Effect of Language on Voice Quality: An Acoustic Study of Bilingual Speakers of Mandarin Chinese and English

Shi Zhu\textsuperscript{a}  Sibie Chong\textsuperscript{b}  Yang Chen\textsuperscript{c}  Tianqi Wang\textsuperscript{b}  Manwa Lawrence Ng\textsuperscript{b}

\textsuperscript{a}Department of Applied Foreign Language Studies, Nanjing University, Nanjing, China; \textsuperscript{b}Speech Science Laboratory, University of Hong Kong, Hong Kong, China; \textsuperscript{c}Duquesne-China Health Institute, Duquesne University, Pittsburgh, PA, USA

Keywords
Long-time average spectra · Bilingual · Chinese · English · Acoustics

Abstract

Introduction: Previous studies reported discrepant vocal qualities associated with different languages. However, possible physical differences associated with speakers of different ethnicities were not accounted for. The present study attempted to examine the effect of language on one’s voice quality by eliminating the potential effects of physical differences associated with speakers of different languages.

Methods: Sixteen Chinese and fourteen Americans who were proficient in both Mandarin Chinese and English were recruited. They were instructed to read one Chinese and one English passage. Time-domain and long-term average spectral analyses were carried out, and speaking fundamental frequency (sF0), jitter, shimmer, and first spectral peak (FSP), mean spectral energy (MSE), and spectral tilt (ST) were measured using Praat. Results: Acoustic measures revealed no differences in sF0, FSP, and ST between Americans and Chinese. However, jitter, shimmer, and MSE values appeared to be affected by ethnicity (Chinese vs. Americans). Jitter and shimmer tended to be greater when someone was speaking his/her mother tongue. For language effect, Chinese was found to be associated with a faster rate of vocal fold vibration than English. MSE was higher for Chinese than English produced by Chinese, but not by American speakers, despite the similar ST in both languages. Discussion/Conclusion: Based on speech samples obtained from the balanced groups of bilingual speakers, the findings confirmed the presence of language effect on one’s voice quality. Laryngeal activity appeared to be affected by the language being spoken.

Introduction

In verbal communication, voice serves as a medium for transmission of linguistic and metalinguistic information. It is a product of human phonation, resulting from the intricate movement of phonatory and respiratory musculatures [1]. It has been documented that voice quality is correlated with the way vocal folds vibrate during sound production. Any change in glottal vibration tends to bring about a change in voice quality. A disordered voice due to some vocal pathology often exhibits a disrupted voice quality such as perceived breathiness, hoarseness, and strain (cf. [1, 2]). With some success, voice quality can objectively be described using acoustic...
parameters such as rate of and perturbation in vocal fold vibration, amount of noise in the signal, etc. [2].

In addition to vocal healthiness, other factors such as speaker’s gender, age, and ethnicity may also affect one’s voice quality (cf. [3–11]). Individuals of different ages, sexes, and ethnicities tend to be associated with different physiques and therefore may demonstrate different phonatory and resonance behaviors. The subtle physical and anatomical differences in the phonatory apparatus and the way it is used are believed to contribute to the discrepancy in vocal quality [6]. Research also suggested that language may have an effect on one’s voice [3, 5–8]. For example, in comparing Spanish and Catalan spoken by Catalan-Spanish and Spanish-Catalan bilingual speakers, Bruyninckx et al. [6] observed a greater between-language variation in vocal quality, compared to the variation within a language. Studies on Cantonese-English bilingual speakers also reported spectral [7] and pitch differences [8] between Cantonese and English spoken. These findings seemed to support the presence of a language effect on voice quality, suggesting that the acoustic differences were attributed by languages spoken. However, Altenberg and Ferrand [12] failed to find F0 difference between Cantonese and English in studying Cantonese-English bilingual adults. Similar findings were also reported by Xue et al. [10, 11] in studying English produced by Euro- and African-American male speakers. In view of these findings, it is still not clear if the language being used could influence the speaking F0, or more generally the voice quality.

