Effects of fundamental frequency changes on spoken sound loudness

Jonathan P. Evans*

Institute of Linguistics, Academia Sinica (Taipei, Taiwan)

Kueihong Lin

Institute of Linguistics, Academia Sinica (Taipei, Taiwan)

Alexander N. Savostyanov

Russian Academy of Medical Sciences

Abstract

This study aimed to investigate the perception of loudness in response to changes in fundamental frequency (F0) in spoken sounds, as well as the influence of linguistic background on this perceptual process. The results revealed that participants consistently perceived changes in F0 have accompanying changes in loudness, with a notable trend of lower F0 sounds being perceived as louder than higher F0 sounds. This finding contrasts with previous studies on pure tones, where increases in frequency typically led to increases in loudness. Furthermore, the study examined differences between two distinct groups of participants: Chinese-speaking and English-speaking individuals. It was observed that English-speaking participants exhibited a greater sensitivity to minor intensity changes compared to Chinese-speaking participants. This discrepancy in sensitivity suggests that linguistic background may play a significant role in shaping the perception of loudness in spoken sound. The study's findings contribute to our understanding of how F0 variations are perceived in terms of loudness, and highlight the potential impact of language experience on this perceptual process.

Keywords

loudness, linguistic background, spoken sound, pitch, tone language, stress, intensity

Corresponding author: Jonathan P. Evans, jonathan@sinica.edu.tw

Introduction

Loudness is a perception of intensity that can only be evaluated psychometrically. It has been claimed that “loudness is related to the total neural activity evoked by a sound” and that it "may depend on a summation of neural activity across different frequency channels (critical bands)” (Moore, 2013).

For pure-tone sounds (sine waves), the wave amplitude correlates heavily with loudness. Nevertheless, with amplitude held constant, pure tones differing in frequency also differ in loudness. The most commonly referenced set of equal loudness contours (ELCs) is ISO 226:2003 (Suzuki & Takeshima, 2004). These contours show that at a conversational loudness level of 60 phons, between approximately 600 Hz and 4 kHz, the human ear is especially sensitive to frequency information (cf. Fletcher & Munson, 1933; Gramming et al., 1988; and the studies referenced therein). This sensitivity amplifies the loudness levels of formants, fricative noise, and other speech information, thereby facilitating comprehension.

ELCs have contributed to the development of loudness meters via A-weighting (IEC 61672:2003). In estimating the loudness of noise, the A-weighted curve is based on the 40-phon ELC taken from Suzuki & Takeshima (2004). The ITU-R 468-weighting curve estimates the loudness of noise in audio systems and environmental noise (Recommendation ITU-R). Another standard, ISO 21727:2016, estimates the level of “subjective loudness and annoyance” during movie showings.

Despite extensive research on loudness, factors contributing to spoken sound loudness remain largely unexplored. This study had three main aims concerning the loudness of spoken sounds: determine whether (i) spoken sounds that differ only in F0 also differ in loudness; (ii) F0 changes of different sizes and directions evoke different loudness properties; and (iii) there is a relationship between F0 and loudness affected by language experience.

In this study potential loudness differences between spoken sounds were investigated. Specifically, the extent to which loudness judgments depend on language experience, the sex of the speaker, the sex of the listener, the difference in F0, and whether a sound was presented first or second were examined. In addition to sounds of equal intensity, study participants were also presented sounds with a 2 dB sound pressure level (SPL) difference in intensity.

To examine the effect of F0 differences between spoken sounds, the perception of loudness was analyzed among a sample of native speakers of either Taiwanese Mandarin Chinese or English. It was predicted that differences in F0 would induce differences in loudness. If general spoken language ability was a factor, it was expected that participants from the two populations would show similar responses. If experience in a particular language was a factor, it was hypothesized that the speakers of the tone language (Chinese) would demonstrate different loudness responses from the speakers of the nontonal language (English). It was unknown whether the sex of the speaker or the participant would have an effect, as the existing research does not support a particular hypothesis. Additionally, different studies suggest different loudness functions when spoken F0 steps up or down; thus, it was unknown whether either sequence might elicit an increase or decrease in loudness.

The experiments were performed on two sets of participants with two different language backgrounds but were otherwise designed to be identical. As this was an exploratory study, the data were analyzed and modeled as separate experiments to facilitate language-specific model discovery. Furthermore, separating the experiments avoided the problem of setting one language's data as the intercept.

Methods

Participants

For the study on native Mandarin Chinese speakers, a total of 34 healthy adults (17 male and 15 female), ranging in age from 20 to 25 years old (M_{age} = 21), were recruited, of which 32 completed the study. Chinese-speaking participants were native speakers of Taiwanese Mandarin and students at Taiwan Central University. English-speaking participants were recruited at Academia Sinica (Taipei, Taiwan). For the study on native English speakers, a total of 39 healthy adults, ranging in age from 23 to 48 years old (M_{age} = 31.6), were recruited, of which 37 (20 male, 17 female) completed the study. All were native or near-native English speakers. Those with extensive knowledge of a tone language (e.g., Chinese) were excluded.

All participants were subjected to standard audiometric testing and found to have normal hearing. The research methods used conformed to the requirements of the Academia Sinica Institutional Review Board (AS-IRB-BM-18048), which approved this study. The participants were advised of their rights, signed consent forms, and received compensation for their study participation.

Sound Recording and Analyses

Recordings were made of young adult male and female speakers producing the sound [ɑ]. The experiment was limited to [ɑ] since vowels have inherent loudness effects correlating with articulatory height (Fletcher, 1972, pp. 82–86). Recordings were selected for processing based on quality. 500 ms selections were isolated since a pilot study found that participants were unable to make loudness comparisons with sounds of shorter duration. Using the speech synthesis tools in Praat (Boersma & Weenink, 2019), amplitude and F0 were fixed at steady levels. Next, 50 ms transitions were added at the beginning and end of the stimuli, and the F0 levels were adjusted to equal either the mean for male/female speakers or ±1 or ±2 SD in semitones (mean and SD values summarized from 13 studies [cf. Simpson, 2009]). The resulting F0 values were as follows: male (86, 101, 119, 140, 164 Hz; step size of 2.8 semitones) and female (151, 177, 207, 240, 283 Hz; step size of 2.7 semitones).

The formant values used in the study fell within the F1 × F2 formant clouds given by Honorof and Whalen (2005) and Fletcher (1972) for American English [ɑ] and Mandarin [ɑ], respectively. F0 manipulation caused slight perturbation of the formants, with an SD <10 Hz for the male stimuli and an SD <30 Hz for the female stimuli (Supplemental Table 1). Across the range of F0 settings, the synthesized vowels were perceptible as [ɑ], suggesting that higher spectrum components were adequately unchanged. This factor was necessary for isolating F0 as an intrinsic variable contributing to any loudness changes. In addition, visual inspection of the spectrograms generated by Praat (Supplemental Figure 1) confirmed that the sound energy distribution was basically unchanged.

