-measured or STRAIGHT) as within-subjects variables. Only speakers for whom all four measurements were available were included, so this comparison is on a slightly smaller set of speakers than comparisons above: 9 males and 11 females for English, and 11 males and 9 females for Mandarin. Significant differences were found between the two kinds of sweeps (unprompted creaky goes lower than prompted falling, $F[1,35] = 76.346$, $p < 0.001$) and the two analysis methods (hand-measured values are lower than STRAIGHT values, $F[1,35] = 32.851$, $p < 0.001$). Significant interaction effects arise because creaky sweeps go even lower in English than in Mandarin (30 versus 55 Hz on average). One reason that the English sweeps go lower on average is that the English women go about as low as the English men do. For example, when measured by STRAIGHT, the English men's mean (and SD) are 38.1 (5.3) Hz, while the English women's are 43.3 (7.3) Hz. In contrast, in the falling sweeps when F0 is measured by STRAIGHT, the sex difference is especially large (in both languages). This is because the two methods differ more for the women's falling sweeps, where STRAIGHT misses more low values.

Thus, telling speakers to allow their voices to creak will, in general, result in lower measured Min F0 (average difference: 40 Hz), and measuring these lowest values by hand from the time waveform will result in lower measured Min

F0 (average difference: 18 Hz). Note that this is not true in every case: sometimes the STRAIGHT measurement agrees with the manual one, and sometimes the STRAIGHT measurement is incorrectly lower than the manual measurement, due to abrupt fluctuations in the STRAIGHT pitchtrack: hence the relatively small average difference. Given this relatively small difference between the methods, we conclude that XMin values can be recovered reasonably well by STRAIGHT, and in the following analyses we will rely on this method without hand-checking.

D. Comparisons of results from sweeps

In sum, there are very few ways in which speakers of the two languages differ in their physiological F0 ranges, generally involving relatively small differences in the lowest F0 values reached in creak. In addition, the Mandarin speakers chose to begin their unprompted rising sweeps at higher F0s than the English speakers did. But there were no language differences in other Min, Max, Range, or Mean measures for either the prompted or the unprompted sweeps.

Table III compares some of our results (from the two languages combined) with previous studies. In our data, prompted sweeps gave higher maximum F0s and thus larger ranges than unprompted sweeps—see also Fig. 2 below—but somewhat smaller ranges than those extracted from data of Hollien and Michel (1968) (as shown in Table VI-20 in Baken and Orlikoff, 2000) and from Reich *et al*. (1990)—even though the latter excluded perceived fry and falsetto speech from their measurements. However, our values are like those in some other previous studies summarized in Baken and Orlikoff's Table VI-16, where MeanRange for (non-elderly) men varies from 484 to 864 Hz, and for (non-elderly) women from 743 to 981 Hz.

Previous studies did not include measures which were specifically about extreme F0 values: XMin, XMax, and XRange. XRange versus MeanRange in our data are shown in Fig. 2, which also graphically compares prompted versus unprompted sweeps. It can be seen that, as expected, the ranges for prompted sweeps are consistently (though not extremely) larger than for unprompted, and the XRange is, by definition, larger than the Mean Range.

E. Isolated words

The sweeps corpora provide information about speakers' maximum physiological F0 range. The isolated words corpus, in contrast, provides information about speakers' use of F0 in their native languages. The part of this corpus analyzed here comprises productions of the English word *sure* and the 4 Mandarin "shi" words, all in two utterance types. F0 was measured by STRAIGHT as described above. Two speakers (1 English woman, 1 Mandarin woman), who had also been excluded from one or both of the previous analyses, were excluded for lack of usable tokens, giving data from 11 each of English men, English women, Mandarin men, and Mandarin women.