That said, there appears to be a pitfall in these studies of language effect on voice. In studying how language might affect voice quality, the use of bilingual speakers of one ethnicity may not be sufficient and results so obtained can be biased, as the potential effect of ethnicity on voice quality is not controlled. With speakers from only one ethnicity, the anthropomorphic influence on voice quality was not accounted for, even when different languages were examined. For instance, studying vocal differences between Cantonese and English spoken by Chinese speakers in Altenberg and Ferrand’s study [12] might not reveal the real effect of language on voice due to the possible interaction between language and ethnicity. Also, in other studies of language effect on voice quality, speech materials were limited to just English; and some other studies relied only on very restricted groups of speakers (for example, developing children of ages 8–10 years in Morris [13] and prepubertal adolescents in Ng et al. [7, 8]; elderly speakers of ages 70–90 years in Xue and Fucci [14]). This greatly limited the generalizability of findings and some of which might be biased. The true effect of language on one’s voice can only be revealed by including a balanced group of proficient bilingual speakers of two ethnicities who were speaking two different languages. Such design could reveal the effect of language on voice bi-directionally, while the potential effects of ethnicity and physical differences associated with speakers of different languages on voice could be eliminated.

The present study adopted a balanced set of participants and examined the acoustical characteristics of Mandarin Chinese and American English produced by proficient Chinese (Asian) and white American bilingual adult speakers. By analyzing Chinese and English produced by Chinese-English (who were Asians) and English-Chinese (who were white Americans) bilingual speakers, and by balancing the potential effect of ethnicity on voice, the true effect of language on voice can be revealed.

Acoustic analysis has been used as a noninvasive means to objectively describe and quantify voice quality [4, 7, 12–16]. Traditional acoustical parameters included fundamental frequency (F0), perturbation of periodicity of vocal fold vibration such as jitter and shimmer, and signal-to-noise ratio. However, the use of these acoustical measures was not without its problems. Recall that voice is predominantly an indication of human phonatory behavior. However, according to the source-filter theory of speech production, the speech sounds that we hear every day are products of sound source (voice) generated by the phonatory system and filter (resonance) of the articulatory system [17]. The supralaryngeal resonance effect needs to be removed before one can faithfully reveal voice quality. This is made possible by using long-term average spectral (LTAS) analysis of speech [18, 19]. According to speech acoustics [17], a spectrum is rendered at an instant of time which is infinitesimal; it reflects the instantaneous frequency distribution of energy as the result of the source and filter of the speech production system. However, LTAS is obtained by overlapping all instantaneous spectra together over a sufficiently long duration. By averaging a number of speech spectra obtained over a long period of time, the effect of supralaryngeal resonance effect on voice source can be eliminated, rendering a depiction of purely phonatory behavior. This is based on the assumption that, in a sufficiently long production, the articulatory (resonance) effect can be neutralized. LTAS analysis of speech has been used previously in evaluating treatment outcomes [20], in quantifying alaryngeal speech [21] and voice quality [7]. In view of its use, LTAS analysis was used in the present study in conjunction with
Effect of Language on Voice Quality

time-domain acoustic analysis to examine the vocal characteristics associated with English and Mandarin Chinese spoken by both white American and Chinese bilingual speakers. The specific objectives of the study were
• to acoustically evaluate the voices associated with English and Mandarin Chinese spoken by white American and Chinese speakers who were proficient speakers of the two languages; and
• based on the acoustical parameters extracted, to determine if the voices associated with English and Mandarin Chinese were significantly different.