For the experiment, the sound files were organized into male–male and female–female sets, yielding 25 pairs for each. A pair of sounds [ɑ.ɑ] (e.g., 86, 101 Hz) was treated as a different stimulus from the same sounds in reverse order (e.g., 101, 86 Hz).

For the presentation, a standard touchscreen Windows computer was paired with Sennheiser HD 280 headphones; sound levels were calibrated to 75 dB SPL ±1 dB. Chinese-speaking participants were tested in a sound-insulated room at the Institute of Cognitive Neuroscience, National Central University in Taoyuan, Taiwan. English-speaking participants were tested in a sound-proof room at the Phonetics Laboratory, Institute of Linguistics, Academia Sinica. During the experiment, the participants were allowed periodic breaks of self-directed duration. The data collection took approximately one hour per participant.

Using a Praat multiple-forced-choice script, the sounds were presented as female or male sound pairs within blocks of same-sex stimuli, randomized across equal- (75 dB), falling- (76 dB, 74 dB), and rising-intensity (74 dB, 76 dB) pairs. The pairs were separated by a 500 ms silent interval. The 2 dB difference was chosen because at sound levels exceeding 40 dB and at frequencies above 100 Hz the just noticeable difference is less than 1 dB (Long, 2014). Thus, two-thirds of the stimuli pairs presented to the participants had detectable intensity differences.

The sequence [ɑ.ɑ] was patterned after two-syllable words, which allowed the participants to perceive a contrast similar to phonological tone and/or stress. After each sound pair was played, the participants were presented with a set of choices on the touchscreen and were instructed to select whether the first or second sound was "definitely louder," "probably louder," or "possibly louder" in their native language (labels patterned after those used in Baines et al., 2013, Lyn-Cook et al., 2007, Sofianou et al., 2013).

After a participant gave their response, there was a one-second pause before the next stimulus pair was presented. Each pair was presented 4 times, yielding 600 total stimuli per session. The responses were scored on a symmetrical 6-point Likert scale comprising three negative values (first sound was louder) and three positive values (second sound was louder).

When participants are presented with similar or identical stimuli, a time-order error (TOE) effect has been observed (Hellström, 1985). In the auditory domain, TOEs in loudness judgments can be induced by the relative loudness of the sound that was presented before the pair under evaluation (Lockhead, 1992). A balanced experimental design is necessary to reduce such errors. This study controlled for TOEs through randomization of stimuli pairs, and by setting the identical sounds condition as the intercept.

The only difference between the two study groups was that native English participants were exposed to two repetitions of each sound pair rather than four. This decision was made due to some participants requiring much more time to evaluate the sound pairs, and a desire to limit the duration of each sampling to within an hour.

Data Analyses

Use of the 6-point Likert scale did not show a clear pattern in the participant responses, perhaps due to difficulties with rating confidence in judgments. The responses were re-coded with the binary values zero (*First-sound-louder*) and one (*Second-sound-louder*). This simplification of participant responses allowed for a clearer interpretation of the results.

The F0 levels were chosen based on the mean and SD for F0 in male and female conversation. To aid in comparison across the two sets of stimuli, distances between F0 levels were converted from Hz to z-scale (Rose, 1987). Thus, the male voice pair (119 Hz, 101 Hz) would be expressed as (μ, -z), yielding the difference of two sounds in each pair in z-scale (*Zdiff*); because *Zdiff* is calculated by taking the first F0 height minus the second, this pair would have a *Zdiff* value of 1.

Analytical Methods

In the data analysis, a logistic mixed-effects model was used. Fixed effects of interest were as follows: the F0 difference between sounds, measured in z-scale (*Zdiff*); the sex of the speaker (*Sound*); and the sex of the participant (*Gender*). Participant (*ID*) was chosen as a random effect. It was unknown whether *Zdiff* would contribute to the response in a linear fashion, so it was treated as categorical. The dependent variable (*Response*) had two possible outcomes, *First-sound-louder* or *Second-sound-*

louder. Therefore, a logistic regression was chosen. Likelihood ratio testing was used to remove minimal relevance effects from the models.

Results

The first data analysis determined whether participants were sensitive to actual intensity differences. A generalized linear model was performed in R (R Core Team, 2020) with *Intensity* as the fixed effect, *ID* as the random effect, and *Response* as the dependent variable (optimized by the BOBYQA algorithm). The independent variable *Intensity* had three levels, *Same* (75 dB, 75 dB), *High–Low* (76 dB, 74 dB), and *Low–High* (74 dB, 76 dB). Both Chinese- and English-speaking participants interpreted actual intensity differences as loudness differences with nearly equally sized, opposite-signed coefficients, which were both statistically significant (Table 1).

Table 1. Effect of intensity differences on loudness judgments.

Chinese-speaking participants:

Predictor	*Second-sound-louder*			
	B	Std. Error	z-value	p-value
Intensity *Same* (75 dB)	0.0497	0.0628	0.792	0.428
Intensity *High–Low* (76 dB, 74 dB)	-0.8935	0.0374	-23.883	<2e-16
Intensity *Low–High* (74 dB, 76 dB)	0.8240	0.0375	21.981	<2e-16

English-speaking participants:

Predictor	*Second-sound-louder*			
	B	Std. Error	z-value	p-value
Intensity *Same* (75 dB)	0.0987	0.0673	1.467	0.142
Intensity *High–Low* (76 dB, 74 dB)	-2.2793	0.0639	-35.681	<2e-16
Intensity *Low–High* (74 dB, 76 dB)	2.3542	0.0689	34.178	<2e-16

The remaining analyses were performed on the same intensity data. The responses were checked for TOE. Across the entire set of same-intensity stimuli, Chinese-speaking participants in the first study identified the first sound as louder 48.8% of the time (SD = 9.4%). English-speaking participants were identified the first sound as louder in 47.6% of trials (SD = 11.3%). Thus, the effect of TOE was determined to be negligible.

The relative loudness of individual sounds was then summarized (Table 2). To evaluate significance, a generalized linear mixed model (GLMM) test with *ID* as the random effect was performed on the loudness of sounds in the first position. As the acoustic intensity levels were the same for all stimuli, the first and second sounds were always equally intense. Thus, the intercept was set to p = 0.5.

Table 2. Number of times that each sound was judged louder than a contrasting sound, arranged by position.

Chinese-speaking participants:

	Louder when in 1st position	% louder responses	Louder when in 2nd position	% louder responses
-2z	623	60.8	691	67.5
-z	477	46.6	538	52.5
m	440	43.0	469	45.8
z	431	42.1	465	45.4
2z	438	42.8	548	53.5

English-speaking participants:

-2z	328	55.4	388	65.5
-z	286	48.3	328	55.4
m	267	45.1	272	45.9
z	237	40.0	269	45.4
2z	257	43.4	328	55.4

Notes: n = 1024 and 592 for Chinese- and English-speaking participants, respectively.

Comparing the percentages of louder responses (Table 2) with the p-values (Table 3), it was found that when occurring in the first position, -2z was significantly louder than other sounds, while z and 2z were significantly softer. For Chinese-speaking participants, the mean was significantly softer, whereas for English-speaking participants the mean F0 approached significance for being softer.