The most directly comparable utterances were selected for initial comparison: just the Mandarin Tone 4 utterances, which have a falling pitch contour that is most similar to the English utterances. Both normal pitch and exclamation utterances were compared; for the exclamation utterances, only the first token was selected, since that was the most comparable elicitation across the languages. Thus the measures here are within a single token for exclamations (so that XMin (etc.) are from single tokens, and there is no MeanMin (etc.)), but across repetitions for normal pitch utterances (giving a full set of measures). Three-way ANOVAs used Language and Sex as between-subjects factors and UtteranceType as a within-subjects factor. In this analysis there were main effects of Language on four measures: XMin (Mandarin 133 Hz versus English 115 Hz, $F[1,40] = 5.037$, $p = 0.03$), XMax (Mandarin 321 Hz versus English 258 Hz, $F[1,40] = 16.162$, $p < 0.001$), XRange (in Hz) (Mandarin 189 Hz versus English 143 Hz, $F[1,40] = 8.425$, $p = 0.006$), and MeanMean (Mandarin 229 Hz versus English 186 Hz, $F[1,40] = 18.383$, $p < 0.001$). All of these differences were due to higher values in Mandarin than in English. That is, the Mandarin speakers did not go as low, but did go higher, with a greater range and mean. These results are shown in Fig. 3.

The Language × UtteranceType interaction was also significant for two of these measures, XMin and MeanMean, with a trend for two more measures, XMax and XRangeST. These interaction effects were explored with *post hoc* ANOVAs (with between-subjects factors Language and Sex) separately on the normal pitch and exclamation utterances (again only the differences due to language are reported). These tests generally showed merely that the language differences are greater in the exclamations, but they also showed that for XRangeST, the only language effect was for women's normal pitch utterances, due to the very high value for the Mandarin women. Finally, the *post hoc* tests showed that while Min values were usually higher in Mandarin than in English (that is, English speakers went lower), that was

TABLE III. MeanMin (from falling sweeps), MeanMax (from rising sweeps), and MeanRange (MeanMax-MeanMin) in the present study, compared with two previous studies as cited by Baken and Orlikoff (2000). English and Mandarin data are combined here; for individual language values, consult the Supplemental Table.

	Mean Min			Mean Max			Mean Range		
	men	women	both	men	women	both	men	women	both
Prompted Sweeps	77	136	106	595	1059	827	518	923	721
Unprompted Sweeps	85	150	118	512	843	678	427	693	560
Hollien and Michel (1968)	94 (modal)	144 (modal)	119	634 (loft)	1131 (loft)	883	540	987	764
Reich *et al*. (1990)	82	150	116	623	1125	874	541	975	758

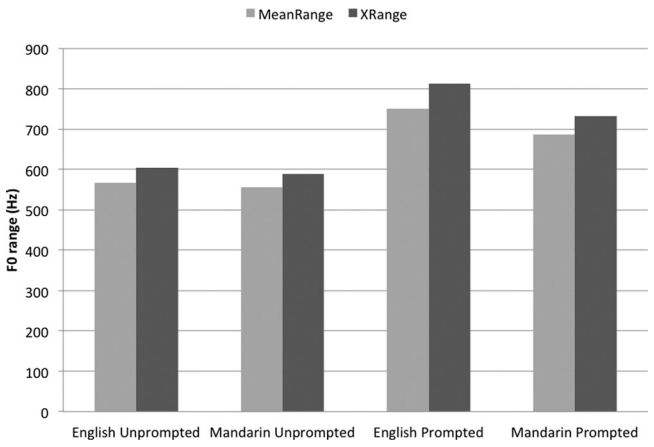

FIG. 2. Mean versus extreme F0 ranges separately by sweep type (unprompted versus prompted) and language of the speakers, averaged across men and women together. In calculating ranges, Minimum values come from falling sweeps while Maximum values come from rising sweeps.

not the case for the women's normal pitch utterances. In sum, however, the results of the *post hoc* tests generally confirm the main effects found, namely that both male and female Mandarin speakers had higher minimums, maximums, and means, and also greater ranges in Hz, such that their overall F0 ranges are shifted up but also expand somewhat. A sample comparison of two women's tokens illustrating these differences is shown in Fig. 4.