Methods

Participants
Two groups of bilingual speakers were recruited for the present study. Group one consisted of 16 (8 M and 8 F) Chinese-English bilingual speakers. They were ethnic Chinese (Asians) who were native speakers of Mandarin Chinese and highly proficient in English. Group two was formed by 14 (8 M and 6 F) English-Chinese bilinguals. They were white Americans and native speakers of American English who were speaking native-like Mandarin Chinese as an L2. All bilingual speakers were highly proficient in both Mandarin Chinese and English, as judged by two investigators (Y.C. and M.L.N.) both of whom were proficient in both Mandarin Chinese and American English. As there is no simple and standardized tool for assessing and comparing proficiency of spoken Mandarin Chinese and American English, subjective evaluation of language proficiency provided by the two investigators was used. All bilingual speakers were adults of ages between 20 and 34 years (mean = 27.11 years, SD = 4.84 years), and they were recruited from different universities in Mainland China and Hong Kong. Individuals were included in the study only when: (1) they were highly proficient bilingual speakers and (2) they had no reported history of problems of speech and voice.

Ethical approval for the study was obtained from the Human Research Ethics Committee for Non-Clinical Faculties (Ref: EA241107). Written informed consent was obtained from the bilingual speakers for participation in the study.

Instrumentation and Procedure

According to Löfqvist [18], in order for LTAS analysis to be valid, the speech samples need to be sufficiently long. As such, continuous Mandarin and English productions were obtained. Participants were instructed to read two short passages, a 180-character Chinese passage “The North Wind and the Sun” in Mandarin Chinese [22] and the 38-word first paragraph of the Rainbow Passage in English [23]. Both speech materials have been used extensively in the field of speech-language pathology, and the sounds in the speech materials should closely resemble the repertoire of sounds used in daily conversation in the respective language. The participants were instructed to produce the speech materials in a comfortable loudness and speech rate. Recording of speech samples was block randomized: a random half of the participants read the Chinese passage first, followed by the English passage; and the remaining participants took the reverse order. The speakers were provided some practice time prior to the experiment to be familiarized with the experiment. Speech samples were recorded using Praat via a professional microphone (SM-010 MV, SONICM Shengmei [HK] Industrial Limited, China) that was located about 10 cm from the speaker’s mouth. The samples were digitalized at a 44.1 kHz sampling rate with 16 bits/sample quantization using Praat.

Data Analysis

Time-domain and LTAS acoustic analyses were performed, from which three time-domain acoustical parameters including average speaking fundamental frequency (sF0), jitter, and shimmer, and three LTAS parameters including the first spectral peak (FSP), mean spectral energy (MSE), and spectral tilt (ST) were obtained from Chinese and English voice samples using Praat.

sF0, jitter, and shimmer have been used widely in describing and quantifying voice. sF0 measures the rate at which vocal folds vibrate (in Hz), and it reflects the perceived pitch. Jitter and shimmer indicate stability of vocal fold vibration during speech. Greater jitter and shimmer correlate with a greater variability in rate (F0) and excursion (amplitude) of vocal fold vibration, respectively. For the LTAS parameters, FSP measures the frequency (in Hz) of the first amplitude peak across the LTAS spectrum, and it should represent the average voice pitch of the production [24] and should be sensitive to vocal fold stiffness [25]. Expressed in negative dB, MSE indicates the average energy (amplitude) between 0 and 8 kHz relative to the maximum amplitude. This measure is believed to correlate with laryngeal tension; a greater MSE value indicates a greater glottal tension and a greater perceived strained quality [25]. ST represents the energy (amplitude) difference between low- and high-frequency ranges (thus called “tilt”). It is calculated by dividing the sum of amplitudes between 0 and 1 kHz and that between 1 and 5 kHz. A greater ST implies relatively less energy in the high frequency, indicating a more rapid decline of energy with frequency [18, 19]. According to Mendoza et al. [26], a smaller ST may reflect vocal hyperfunction.

Both time-domain and LTAS analyses can only be carried out from periodic segments. Therefore, all silent and unvoiced segments were removed from the speech samples before analysis. To do this, the speech waveform was first displayed with period markers superimposed. The unvoiced segments void of period markers were removed from the speech sample. With the edited speech samples, cycle-to-cycle F0, percent jitter (in %), and shimmer (in dB) values were calculated. It was followed by LTAS analysis using a bandwidth of 10 Hz was used. After the LTAS spectra were obtained, FSP, MSE, and ST values were calculated. All editing of waveforms and subsequent analyses were carried out using Praat.