Table 3. The relative loudness of sounds in the first position.

Chinese-speaking participants:

Predictor	B	Std. Error	z-value	p-value
First Pos.: –2z	–0.45099	0.08589	–5.251	1.51E-07
First Pos.: –z	0.14087	0.08484	1.661	0.096802
First Pos.: mean	0.29066	0.0852	3.412	0.000646
First Pos.: z	0.32749	0.08533	3.838	0.000124
First Pos.: 2z	0.29883	0.08522	3.506	0.000454

English-speaking participants:

Predictor	B	Std. Error	z-value	p-value
First Pos.: -2z	-0.2255	0.114	-1.977	0.048
First Pos.: -z	0.074	0.1137	0.651	0.5152
First Pos.: mean	0.2096	0.114	1.839	0.0659
First Pos.: z	0.4274	0.1149	3.719	0.0002
First Pos.: 2z	0.2815	0.1142	2.465	0.0137

Notes: Random effects of Chinese-speaking participants (variance = 0.1016, SD = 0.3188). Random effects of English-speaking participants (variance = 0.2151, SD = 0.4638).

To evaluate the loudness of sounds in the second position, the GLMM test was performed using the values from the right half of Table 2, with the intercept again set to 0.5. Comparing the percentages of louder responses (Table 2) with the p-values (Table 4), when occurring in the second position, –2z was significantly louder than the other sounds for Chinese-speaking participants, while *mean* and z were significantly softer. For English-speaking participants, –2z*,* –z*,* and 2z were significantly louder than the other sounds.

Table 4. The relative loudness of sounds in the second position*.*

Chinese-speaking participants:

Predictor	B	Std. Error	z-value	p-value
Second Pos.: –2z	0.74808	0.08817	8.485	<2e-16
Second Pos.: –z	0.10475	0.08503	1.232	0.218
Second Pos.: mean	–0.1721	0.08514	–2.021	0.0432
Second Pos.: z	–0.18825	0.08517	–2.21	0.0271
Second Pos.: 2z	0.14495	0.08509	1.703	0.0885

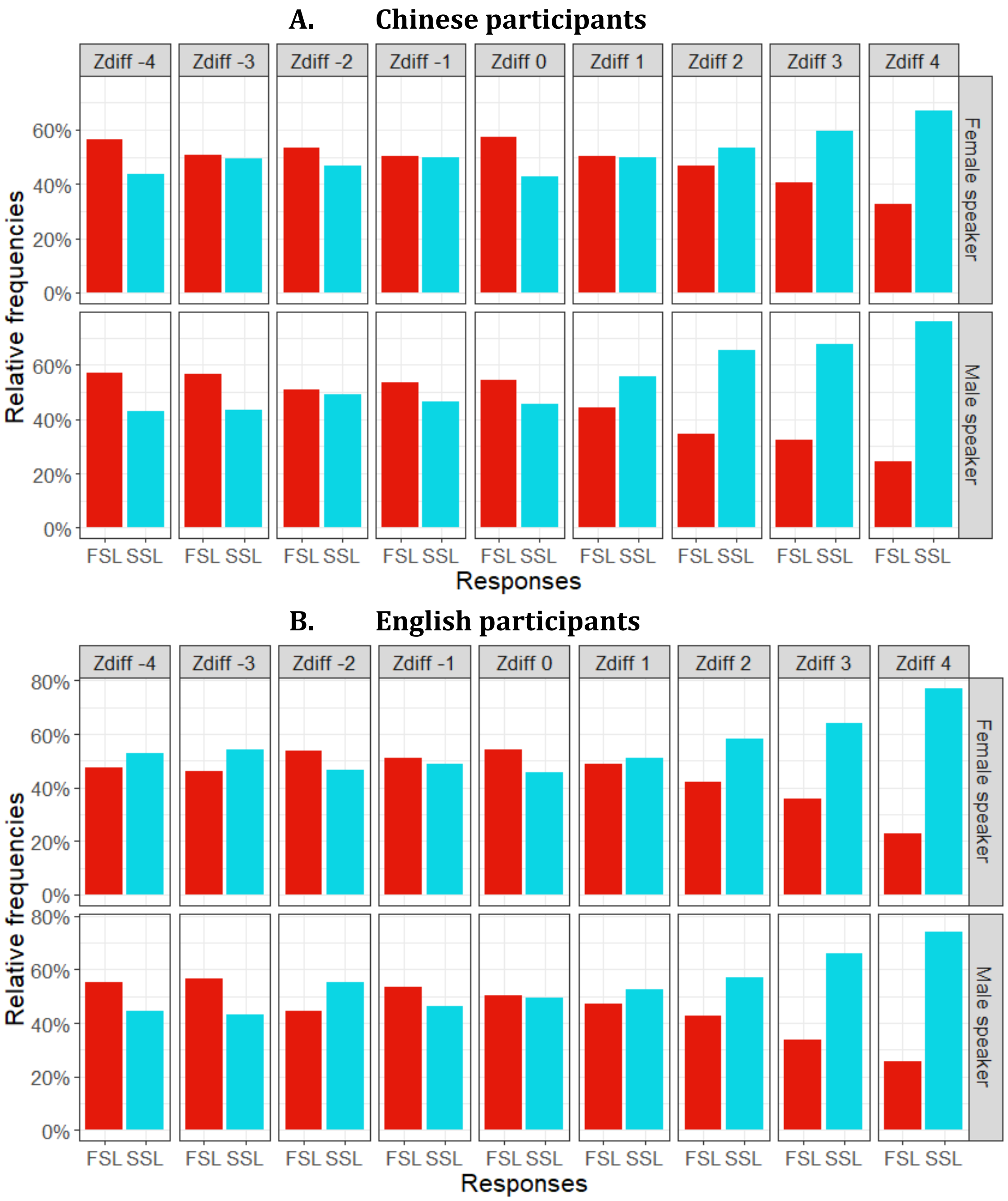
English-speaking participants:

Predictor	B	Std. Error	z-value	p-value
Second Pos.: -2z	0.6783	0.1176	5.768	8.03E-09
Second Pos.: -z	0.2315	0.1147	2.018	0.0436
Second Pos.: mean	-0.1682	0.1146	-1.468	0.1422
Second Pos.: z	-0.1898	0.1147	-1.655	0.098
Second Pos.: 2z	0.2315	0.1148	2.018	0.0436

Note: Random effects of Chinese-speaking participants (variance = 0.1029, SD = 0.3208). Random effects of English-speaking participants (variance = 0.2209, SD = 0.47).

This experiment also sought to answer whether shifts between F0 levels yielded consistent loudness differences. For the statistical modeling, the respondents' selection of the first sound as louder and *Zdiff* = 0 were taken together as the intercept of the binary response variable. The independent variables chosen for evaluation included the F0 difference between the two sounds in each pair, normalized to *Zdiff*; the integer values ranged from –4 to 4. The variable *Zdiff* was treated as categorical to avoid assumptions of linearity that would be imposed by a continuous variable. The other independent variables were *Gender* and *Sound*.