In order to see the effect that lexical tones have on the language comparison, the same analyses were then performed using the data from all four Mandarin tones (rather than just the falling Tone 4 as above). The results are very similar: compared to the English speakers, Mandarin speakers had higher XMax (332 versus 258 Hz) and MeanMean (225 versus 186 Hz) values, and greater XRanges (221 versus 143 Hz and 19 versus 14 ST), and, in Exclamations only, higher XMin values (140 versus 124 Hz); however, the English speakers had larger MeanSDs (43 versus 27 Hz). Some of these results are also shown in Fig. 3.

That is, there is little difference in F0 measures between when Mandarin speakers produce all their tones, versus just their falling Tone 4. The main exception is the standard deviation measure, which for Mandarin is much smaller in the all-tone data than in the Tone 4-only data, presumably due mostly to a lack of variation within the level tone, Tone 1. As a result, while the variability is the same for the two languages when only Tone 4 is included for Mandarin, overall

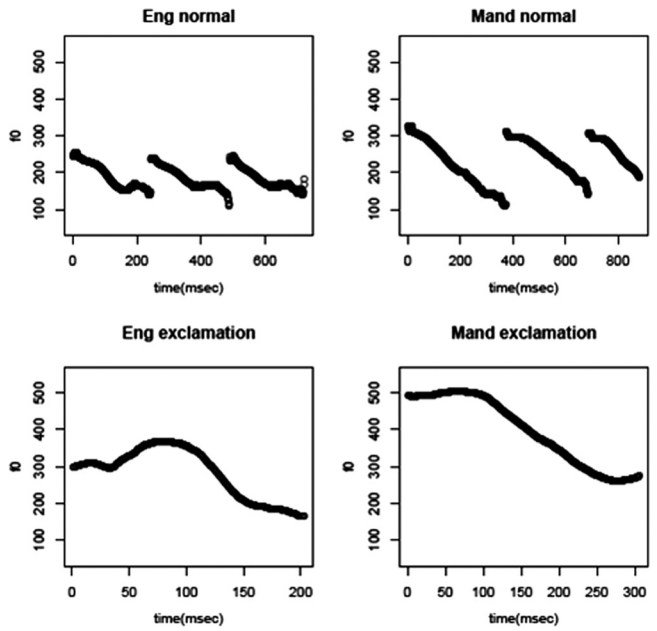

FIG. 4. Sample pitchtracks of normal pitch versus exclamation utterances from one English woman and one Mandarin woman.

variability as measured by the standard deviation is greater in English when compared to all the tones in Mandarin.

F. Rainbow passage

Like the isolated word corpus, the prose reading passage corpus provides information about speakers' use of F0 in their native languages, but in this case in connected reading. After exclusions for problems with recordings (including 4 speakers who had been excluded from one or more previous analyses), passages from 18 Mandarin speakers (9 women, 9 men) and 18 English speakers (10 women, 8 men) were available for analysis. As before, Praat TextGrids were used to segment all the voiced portions of the recordings, and they were checked to prevent spurious or null F0 measurements, generally by removing the segment labels of problematic segments. That is, not all voiced segments were tracked in their entirety. The STRAIGHT F0 values of the pitchtracked segments were then output every 10 ms, and zero values were removed. (The typical recording yielded about 4800 F0 values.) Descriptive statistics (Min, Max, Mean, SD, Range, RangeST - one value of each measure per speaker) over all of each speaker's F0 values were calculated in Excel and then