Statistical Analysis

To reveal the acoustical differences between languages (Chinese vs. English) and ethnicities (Chinese vs. Americans), two-way repeated-measures analyses of variance were performed for sF0, jitter, shimmer, FSP, MSE, and ST values.

Reliability Measures

As human decision was involved in editing speech segments prior to LTAS, reliability of such a procedure was assessed. Relative percent errors were calculated to reflect inter- and intra-rater reliability in order to determine if editing of speech samples was done reliably and consistently. A collection of 20% of all speech samples

Effect of Language on Voice Quality

Folia Phoniatr Logop 2022;74:421–430
DOI: 10.1159/000525649
Table 1. Mean and standard deviation values of various acoustic parameters associated with Mandarin and English produced by American and Chinese speakers

|                      | American speakers |               | Chinese speakers |               |
|----------------------|-------------------|--------------|------------------|--------------|
|                      | Mandarin          | English      | Mandarin          | English      |
| F0, Hz               | 171.14 (60.06)    | 162.45 (60.99)| 171.45 (47.22)   | 172.56 (44.57)|
| Jitter, %            | 1.99 (0.54)       | 2.30 (0.74)  | 2.39 (0.59)      | 2.07 (0.40)  |
| Shimmer, dB          | 0.78 (0.14)       | 0.88 (0.13)  | 0.91 (0.10)      | 0.85 (0.09)  |
| FSP, Hz              | 187.86 (66.15)    | 166.43 (55.03)| 190.63 (51.38)   | 165.00 (60.66)|
| MSE, dB              | −33.79 (4.33)     | −33.04 (3.39)| −34.83 (4.45)    | −36.54 (4.97) |
| ST                   | 0.56 (0.11)       | 0.55 (0.12)  | 054 (0.12)       | 0.56 (0.17)  |

Fig. 1. Mean F0 values (in Hz) associated with American and Chinese speakers when speaking Mandarin and English.

Fig. 2. Mean jitter values (in %) associated with American and Chinese speakers when speaking Mandarin and English.
was randomly selected and reedited by the same researcher and a second researcher who was also experienced in using Praat for acoustical analysis. Intra-judge reliability was calculated by comparing the values obtained in the first and the second edits by the primary researcher. Inter-judge reliability was calculated by comparing the values obtained by the first and the second researcher. Results indicated the intra- and inter-judge relative percentage error values were 1.72% and 1.63%, respectively, indicating editing of speech samples was reliable and consistent.

**Results**

Descriptive data including average and standard deviation values of sF0, percent jitter, shimmer in dB obtained from time-domain analysis and those of FSP, MSE, and ST obtained from LTAS analysis associated with Mandarin and English produced by all bilingual speakers are depicted in Table 1. It should be noted that the values in Table 1 are averaged across male and female speakers. Considering the known difference in sF0 between men and women, the numeric data of sF0 and possibly FSP in Table 1 should be interpreted with caution. To facilitate comparison between languages and ethnicity, data on sF0, jitter, shimmer, FSP, MSE, and ST are represented in Figures 1–6.

**Time-Domain Analysis**

For sF0, jitter, and shimmer values, rmANOVA results consistently revealed a significant interaction effect between language and ethnicity (ps < 0.01). As such, dependent-samples *t* tests were subsequently performed, one for each measure. Significantly higher sF0 was found in Chinese than English when spoken by American speakers (*t*(13) = 3.667, *p* < 0.01), but not Chinese speakers (*p* > 0.05). Both jitter and shimmer associated were significantly greater for Chinese than English when spoken by Chinese speakers (jitter: *t*(15) = 4.516, *p* < 0.001; shimmer: *t*(15) = 2.420, *p* < 0.05). In addition, while no significant difference in jitter was found between English and Chinese produced by American speakers (*p* > 0.05), they exhibited greater shimmer in dB values when speaking English (*t*(13) = −4.266, *p* < 0.01), but not when speaking Chinese (*p* > 0.05).