As shown in Figure 1, in contexts where the *Zdiff* values were above zero (the first sound has a higher F0 than the second sound), a shift toward more *Second-sound-louder* judgments was observed. In contexts where *Zdiff* values were negative, no obvious trend emerged.


A. Chinese participants

B. English participants

Figure 1. The percentage of trial counts in which the first sound (*red*) was selected as louder versus the second sound (*blue*), arranged by fundamental frequency difference in z-scale.

To evaluate the initial observations, the data were analyzed using GLMM in R with the lme4 package (Bates et al., 2015), and p-values were obtained using the lmerTest package (Kuznetsova et al., 2017). Here, *ID* was chosen as a random variable, and the independent fixed variables were *Zdiff* and *Sound*. *Gender* was found to have an insignificant effect on *Response* through likelihood ratio testing (p = 0.37). The random-effect structure of the final model for Chinese-speaking participants (Table 5) contained the random intercept for participants (SD^2 = 0.132, SD = 0.363).

Table 5. Results of generalized linear mixed model by fundamental frequency difference in z-scale (Chinese-speaking participants).

Predictor	*Second-sound-louder*			
	B	Std. Error	z-value	p-value
(Intercept = *Zdiff* 0)	−0.2975	0.1034	−2.876	0.004
Zdiff −4	0.0394	0.1983	0.199	0.8426
Zdiff −3	0.2665	0.1507	1.769	0.0769
Zdiff −2	0.1588	0.1318	1.205	0.2284
Zdiff −1	0.2907	0.1210	2.403	0.0163
Zdiff 1	0.2826	0.1210	2.336	0.0195
Zdiff 2	0.4279	0.1318	3.247	0.0012
Zdiff 3	0.6901	0.1525	4.524	0.0000
Zdiff 4	1.0370	0.2074	5.001	0.0000
Sound Male	0.1112	0.1143	0.972	0.3308
Zdiff −4:*Sound* Male	−0.1440	0.2805	−0.513	0.6077
Zdiff −3:*Sound* Male	−0.3547	0.2135	−1.661	0.0967
Zdiff −2:*Sound* Male	−0.0142	0.1861	−0.077	0.9390
Zdiff −1:*Sound* Male	−0.2485	0.1710	−1.453	0.1461
Zdiff 1:*Sound* Male	0.1478	0.1712	0.864	0.3878
Zdiff 2:*Sound* Male	0.4141	0.1889	2.193	0.0283
Zdiff 3:*Sound* Male	0.2540	0.2192	1.159	0.2466
Zdiff 4:*Sound* Male	0.3235	0.3049	1.061	0.2888

Notes: Zdiff: Difference in frequency between the two sounds, measured in z-scale. *Zdiff* > 0 indicates a drop in frequency. *Sound*: Sex of the speaker of the stimulus

For English-speakers (Table 6), *Sound* was found to have no effect on *Response* through likelihood ratio testing (p = 0.282); thus, the sex of the speaker was not included in the final model. The random-effect structure of the final model contained the random intercept for participants (SD^2 = 0.174, *SD* = 0.417). Note that while the English-speaking participant model included *Gender* but not *Sound*, the Chinese-speaking participant model included *Sound* but not *Gender*.

Table 6. Results of generalized linear mixed model by fundamental frequency difference in z-scale (English-speaking participants).

Predictor	*Second-sound-louder*			
	B	Std. Error	z-value	p-value
(Intercept = *Zdiff* 0)	0.0364	0.1500	0.243	0.8083
Zdiff -4	-0.2211	0.2720	-0.813	0.4163
Zdiff -3	-0.0368	0.2304	-0.160	0.8732
Zdiff -2	0.1888	0.1814	1.041	0.2979
Zdiff -1	0.1407	0.1611	0.873	0.3825
Zdiff 1	0.1544	0.1612	0.958	0.3380
Zdiff 2	0.2718	0.1820	1.494	0.1351
Zdiff 3	0.4196	0.2345	1.789	0.0736
Zdiff 4	0.8011	0.2892	2.771	0.0056
Gender Male	-0.2516	0.2043	-1.232	0.2181
Zdiff -4:*Gender* Male	0.4931	0.3697	1.334	0.1821
Zdiff -3:*Gender* Male	0.2914	0.3136	0.929	0.3528
Zdiff -2:*Gender* Male	-0.0907	0.2468	-0.368	0.7130
Zdiff -1:*Gender* Male	-0.2600	0.2200	-1.182	0.2373
Zdiff 1:*Gender* Male	0.0539	0.2193	0.246	0.8060
Zdiff 2:*Gender* Male	0.2805	0.2478	1.132	0.2576
Zdiff 3:*Gender* Male	0.5989	0.3246	1.845	0.0650
Zdiff 4:*Gender* Male	0.9337	0.4220	2.213	0.0269

Notes: Zdiff: Difference in frequency between the two sounds, measured in z-scale. *Zdiff* > 0 indicates a drop in frequency. *Gender:* Sex of the participant.

For Chinese-speaking participants, the effect of F0 difference predicted loudness judgments. The model specified *Zdiff* = 0 as the intercept, yielding the following results: β = –0.30, *SE* = 0.10, z = –2.88, and p < 0.004. The results demonstrated that as *Zdiff* increased from the intercept (zero) to four the participants gradually responded less frequently with *First-sound-louder* when they were presented with trials produced by the female speaker (Table 7). The coefficient was the highest in trials with a difference of *Zdiff* = 4 (β = 1.03, p < 0.000) and gradually dropped to *Zdiff* = 0 for female speakers. A similar, yet more prominent, trajectory occurred when stimuli were produced by male speakers, obtained through a calculation of the coefficients of *Zdiff* and those of their interactions with *Sound* = Male. When participants were presented with trials in which the second sound was higher in frequency than the first (*Zdiff* < 0), their responses were diverse, and no significant trend was detected by the model (p > 0.05).

Table 7. Percentage of *First-sound-louder* responses by sound pair for Chinese-speaking participants.