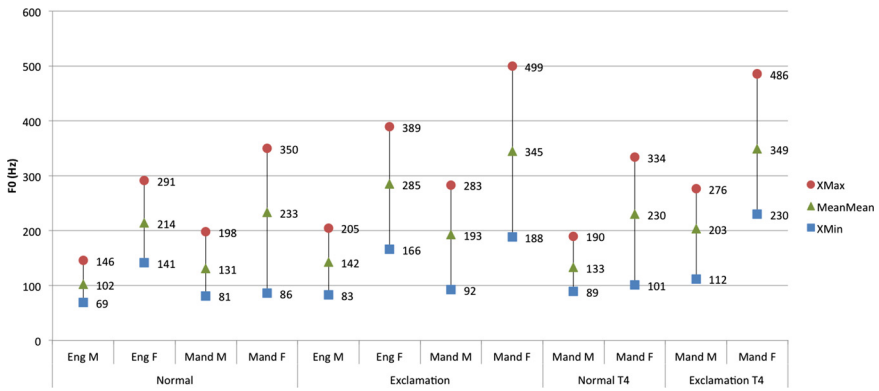

FIG. 3. (Color online) Three F0 measurements from the isolated words, separately for each combination of Mandarin tone (all tones versus only Tone 4, vacuous for English), utterance type (normal pitch versus exclamation), and language and sex of the speakers. On each bar, the bottom number is the mean XMin (which is the sole Min for the exclamations), the middle number is the mean Mean, and the top number is the mean XMax (which is the sole Max for the exclamations).

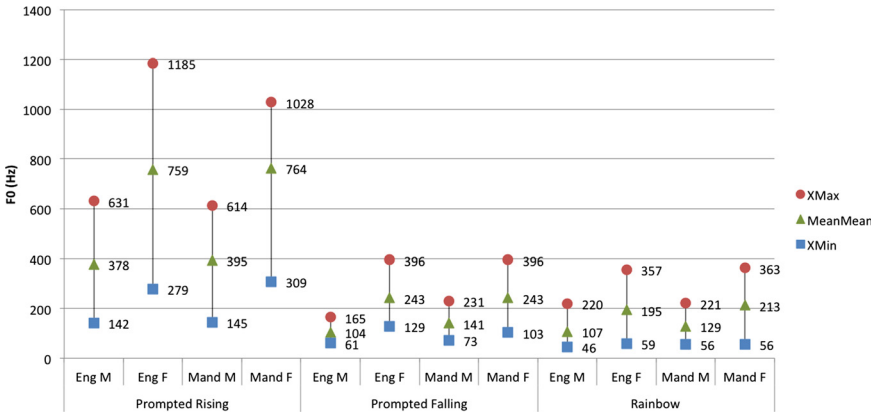

FIG. 5. (Color online) Three F0 measurements from the prompted sweeps (rising and falling) versus the Rainbow passage, separately by language and sex of the speakers. On each bar, the bottom number is the mean XMin (which is the sole Min from the Rainbow passage), the middle number is the mean Mean, and the top number is the mean XMax (which is the sole Max from the Rainbow passage).

analyzed as above by univariate ANOVAs, with Language and Sex as independent variables. Because there is only one Min and one Max value for each speaker, there are no separate XMin/XMax or MeanMin/MeanMax measures.

These analyses showed that, as expected, women had significantly higher Max, Mean, SD, and Range, and also RangeST; but the sexes did not differ on the Min F0 measures. There were no significant Sex × Language interactions. The two languages differed on only one measure, Mean ($F[1,32] = 12.437$, $p = 0.001$), with Mandarin having the higher values (171 versus 151 Hz). Thus, in this neutral reading passage, the Mandarin speakers used the same overall F0 ranges and variability as the English speakers, but they had higher average F0s within that range.

These results can be compared with those from the prompted sweeps, which tend to display speakers' widest physiological ranges. This comparison of average Min, Mean and Max F0 by language and sex is shown in Fig. 5, with the rising and falling prompted sweeps given separately. The values for the falling prompted sweeps and the reading passage are fairly similar. That is, in neutral reading, speakers on average approach their lowest F0s, and go about as high as the prompted starting F0s of the sweeps (as set by Reich *et al.*, 1990). However, in the passage reading the low F0s were very low and not different between the sexes, in accord with Blomgren *et al.* (1998), who showed that Min F0 does not differ between the sexes in vocal fry, and with our own results for English speakers on the unprompted creaky sweeps.