Similarly, independent-samples *t*-tests were carried out to assess possible differences in sF0, jitter, and shimmer between American and Chinese speakers. Results revealed that, regardless of language, both sF0 and jitter were not significantly different between American and Chinese speakers (*p* > 0.05). However, while shimmer was not significantly different between Chinese and American speakers (*p* > 0.05) when speaking English, Chinese speakers exhibited a significantly higher shimmer than American speakers when speaking Mandarin Chinese (*t*(28) = −2.744, *p* < 0.05).

**LTAS Analysis**

For FSP, interaction effect between language and ethnicity (*p* > 0.05) and main effect for ethnicity (*p* > 0.05) were not significant. However, a significant main effect for language (*F*(1, 28) = 8.271, *p* < 0.01) was found. As
shown in Table 1, for both American and Chinese speakers, FSP associated with Mandarin Chinese was significantly greater than that with American English.

However, a significant interaction effect between language and ethnicity was found for MSE ($F(1, 28) = 7.222$, $p < 0.05$). Subsequent paired-samples $t$-tests revealed that MSE associated with Mandarin Chinese spoken by Chinese speakers was closer to zero (therefore greater in value) than English ($t(15) = 3.039$, $p < 0.01$), but not for American speakers ($p > 0.05$). A significantly greater MSE (less negative) was associated with American than Chinese speakers when speaking English ($t(28) = 2.222$, $p < 0.05$). Yet, no significant difference in MSE values between them when speaking Chinese ($p > 0.05$). With regard to ST, results indicated a lack of significant interaction effect between language and ethnicity ($p > 0.05$) and no main effect for both language and ethnicity ($ps > 0.05$).
Discussion

The present study examined American and Chinese speakers who were highly proficient in both Mandarin Chinese and English. Such balanced design was adopted in order to eliminate or control for the possible influences to voice quality that might originate from ethnicity. Acoustic analyses were performed on continuous speech samples produced in Chinese and English by speakers of two ethnicities: white Americans (Caucasians) and Chinese (Asians), which allowed us to understand how language and voice quality are related, with the possible effect of ethnicity being canceled out. As language proficiency has been reported to contribute to vocal differences in previous studies, especially for bilingual speakers (cf. [7, 8]), only bilingual speakers who spoke both Mandarin Chinese and English equally proficiently were recruited for the study. While it was easy to identify Chinese who spoke good English, recruiting American speakers who were proficient in Mandarin Chinese was quite difficult, possibly due to the lack of popularity of learning to speak Chinese among American speakers. This explains the small pool of American bilingual speakers.

The array of acoustical parameters included those extracted from time-domain acoustic analysis: (1) sF0, (2) jitter, and (3) shimmer, and LTAS analysis: (1) FSP, (2) MSE, and (3) ST.

Voice Quality Difference between Americans and Chinese

Acoustic findings revealed no significant differences in sF0 and FSP values between Americans and Chinese, regardless of language used (see Fig. 1, 4). Both sF0 and FSP should reflect the average rate of vocal fold vibration during continuous speech production. The present finding suggests that, when white Americans and Chinese were speaking Chinese or English, their vocal folds were vibrating with a similar stiffness and therefore a similar rate. This appears to contradict with those reported by Andrianopoulos et al. [3] in which Chinese were found to exhibit a higher F0 than Indians, Caucasians, and African Americans. However, it should be noted that, in Andrianopoulos et al.’s [3] study, participants of different ethnicities were physiologically matched with their age, height, and weight to be within 15%. By doing so, the morphological or physical factor due to different ethnicities was eliminated.