Female speakers:

First sound	Second sound				
	151 Hz -2z	177 Hz -z	207 Hz μ	240 Hz z	283 Hz 2z
151 Hz (-2z)	53.9 [0]	64.1 [-1]	57.8 [-2]	58.6 [-3]	56.3 [-4]
177 Hz (-z)	40.6 [1]	57.0 [0]	50.0 [-1]	57.0 [-2]	42.0 [-3]
207 Hz (μ)	44.5 [2]	57.8 [1]	58.6 [0]	46.1 [-1]	45.3 [-2]
240 Hz (z)	40.6 [3]	46.1 [2]	49.2 [1]	61.7 [0]	40.6 [-1]
283 Hz (2z)	32.8 [4]	40.6 [3]	50.0 [2]	53.9 [1]	54.7 [0]

Male speakers:

First sound	Second sound				
	86 Hz -2z	101 Hz -z	119 Hz μ	140 Hz z	164 Hz 2z
86 Hz (-2z)	47.7 [0]	64.1 [-1]	61.7 [-2]	67.2 [-3]	57.0 [-4]
101 Hz (-z)	25.0 [1]	59.4 [0]	57.8 [-1]	53.1 [-2]	46.1 [-3]
119 Hz (μ)	26.6 [2]	38.3 [1]	53.1 [0]	46.9 [-1]	38.3 [-2]
140 Hz (z)	25.8 [3]	29.7 [2]	59.4 [1]	56.3 [0]	45.3 [-1]
164 Hz (2z)	24.2 [4]	39.1 [3]	47.7 [2]	53.9 [1]	56.3 [0]

Note: [*Zdiff* value]

Table 7 indicates a statistically significant loudness effect that corresponds to *Zdiff* values ≥ -1. That is, changes in F0 evoked a loudness response in which the second sound with lower F0 was judged to be louder than the first sound. As seen in Figure 1, the effect monotonically increased from *Zdiff* = 0 to *Zdiff* = 4. However, when the first sound had a lower F0 than the second (negative *Zdiff* values), there was no significant correlation between *Zdiff* and loudness judgments.

For English-speaking participants, the main effect of the F0 difference predicted loudness judgments (Table 8). The model specified *Zdiff* = 0 and female participants as the intercept, yielding the following: $\beta = 0.04$, $SE = 0.15$, z = 0.24, p = 0.81. The results showed that responses of *First-sound-louder* from the female participants slowly decreased as the *Zdiff* increased in trials where the first sound was higher in frequency than the second. The coefficient was the highest in trials with *Zdiff* = 4 ($\beta = 0.80$, $p < 0.005$) and gradually dropped to *Zdiff* = 0 (Figure 2C); only *Zdiff* = 4 had $p < 0.05$. The coefficients from the male participants shared this trend, but the trajectory was sharper (Figure 2D). Their values were obtained through the addition of *Zdiff* and their interactions with *Gender*. When the participants were presented with trials in which the second sound was higher in frequency than the first (*Zdiff* < 0), no obvious pattern was indicated by the model (p > 0.05). The results indicated a statistically significant loudness effect corresponding to *Zdiff* = 4. That is, when the F0 stepped down from 2z to –2z, the second sound was deemed louder by both male and female participants.

Table 8. Percentage of *First-sound-louder* responses by sound pair for English-speaking participants.

Female speakers:

First sound	Second sound				
	151 Hz -2z	177 Hz -z	207 Hz μ	240 Hz z	283 Hz 2z
151 Hz (-2z)	51.4 [0]	59.5 [-1]	50.0 [-2]	51.4 [-3]	47.3 [-4]
177 Hz (-z)	39.2 [1]	55.4 [0]	54.1 [-1]	60.8 [-2]	40.5 [-3]
207 Hz (μ)	40.5 [2]	45.9 [1]	60.8 [0]	50.0 [-1]	50.0 [-2]
240 Hz (z)	33.8 [3]	36.5 [2]	50.0 [1]	51.4 [0]	40.5[-1]
283 Hz (2z)	23.0 [4]	37.8 [3]	48.6 [2]	60.8 [1]	52.7[0]

Male speakers:

First sound	Second sound				
	86 Hz -2z	101 Hz -z	119 Hz μ	140 Hz z	164 Hz 2z
86 Hz (-2z)	48.6 [0]	60.8 [-1]	62.2 [-2]	56.8 [-3]	55.4 [-4]
101 Hz (-z)	44.6 [1]	51.4 [0]	56.8 [-1]	47.3 [-2]	43.2 [-3]
119 Hz (μ)	35.1 [2]	44.6 [1]	56.8 [0]	51.4 [-1]	43.2 [-2]
140 Hz (z)	33.8 [3]	35.1 [2]	54.1 [1]	45.9 [0]	36.5[-1]
164 Hz (2z)	25.7 [4]	36.5 [3]	56.8 [2]	58.1 [1]	50.0 [0]

Note: [*Zdiff* value]