G. Story dialog

Speakers read the story of Little Red Riding Hood twice, once neutrally and once using different voices for the dialog by the different characters in the story. From these second dialog readings, four intervals of speech involving high-pitched voices were selected: (1) the conversation of Grandma with Wolf pretending to be Little Red Riding Hood; (2) an utterance to Little Red Riding Hood by Wolf pretending to be Grandma; (3) the conversation of Little Red Riding Hood with Wolf pretending to be Grandma; (4) an utterance to the Woodsman by Little Red Riding Hood. These intervals included both the character dialog and any narration that intervened. The pitchtracking and measuring method used for the Rainbow readings was applied to these intervals to yield the same set of F0 measures. After exclusions for problems with recordings (all of these speakers had also been excluded from one or more of the previous analyses), passages from 20 Mandarin speakers (10 women, 10 men) and 21 English speakers (11 women, 10 men) were available for analysis. As before, ANOVAs with Language and Sex as between-subjects variables were carried out on all the F0 measures.

These analyses showed that the two languages differed on most measures: Min—English 58 Hz versus Mandarin 72 Hz ($F[1,37] = 4.907$, $p = 0.033$), Max—English 627 Hz versus Mandarin 490 Hz ($F[1,37] = 12.512$, $p = 0.001$), SD—English 106 Hz versus Mandarin 78 Hz ($F[1,37] = 10.815$, $p = 0.002$), Range—English 569 Hz versus Mandarin 418 Hz ($F[1,37] = 15.644$, $p < 0.001$), and RangeST—English 41.3 ST versus Mandarin 33.5 ST ($F[1,37] = 17.834$, $p < 0.001$), but not for Mean—English 241 Hz versus Mandarin 228 Hz ($F[1,37] = 1.13$, $p = 0.295$). Many speakers, both English and Mandarin, used very high voices for all of the story characters in the sample analyzed here. Overall, though, these statistical results indicate that English speakers of both sexes went both lower and higher than the Mandarin speakers, and thus had larger F0 ranges, in both Hz and ST, as well as larger F0 standard deviations. The Mean F0, however, did not differ between the two languages.

IV. DISCUSSION AND CONCLUSION

Here we summarize and discuss the results reported above, in the context of the study's research questions:

(1) Do English and Mandarin F0 profiles differ? In the pitch sweeps (prompted, unprompted), there were very few language differences. That is, the English and Mandarin speakers in this study appear to have essentially the same physical capabilities with respect to rate of vocal fold vibration. In contrast, the speech samples showed several differences between the languages. Most often when there was a language difference the Mandarin speakers used higher F0s and/or larger F0 ranges. Previous studies, mostly based on reading passages or other connected speech, have given a variety of results. Our reading passage result, that the average F0 is higher in Mandarin, agrees with Eady (1982) and Mang (2001). Like previous studies, we also found a greater F0 range

for Mandarin, but only for the single word utterances. Note that in our data, standard deviation often does not pattern together with the range measures, indicating that the pattern of variability around the mean can be very different from the total extent of F0s produced, and thus one measure cannot be substituted for the other.

Crucially, the different speech samples showed different patterns of results. That is, whether the two languages will appear to have similar or different F0 profiles very much depends on the speech corpus or task, as well as on the acoustic measures. The inconsistency of results from previous studies may be partly due to such methodological differences.