Findings revealed a lack of significant differences in jitter and shimmer between Americans and Chinese, with the exception of shimmer associated with Mandarin Chinese spoken by Chinese speakers. Jitter and shimmer are related to steadiness of vocal fold vibration. Greater jitter and shimmer values are associated with a less stable phonation with perturbations in rate and excursion of vocal fold vibration, respectively. Though not directly related, Chen [27] identified a difference in vocal steadi-

![Fig. 6. Mean ST values associated with American and Chinese speakers when speaking Mandarin and English.](image-url)
ness associated with Chinese and Americans when speaking English. Articulation appears to have an effect on steadiness of production. In fact, in their kinematic study, Fuchs and Perrier [28] observed a greater variation in producing posterior sounds when compared with anterior sounds. However, in the present context, jitter and shimmer only refer to phonatory steadiness during speech production.

When speaking English, American speakers exhibited a significantly higher MSE value (less negative) than Chinese speakers. A similar trend was observed when they were speaking Mandarin Chinese (see Fig. 5). According to previous studies (cf. [24, 25]), MSE generally indicates tension in the larynx during speech production. The higher MSE in American speakers implies that they spoke both English and Chinese with tenser vocal folds than Chinese speakers. This finding of different vocal physiology between Americans and Chinese could be combined with Xue, Hao, and Mayo [10], which suggested that the vocal differences between American and Chinese speakers were likely due to their different vocal tract volumes. Americans were likely to possess a larger vocal tract volume, offering a slightly different resonance than Chinese. With the source-filter theory stating that all speech sounds are products of the phonatory and articulatory (resonance) systems, the different articulatory systems between Americans and Chinese rendered a discrepancy in the speech output.

Regarding ST, American and Chinese speakers exhibited no significant differences in ST when speaking Chinese and English. ST is defined as the relative energy of lower and higher harmonics; a smaller ST resembles a hypofunctional voice [19]. As discussed in Goberman and Robb [24], ST is correlated with the perceived breathiness and the level of aspiration in the voice. It follows that our American and Chinese speakers demonstrated similar level of strain in the laryngeal system when speaking Mandarin Chinese and English.

Language Effect on Voice Quality

Findings on sF0 and FSP generally suggest that Mandarin was associated with a faster rate of vocal fold vibration when compared to English, with the exception of sF0 in Chinese speakers. This somehow demonstrates a language effect on voice, in which voice pitch to some extent was found to vary with the language being spoken. This finding reconfirms the language effect on speaking F0 reported by Ng et al. [7] which compared F0 of Cantonese and English spoken by bilingual speakers. The finding of higher F0 in Chinese echoes to what was reported by Andrianopoulos et al. [3], despite the use of vowels instead of continuous speech in their study. Based on the above, it appears that both Cantonese and Mandarin Chinese are likely to exhibit a higher F0 than English. Vocal folds needed to vibrate at a faster rate when speaking Cantonese and Mandarin Chinese when compared to English. This may be related to tonality of the language; both Cantonese and Mandarin, as different dialects of Chinese, are tonal in nature. In a tone language, the same phonetic segment can carry different meanings when produced at different tones. In Mandarin, there are four lexical tones: high-level, rising, falling-rising, and high-falling tones, whereas in Cantonese, the six lexical tones are high-level, high-rising, mid-level, low-level, low-falling, and low-rising tones [29]. However, Altenberg and Ferrand [12] reported comparable average F0 between English and Cantonese spoken by English-Chinese bilingual speakers. Such discrepancy might be attributed to the fact that only female speakers were recruited for their study.

Perturbation measures showed an interesting pattern. For Chinese speakers, jitter and shimmer were greater when speaking Mandarin Chinese, but not English. The reverse was true for American speakers for shimmer. There was a tendency that American speakers had a greater jitter when speaking English, but not Mandarin. It follows that the speakers somehow showed greater variation in frequency and amplitude of glottal vibration when speaking their mother tongues. This shows a concrete language effect on voice, at least regarding steadiness of voice production.