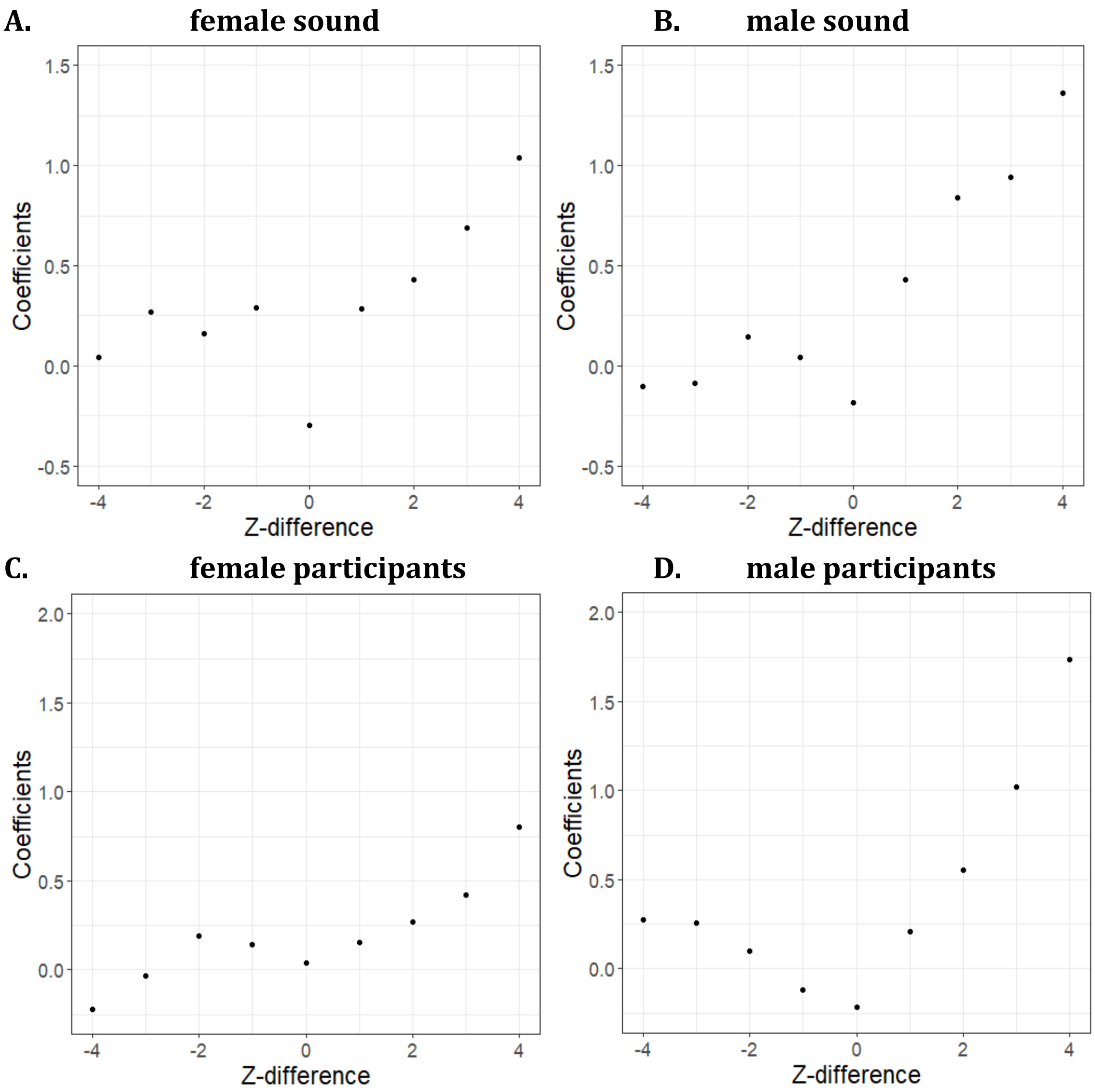


Figure 2. Logit values for predicting *Second-sound-louder*. A, B, Chinese-speaking participants. C, D, English-speaking participants. *Zdiff* > 0 indicates a drop in frequency. Coefficients > 0 indicate second sound was judged louder than the first.

Interim summary

Both Chinese and English speakers were sensitive to small intensity differences (2 dB) and judged the lowest F0 sounds to be louder. For Chinese speakers, downward F0 steps made the second sound seem louder. For English speakers only the largest downward step (4z) caused the second sound to be significantly louder.

Discussion

The first aim of the study was to investigate whether spoken sounds that differ only in F0 also differ in loudness. Unlike pure tone studies, the present study was not designed to find inherent loudness of fixed F0 levels (e.g., 100 Hz). Instead, the experiment was designed to model perception of spoken sounds in the immediate context of a preceding or subsequent syllable, and in the greater context of a speaker's conversational F0 range. Loudness properties of spoken sounds are summarized as in Table 9.

Table 9. Summary of statistically significant loudness distinctions by F0 level.

	Chinese Speakers		English Speakers	
	First Position	Second Position	First Position	Second Position
-2z	louder	louder	louder	louder
-z				louder
m	softer	softer	(softer)	
z	softer	softer	softer	
2z	softer	(louder)	softer	louder

Notes: (-2z, -z, m, z, 2z) correspond to F0 = 151, 177, 207, 240, 283 Hz (female voice) and 86, 101, 119, 140, 164 Hz (male voice). Parentheses indicates a p value that approaches significance. Calculated using data from Tables 3 and 4.

Table 9 shows that the lowest sound (-2z) was louder for both sets of participants, whether in initial or final position. For both sets of participants, the highest sound (2z) was softer in initial position and louder in final position (although not significantly louder for the Chinese-speaking participants). For Chinese-speaking participants, (m, z) were softer in both positions. For all other sound/speaker pairings, loudness of a particular F0 level was dependent on whether the sound was presented first or second in its pair.

Equal loudness contours (ELCs; e.g., Suzuki & Takeshima, 2004) show that in the range of F0 used in these experiments, higher frequency pure tones are louder than lower frequency comparanda. The difference between participant responses in the present study versus pure tone studies corroborates the hypothesis that perception of spoken loudness is subject to constraints which differ from those activated during pure tone perception. This part of the study verifies that differences in F0 lead to differences in loudness.

Crosslinguistically, F0 and intensity tend to decline continuously across an utterance (Ladd 1984). Thus, an utterance-final sound with F0 +2z above the mean is highly atypical. In studying the size-weight illusion, Yasegawa & Mikami (2015) found that when the prediction error increased to the point that the difference is much greater than expected, participants perceived a greater-than-actual weight difference. In normal speech conditions, +2z F0 sounds tend to be much more intense than normal F0 sounds. In our case, intensity was fixed, and the large difference between expected higher loudness and the actual loudness induced an enhanced perception. The fact that the increased loudness did not reach the level of significance for Chinese-speaking participants is discussed below.

Previous studies have found that when vocal sounds are produced with different F0 levels, intensity typically changes in the same direction (Scharine &

McBeath 2018, Gramming et al., 1988; Watson & Hughes, 2006). Thus, perception of speech may be enhanced by imputing positive correlation between F0 and loudness. This "expectation effect" has been found in studies of the size-weight illusion (Yasegawa & Mikami, 2015). In these studies, participants' life experience with correlations between size and weight leads them to estimate that a larger object is lighter than a smaller object of the same weight. In our experiment it seems that expectations of positive correlation between F0 and loudness lead participants to judge that the lowest F0 sound had greater loudness than other F0 levels of the same intensity.

In spoken language experience, it is typical to have higher F0 at the beginning of an utterance, with concomitant increase in intensity. Thus, +2z F0 at the initial position is analogous to -2z F0 in final position; it is in the predicted direction of change, although it is more extreme. The higher F0 sound is expected to be louder, so the expectation effect yields decreased loudness in this case. In final position, both F0 and loudness are expected to decline. The greater difference between stimulus and expectation seems to have caused a perceptual shift from assimilation to contrast (Yasegawa and Mikami, 2015). This paper found that when the prediction error is small, an individual's prior expectation aligns well with sensory information, leading to an assimilation effect. This bias causes the individual's perception to lean towards the prior expectation, reducing the disparity between the expectation and sensory data. Conversely, with an increasing prediction error, the prior expectation becomes less accurate. The individual then experiences a contrast effect, biasing perception away from the prior expectation and amplifying the difference between the prior expectation and sensory information. Their finding is consistent with +2z F0 being softer in initial position and louder in second position.

Other loudness effects seem to be affected more by particular language experience and are thus discussed below.

The second aim of the study was to investigate whether F0 changes of different sizes and directions evoke different loudness properties. F0 changes are part of linguistically meaningful signals; thus, it is important to understand whether shifts in F0 induce loudness changes. When the sound sequence comprised a lowering of F0 (*Zdiff* > 0), loudness coefficients increased monotonically for both groups of participants. Rising coefficients indicate increasing loudness for the second (lower) sound versus the first (higher) sound. Sound pairs with rising F0 (*Zdiff* < 0) did not elicit a pattern of responses. In previous studies, with intensity held constant, increases in frequency of pure tones led to increases in loudness (Neuhoff, McBeath & Wanzie 1999).

In the quest for a possible speech perception mechanism for a decrease in F0 corresponding to an increase of loudness, we note that each presentation of the stimuli pairs was followed by silence. In this way the temporal location of the second sound resembles prepausal position in spoken utterances. This position has been noted universally as a location for both F0 decline and a decrease in intensity (Arsikere, Shriberg, Ozertem 2015). If the second sound is perceived as prepausal, then the aforementioned expectation effect predicts that participants would evaluate the loudness of the second sound relative to an expected reduction of both F0 and loudness. Thus, intensity being held constant would be perceived as a loudness increase. The lack of a clear trend for increases in F0 (*Zdiff* < 0) remains an object for further study.