(2) How does the type of speech sample elicited affect the F0 characterization? As just noted, physiological versus real-speech comparisons showed very different results, and even within the real-speech samples, differences were found across corpora. For single-word utterances (whether neutral or exclamatory, all Mandarin tones or Tone 4 only), language differences were clear and consistent: the Mandarin speakers had higher values on most measures. We discuss this result below, with respect to the Mandarin tones. However, for the Rainbow passage reading, which was neutral in style, there was no difference in the F0 ranges or standard deviations, but the Mandarin mean F0 was higher. Thus the two languages appear most similar in the prose passage reading. We suggest below that this pattern of results would follow if the overall reading styles are similar across the languages, but more of the Mandarin text carries high (level or falling) tones. Within their F0 range, the Mandarin speakers would then spend more of their time in the higher part of the range, giving a higher mean. Finally, for the character voices in the Little Red Riding Hood story, quite the opposite result obtained: the means were the same between the languages, but for the other measures the English values were more extreme. Here, we suppose that for whatever reason (e.g., cultural conventions about story-telling in general, or more specifically about rendering high-pitched character voices), the English speakers were more theatrical in their readings, using more extreme F0s and larger F0 ranges. Their mean F0 would also rise as a result. However, we have already noted that within similar F0 ranges, the mean F0 is higher in Mandarin than in English. Thus it is possible here that the Mandarin higher mean in a lower range happens to be about the same as the English lower mean in a higher range.

Comparison of the overall F0 characteristics of some of the samples (e.g., as seen in the figures) suggests that the Rainbow passage readings, the prompted falling sweeps, and the isolated words (in normal style) are broadly similar within each language. That is, these kinds of samples seem to give similar information about voices—how low a speaker can go, how high is comfortable for non-expressive speech, and what is a comfortable average F0—across the languages.

(3) Can any differences be related to the fact that Mandarin is a lexical tone language? In the single-word utterances, English falling intonation contours were directly compared with the four Mandarin lexical tones. It is this corpus that shows the most difference between the two languages. Even when only the Mandarin high-falling Tone 4 is compared with the falling intonation contour in English—when the overall F0 contours are most similar between the languages—the differences are robust. Thus the language differences cannot be due to the fact that Mandarin has multiple lexical tones, or that it has a high level tone. Instead, the differences must be due to the way Mandarin speakers pronounce their high-falling tone, with a higher F0 peak (perhaps serving to enhance the falling pitch contour). The higher peak results in an overall larger F0 range and higher average F0.

For the prose passage readings, the only difference between the languages is in the Mean F0, which was higher in Mandarin. The Mandarin speakers, though they covered the same F0 range as the English speakers, spoke about 20 Hz higher on average. This difference in the mean F0 might be due to the Mandarin tones: almost half (44%) of the lexical tones in the text are high level or high falling, and so involve a high pitch, and only the low-dipping tone stays in the lower half of the pitch range. Thus Mandarin pitches will tend to lie in the higher part of the overall tonal pitch range. In contrast, readings of the English passage are very unlikely to put high-toned pitch accents on half of the syllables, and so English pitches will tend to lie more in the middle part of the passage's pitch range. Such an effect of the tonal inventory of Mandarin (relatively denser high tones) versus the intonational structure of English (relatively sparser high tones) could possibly account for the 20 Hz difference in the means in the passages. Different rates of occurrence of voiceless consonant contexts might also play a small role. However, studies which tightly control the properties of the texts would be needed to establish such effects.

Our results thus provide some support for the hypothesis that a tone language like Mandarin can have an overall higher average F0, as this is what was found with both the single words and the prose passage, though not with the story voices. They also provide some, albeit weaker, support for the hypothesis that a tone language can have an overall larger F0 range, as this is what was found with the single words, though not with the prose passage or the story voices. The Tone 4 data from the single word comparisons indicates that this tone alone can produce the observed language differences. That is, we have suggested that the F0 range differences are not due to some generic property of tone languages, but specifically because of the phonetic properties of Mandarin's high falling tone; the occurrence of a high level tone may also contribute to mean F0 differences.

In addition, we asked if there is a benefit to manually measuring the lowest F0 values from the time-domain waveform. While hand-measuring did improve detection of the lowest F0 values, the overall results were not greatly different, and we concluded that they did not justify the much greater effort required. Especially for the connected prose passages, hand-measuring would have been extremely time-

consuming, and in any event could not readily provide continuous measurements over intervals of speech.