It has been suggested that a higher MSE value is correlated with an increased laryngeal tension during speech production [25]. In the present study, despite the lack of MSE between Chinese and English spoken by American speakers, it was significantly higher for Mandarin Chinese than English produced by Chinese speakers. In a study of accent, Kerr [30] reported that speakers of an Asian language such as Vietnamese, Cantonese, and Korean tended to produce sounds with greater posterior resonance. As glottal and pharyngeal musculatures are anatomically located toward the back of vocal tract, it is not sure if the perceived posterior resonance is correlated with Mandarin Chinese as the language being spoken, which may seem to correlate with the greater laryngeal tension seen in the present study. Apparently, further study examining relevant muscular activity is needed to confirm this.

ST value obtained from LTAS tends to indicate the breathiness of voice [31]. A lower ST indicates a greater
amount of breathiness and aspiration noise in the voice [26]. The lack of significant difference in ST between Chinese and English observed in the present study suggests that language does not affect perceived breathiness of voice.

**Limitations of the Study**

Several limitations could be identified. First, as it was not easy to recruit Americans who can speak Mandarin Chinese proficiently, the study suffered a small pool of bilingual Americans who were proficient in both Chinese and English. This could have limited the generalizability of results from the study. Future study needs to involve an extended pool of bilingual participants. In addition, bilingual speakers of other languages should also be studied in the future, in order to determine in what specific ways does a language affect one’s voice quality.

**Conclusion**

The present study investigated the language effect on voice quality using an array of acoustical parameters, based on balanced groups of American and Chinese bilingual speakers who were proficient in both Chinese and English. The present data confirmed the presence of a language effect on voice quality. Acoustical measures revealed no differences in sF0, FSP, and ST between Americans and Chinese. Jitter, shimmer, and MSE, however, appeared to be different between American and Chinese bilingual speakers. Jitter and shimmer tended to be greater when someone was speaking his/her mother tongue. For language effect, Mandarin was found to be associated with a faster rate of vocal fold vibration than English. MSE was higher for Mandarin Chinese than English produced by Chinese, but not American speakers, despite that ST was similar for both languages.

Results provide supporting evidence that language should be taken into consideration when it comes to clinical practice. For clinical purposes, clinical norms should be developed for different languages.

**Statement of Ethics**

The study was approved by the Human Research Ethics Committee for Non-Clinical Faculties (HRECNF) of the University of Hong Kong (Ref: EA241107). Written informed consent was obtained from the bilingual speakers for participation in the study.

**Conflict of Interest Statement**

This is to declare that the corresponding author of the article is currently an Associate Editor of Folia Phoniatrica et Logopaedica. All authors have no conflicts of interest to declare.

**Funding Sources**

The research was supported by the Education Faculty Research Fund, Faculty of Education, University of Hong Kong (21400035).

**Author Contributions**

Dr. Shu Zhu was responsible for literature review, recruiting Chinese bilingual participants, data collection, preliminary data analysis, and writing of part of the manuscript. Ms. Sibie Chong was responsible for literature review, acoustic measurements, data and statistical analysis, and preparation of the manuscript. Dr. Yang Chen was responsible for recruiting American bilingual participants, data collection, and preliminary data analysis. Mr. Tianqi Wang was responsible for data analysis and reliability measurements. Dr. Manwa L. Ng oversaw the entire project, from the concept and design of the study, literature review, data analysis, and manuscript preparation.

**Data Availability Statement**

All data generated and analyzed during this study are included in the article. Further inquires can be directed to the corresponding author.