The third aim of the study was to evaluate the extent to which specific language experience affected loudness perception. Both the Chinese- and English-speaking participants were sensitive to a 2 dB SPL difference in intensity over the range of 86–283 Hz, which covers the typical conversational range of F0 for male and female speakers (Simpson, 2009). However, the English-speaking participants showed a stronger sensitivity to the intensity difference, as evidenced by their larger coefficients (Table 10). English has prominent stress effects, in which one syllable of a word receives more emphasis than the others, such as the first syllable of "*em*phasis." For example, English speakers have been shown to be more likely than Estonian speakers to interpret amplitude changes as stress (Lehiste & Fox, 1992). Likewise, Yu and Andruski (2010) found that English speakers rely on a variety of signals for detecting stress, while Chinese speakers primarily cue in to F0 changes (Li et al., 2021).

Table 10. Coefficients and z-values of generalized linear mixed models of participant sensitivity to 2 dB differences in spoken sounds, ranging from 86 to 283 Hz.

Participant	*High–Low* Coefficient	*High–Low* z-value	*Low–High* Coefficient	*Low–High* z-value
Chinese	-0.8935	-23.883	0.8240	21.981
English	-2.2793	-35.681	2.3542	34.178

Note: calculated from data in Table 1.

In the final +2z F0 condition, Chinese-speakers' judgment of loudness did not reach significance (p = 0.0885), unlike English-speaker judgments. Mandarin Chinese has sentence-final discourse markers that are pronounced with a high tone; e.g., [o^{55}], [a^{55}]. Although these syllables are somewhat infrequent, their occurrence in the grammar might have reduced Chinese-speakers' sense of increased loudness for final +2z F0 sounds.

Conclusion

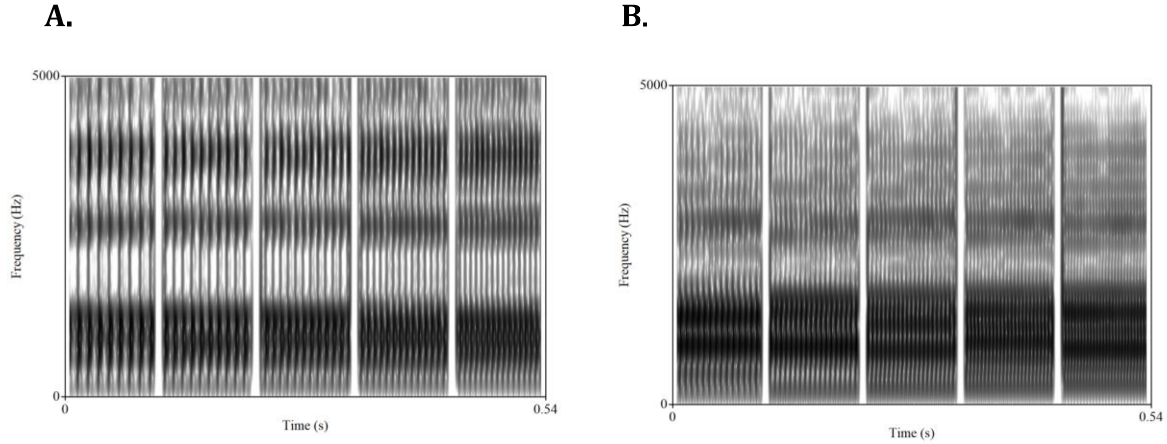
In this study, spoken sound loudness perception tests were performed with Chinese- and English-speaking participants, separated into two groups. The stimuli were based on the mean and ±1 and ±2 SD of conversational F0 levels. In sound pairs with minor differences in intensity (±2 dB SPL), English speakers showed more sensitivity to intensity changes than did Chinese speakers; cf. Table 10. This may have been due to the increased prominence of intensity in English stress patterns compared to acoustic markers of stress in Mandarin Chinese.

Tests were also performed on participant responses to spoken sounds of equal intensity. The loudness effects differed from those of previous pure tone models (e.g., Suzuki & Takeshima, 2004). In pure tone studies covering the present frequency range, lower-frequency sounds were softer than higher-frequency sounds. However, in the current study, the general trend was that low F0 sounds were louder than higher F0 sounds. In addition, downward shifts in F0 induced an effect of increasing loudness, but upward shifts did not. In spoken language, sounds with lower F0 tend to have concomitant reduced intensity. Also, in the prepausal position, there is a trend toward lowered F0 and lowered intensity.

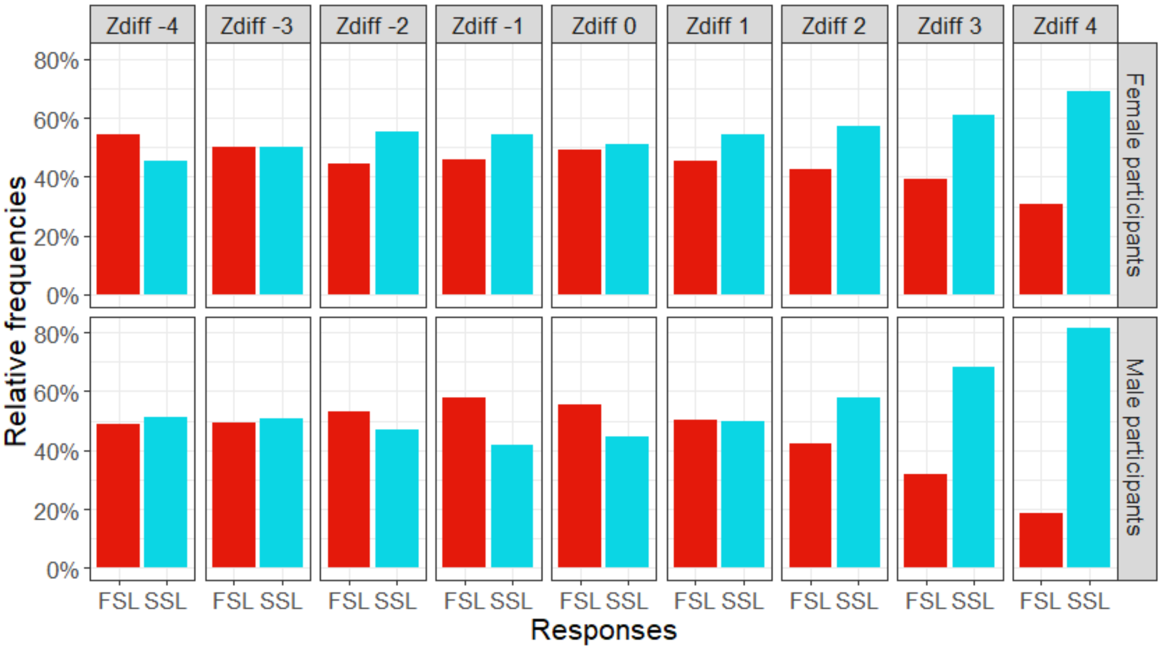
It appeared that the participants boosted the loudness of F0 sounds at the low end of the stimulus ranges and also detected loudness when F0 shifted downward, as a compensation for the typically decreased intensity of spoken sounds in those positions; cf. Table 9. The results from the English-speaking participants may reflect English stress properties. For example, higher F0 sounds in the initial position were judged to be softer, and lower sounds in the second position were judged to be louder. These loudness judgments may have reflected a response to the most common pattern of disyllabic stress, in which the first syllable is stressed (higher F0, greater intensity) and the second syllable is unstressed (lower F0, lower intensity). On the other hand, the loudness judgments of Chinese participants did not reach statistical significance, possibly due to the presence of sentence-final discourse markers in Mandarin Chinese, which may have influenced their perception of loudness in such contexts.