In conclusion, we found differences in the F0 profiles of Mandarin versus English speakers, but these differences depended on the particular speech samples being compared. Most notably, the overall physiological F0 ranges of the speakers, as determined from tone sweeps, did not differ between the two languages, indicating that the speakers' voices are comparable. Their use of speaking F0 in single-word utterances was, however, quite different, with the Mandarin speakers having higher maximums and means, and larger ranges, even when only the Mandarin high falling tone was compared with English. In contrast, in a prose passage, the two languages were more similar, differing only in the mean F0, Mandarin again being higher. Finally, however, in a lively reading of high-pitched story character voices, the English speakers had the higher maximums and the larger ranges. We have suggested that these language differences could be due to a combination of linguistic and cultural differences related to these speech samples.

The study thus contributes to the growing literature showing that languages can differ in their F0 profile, but highlights the fact that the choice of speech materials to compare can be critical.

ACKNOWLEDGMENTS

This research was supported by NSF Grant No. BCS-0720304 to the first author. A very preliminary version was presented at the Spring 2009 meeting of the Acoustical Society of America in Portland. We thank Yen-Liang Shue for all his help with VoiceSauce, undergraduate research assistants Ting Fan, Spencer Lin, Larina Luu, and Caitlin Smith for help with recording and analysis, and two reviewers for helpful comments.

[1]See supplementary material at http://dx.doi.org/10.1121/1.4730893 for the text of the English version of the story "Little Red Riding Hood," and means for all measures for all corpora, separately by speaker language and sex.

Baken, R., and Orlikoff, R. (**2000**). *Clinical Measurement of Speech and Voice*, 2nd ed. (Singular Publishing, San Diego), pp. 145–223.

Blomgren, M., Chen, Y., Ng, M., and Gilbert, H. (**1998**). "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," J. Acoust. Soc. Am. **103**, 2649–2658.

Boersma, P. (**2001**). "Praat, a system for doing phonetics by computer," Glot Int. **5**, 341–345.

Chen, G. T. (**1972**). "A comparative study of pitch range of native speakers of Midwestern English and Mandarin Chinese: An acoustic study," doctoral dissertation, University of Wisconsin-Madison, Madison.

Chen, S. (**2005**). "The effects of tones on speaking fundamental frequency and intensity ranges in Mandarin and Min dialects," J. Acoust. Soc. Am. **117**, 3225–3230.

Crystal, D. (**1969**). *Prosodic Systems and Intonation in English* (Cambridge University Press, London), pp. 126–194.

Deutsch, D., Jinghong, L., Sheng, J., and Henthorn, T. (**2009**). "The pitch levels of female speech in two Chinese villages," J. Acoust. Soc. Am. **125**, 208–213.

Dolson, M. (**1994**). "The pitch of speech as a function of linguistic community," Music Percept. **11**, 321–331.

Eady, S. J. (**1982**). "Differences in the F0 patterns of speech: Tone language versus stress language," Lang Speech **25**, 29–42.

Fairbanks, G. (**1940**). *Voice and Articulation Drill Book* (Harper, New York), p. 168.

Fairbanks, G. (**1960**). *Voice and Articulation Drill Book*, 2nd ed. (Harper, New York), pp. 124–139.

Graddol, D., and Swann. J. (**1983**). "Speaking fundamental frequency: Some physical and social correlates," Lang Speech **26**, 351–366.

Hanley, T. D., and Snidecor, J. C. (**1967**). "Some acoustic similarities among languages," Phonetica **17**, 141–148.

Hanley, T. D., Snidecor, J. C., and Ringel, R. (**1966**). "Some acoustic differences among languages," Phonetica **14**, 97–107.

Harrington, J. (**2010**). *Phonetic Analysis of Speech Corpora* (John Wiley and Sons, New York), Chap. 2.