**References**

1. Zemlin WR. Speech and hearing science: anatomy and physiology. 4th ed. Needham Heights: Allyn & Bacon: 1998.
2. Colton RH, Casper JK, Leonard R. Understanding voice problems: A physiological perspective for diagnosis and treatment. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2006.
3. Andrianopoulos AV, Darrow KN, Chen J. Multimodal standardization of voice among four multicultural populations: fundamental frequency and spectral characteristics. J Voice. 2001;15(2):194–219.
4. Awan SN, Mueller PB. Speaking fundamental frequency characteristics of white, African American, and Hispanic kindergartners. J Speech Hear Res. 1996;39(3):573–7.
5. Bahmanbiglu SA, Mojiri F, Abnavi F. The impact of language on voice: aAn LTAS study. J Voice. 2017;31(2):249.e9–e12.
6. Bruynincks M, Harmegnies B, Listerri J, Poch-Oiév D. Language-induced voice quality variability in bilinguals. J Phonetics. 1994; 22(1):19–31.
7 Ng ML, Chen Y, Chan EYK. Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers: a long-term average spectral analysis. J Voice. 2012;26(4):e171–6.
8 Ng ML, Hsueh G, Sam Leung CS. Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children. Int J Speech Lang Pathol. 2010;12(3):230–6.
9 Walton JH, Orlikoff RF. Speaker race identification from acoustic cues in the vocal signal. J Speech Hear Res. 1994;37(4):738–45.
10 Xue SA, Hao GJP, Mayo R. Volumetric measurements of vocal tracts for male speakers from different races. Clin Linguist Phon. 2006;20(9):691–702.
11 Xue SA, Neeley R, Hagstrom F, Hao J. Speaking F0 characteristics of elderly Euro-American and African-American speakers: building a clinical comparative platform. Clin Linguist Phon. 2001;15(3):245–52.
12 Altenberg EP, Ferrand CT. Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. J Voice. 2006;20(1):89–96.
13 Morris RJ. Speaking fundamental frequency characteristics of 8- through 10-year-old white- and African-American boys. J Commun Disord. 1997;30(2):101–6.
14 Xue SA, Fucci D. Effects of race and sex on acoustic features of voice analysis. Percept Mot Skills. 2000;91:951–8.
15 Brockmann M, Drinnan MJ, Storck C, Carding PN. Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. J Voice. 2011;25(1):44–53.
16 Orlikoff RF. Vocal stability and vocal tract configuration: an acoustic and electroglottographic investigation. J Voice. 1995;9(2):173–81.
17 Pickett JM. The acoustics of speech communication: fundamentals, speech perception theory, and technology. Boston: Allyn and Bacon; 1999.
18 Löfqvist A. The long-time-average spectrum as a tool in voice research. J Phon. 1986;14:471–5.
19 Löfqvist A, Mandersson B. Long-time average spectrum of speech and voice analysis. Folia Phoniatr Logop. 1987;39(5):221–9.
20 Tanner K, Roy N, Ash A, Buder EH. Spectral moments of the long-term average spectrum: sensitive indices of voice change after therapy? J Voice. 2005;19(2):211–22.
21 Ng ML, Liu H, Zhao Q, Lam PK. Long-term average spectral characteristics of Cantonese laryngeal speech. Auris Nasus Larynx. 2009;36(3):371–7.
22 Lee WS, Zee E. Standard Chinese (Beijing). J Int Phonetic Assoc. 2003;33(1):109–12.
23 Fairbanks G. Voice and articulation drill-book. 2nd ed. New York: Harper & Row; 1960. p. 124–39.
24 Goberman AM, Robb MP. Acoustic examination of preterm and full-term infant cries: the long-time average spectrum. J Speech Lang Hear Res. 1999;42(4):850–61.
25 Fuller BF, Horii Y. Spectral energy distribution in four types of infant vocalizations. J Commun Disord. 1988;21(3):251–61.
26 Mendoza E, Valencia N, Muñoz J, Trujillo H. Differences in voice quality between men and women: use of the long-term average spectrum (LTAS). J Voice. 1996;10(1):59–66.
27 Chen Y. Estimation of vocal tract steadiness in speakers of Mandarin-accented English. J Int Phonetic Assoc. 2009;39(2):189–204.
28 Fuchs S, Perrier P. On the complex nature of speech kinematics. Zaspiil. 2005;42:137–65.
29 Zee E. Chinese (Hong Kong Cantonese). In: International Phonetic Association, editor. Handbook of the International Phonetic Association. Cambridge: Cambridge University Press; 1999.
30 Kerr J. Articulatory setting and voice production: issues in accent modification. Prospect. 2000;15(2):4–14.
31 Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. J Speech Hear Res. 1996;39:311–21.