Overall, this study's findings suggested that some loudness perception is shared by speakers of diverse linguistic and cultural backgrounds (e.g., the lowest sound was judged as louder). However, the remainder of such perception could be affected by language experience, such as exposure to tones and/or stress.

Supplementary Figures and Tables



Supplemental Figure 1. Synthesized voice spectrograms. (A) The 100-ms spectrograms of the synthesized male voice for the vowel sound [ɑ] at the fundamental frequency values of 86, 101, 119, 140, and 164 Hz. (B) The 100-ms spectrograms of the synthesized female voice sound for [ɑ] at the fundamental frequency values of 151, 177, 207, 240, and 283 Hz.



Supplemental Figure 2. The percentage of trial counts in which the first sound (*red*) was selected as louder versus the second sound (*blue*), arranged by fundamental frequency difference in z-scale and by sex of the participant.

Supplemental Table 1. Mean formant values of the stimuli used in this study, with American English and Mandarin average values for reference.

Male	F1	F2	F3	F4
86 Hz	730	1146	2708	3749
101 Hz	734	1153	2714	3750
119 Hz	729	1151	2718	3748
140 Hz	738	1131	2704	3749
164 Hz	744	1132	2711	3747
English	768	1333	2522	3687
Mandarin	957	1328	2813	

Female	F1	F2	F3	F4
151 Hz	879	1410	2870	3941
177 Hz	881	1480	2862	3938
207 Hz	866	1452	2875	3935
240 Hz	922	1464	2878	3948
283 Hz	854	1460	2868	3913
English	936	1551	2815	4299
Mandarin	1104	1593	3188	

Note: F1~F4 are the first four formants of the speech signal.

Acknowledgments

We want to thank Kevin Chun-Hsien Hsu and I-Hui Hsieh for their discussions with Jonathan Evans about the data. We also want to thank William Ching (靖其樂), Chunpai Niu (牛君白), Yizong Huang (黃一中), Yinqi Zhang (張殷綮), and Yu-hsuan Lin (林雨萱), for their valuable assistance. The research would have been impossible without the support of the National Science and Technology Council (MOST 107-2410-H-001-066) and the Academia Sinica Institute of Linguistics. Finally, we owe a debt of gratitude to the participants who patiently evaluated the stimuli.

References

Arsikere, H., Shriberg, E. & Ozertem, U.. "Enhanced end-of-turn detection for speech to a personal assistant." *2015 AAAI Spring symposium series*. 2015.

Baines, R. J., Langelaan, M., de Bruijne, M. C., Asscheman, H., Spreeuwenberg, P., van de Steeg, L., Siemerink K. M., van Rosse F., Broekens M., & Wagner, C. (2013). Changes in adverse event rates in hospitals over time: a longitudinal retrospective patient record review study. *BMJ quality & safety, 22*(4), 290-298.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67, 1–48.

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program] (Version 6.0.52). Retrieved from http://www.praat.org/

Fletcher, H. (1972). *Speech and hearing in communication*. Huntington: Robert E. Krieger.

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Bell System Technical Journal, 12*(4), 377-430.

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., & Perkins, W. H. (1988). Relationship between changes in voice pitch and loudness. *Journal of voice, 2*(2), 118-126.

Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, *97*(1), 35.

Honorof, D. N., & Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *The Journal of the Acoustical Society of America, 117*(4), 2193-2200.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13), 1-26.

Ladd, D. Robert. "Declination.: a review and some hypotheses." *Phonology* 1 (1984): 53-74.

Lehiste, I., & Fox, R. A. (1992). Perception of prominence by Estonia and English listeners. Language and Speech, 35, 419–434.

Li, 11., Tang, C., Lu, J., Wu, J., & Chang, E. F. (2021). Human cortical encoding of pitch in tonal and non-tonal languages. *Nature communications*, *12*(1), 1-12. https://doi.org/10.1038/s41467-021-21430-x

Lockhead, G. (1992). Psychophysical scaling: Judgments of attributes or objects? Behavioral and Brain Sciences, 15(3), 543-558. doi:10.1017/S0140525X00069934

Long, M.. "Human perception and reaction to sound." Architectural acoustics. Vol. 3. New York, NY, USA: Academic, 2014. 81-127.

Lyn-Cook, R., Halm, E. A., & Wisnivesky, J. P. (2007). Determinants of adherence to influenza vaccination among inner-city adults with persistent asthma. *Primary Care Respiratory Journal, 16*(4), 229-235.

Moore, B. C. (2013). *An introduction to the psychology of hearing (6th ed.)*: Brill.

Neuhoff J. G., McBeath M. K., &, Wanzie W. C. (1999). Dynamic frequency change influences loudness perception: A central, analytic process. Journal of Experimental Psychology: Human Perception and Performance, 25, 1050–1059.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rose, P. (1987). “Considerations in the normalisation of the fundamental frequency of linguistic tone.” *Speech communication*, *6*(4), 343-352.

Scharine, A. A., & McBeath M. K. "Natural regularity of correlated acoustic frequency and intensity in music and speech: Auditory scene analysis mechanisms account for integrality of pitch and loudness." *Auditory Perception & Cognition* 1.3-4 (2018): 205-228.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and linguistics compass, 3*(2), 621-640.

Sofianou, A., Martynenko, M., Wolf, M. S., Wisnivesky, J. P., Krauskopf, K., Wilson, E. A., Goel, M. S., Leventhal H., Halm E. A., & Federman, A. D. (2013). Asthma beliefs are associated with medication adherence in older asthmatics. *Journal of general internal medicine, 28*(1), 67-73.

Suzuki, Y., & Takeshima, H. (2004). “Equal-loudness-level contours for pure tones.” *The Journal of the Acoustical Society of America*, 116(2), 918-933.

Watson, P. J., & Hughes, D. (2006). The relationship of vocal loudness manipulation to prosodic F0 and durational variables in healthy adults. *Journal of Speech, Language and Hearing Research*.

Yanagisawa, H., & Mikami, N. (2015). How does expectation change perception? A simulation model of expectation effect. In DS 80-9 Proceedings of the 20th International Conference on Engineering Design (ICED 15) Vol 9: User-Centred Design, Design of Socio-Technical systems, Milan, Italy, 27-30.07. 15 (pp. 149-158).

Yu, V., & Andruski, J. E. "A cross-language study of perception of lexical stress in English." *Journal of psycholinguistic research* 39 (2010): 323-344.