Henton, C. (**1989**). "Fact and fiction in the description of female and male pitch," Lang. Commun. **9**, 299–311.

Hollien, H., and Michel, J. F. (**1968**). "Vocal fry as a phonational register," J. Speech Hear. Res. **11**, 600–604.

Honorof, D. N., and Whalen, D. H. (**2005**). "Perception of pitch location within a speaker's F0 range," J. Acoust. Soc. Am. **117**, 2193–2200.

Huang, Y.-H., and Fon, J. (**2011**). "Investigating the effect of Min on dialectal variations of Mandarin tonal realization," in *Proc. 17th ICPhS Hong Kong*, pp. 918–921.

Kawahara, H., de Cheveign, A., and Patterson, R. D. (**1998**). "An instantaneous-frequency-based pitch extraction method for high quality speech transformation: Revised TEMPO in the STRAIGHT-suite," in *Proc. ICSLP'98*, Sydney, Australia (December 1998).

Lehiste, I. (**1970**). *Suprasegmentals* (MIT Press, Cambridge), pp. 68–105.

Liu, H.-M. (**2002**). "The acoustic-phonetic characteristics of infant-directed speech in Mandarin Chinese and their relation to infant speech perception in the first year of life," doctoral dissertation, University of Washington, pp. 109–113.

Loveday, L. (**1981**). "Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae," Lang. Speech **24**, 71–89.

Majewski, W., Hollien, H., and Zalewski, J. (**1972**). "Speaking fundamental frequency of Polish adult males," Phonetica **25**, 119–125.

Mang, E. (**2001**). "A cross-language comparison of preschool children's vocal fundamental frequency in speech and song production," Res. Studies Music Educ. **16**, 4–14.

Mennen, I., Schaeffler, F., and Docherty, G. (**2012**). "Cross-language differences in fundamental frequency range: A comparison of English and German," J. Acoust. Soc. Am. **131**, 2249–2260.

Ohara, Y. (**1992**). "Gender-dependent pitch levels: A comparative study in Japanese and English," in *Locating Power: Proceedings of the SECOND BERKELEY WOMEN and Language Conference*, edited by K. Hall, M. Bucholtz, and B. Moonwoman (Berkeley: Berkeley Women and Language Group, Berkeley), pp. 469–477.

Podesva, R. J. (**2007**). "Phonation type as a stylistic variable: The use of falsetto in constructing a persona," J. Sociolinguistics **11**, 478–504.

Reich, A. R., Frederickson, R. R., Mason, J. A., and Schlauch, R. S. (**1990**). "Methodological variables affecting phonational frequency range in adults," J. Speech Hear. Disord. **55**, 124–131.

Shue, Y.-L. (**2010**). "The voice source in speech production: Data, analysis and models," doctoral dissertation, UCLA, pp. 63–66.

Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (**2011**). "VoiceSauce: A program for voice analysis," in *Proc. 17th ICPhS Hong Kong*, pp. 1846–1849.

Todaka, Y. (**1993**). "A cross-language study of voice quality: bilingual Japanese and American speakers," doctoral dissertation, University of California, Los Angeles, pp. 145–147.

Torgerson, R. C. (**2005**). "A comparison of Beijing and Taiwan Mandarin tone register: An acoustic analysis of three native speech styles," master's thesis, Brigham Young University, pp. 73–82.

Yamazawa, H., and Hollien, H. (**1992**). "Speaking fundamental frequency patterns of Japanese women," Phonetica **49**, 128–140.

Xu, Y., and Xu, C. (**2005**). "Phonetic realization of focus in English declarative intonation," J. Phon. **33**, 159–197.

Xue, A., Hagstrom, F., and Hao, G. (**2002**). "Speaking F0 characteristics of bilingual Chinese-English speakers: A functional system approach," Asian Pacific J. Speech, Language Hearing **7**, 55–62.