Vocal communication of magnitude across language, age, and auditory experience

Perlman, Marcus; Paul, Jing; Lupyan, Gary

DOI:
10.1037/xge0001103

Document Version
Peer reviewed version

Citation for published version (Harvard):
Perlman, M, Paul, J & Lupyan, G 2021, 'Vocal communication of magnitude across language, age, and auditory experience', *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001103

Vocal communication of magnitude across language, age, and auditory experience

Marcus Perlman[a], Jing Paul[b], and Gary Lupyan[c]

[a]University of Birmingham

[b]Agnes Scott College

[c]University of Wisconsin-Madison

Corresponding author:

Marcus Perlman

m.perlman@bham.ac.uk

Abstract

Like many other vocalizing vertebrates, humans convey information about their body size through the sound of their voice. Vocalizations of larger animals are typically longer in duration, louder in intensity, and lower in frequency. We investigated people's ability to use voice-size correspondences to communicate about the magnitude of external referents. First, we asked hearing children, as well as deaf children and adolescents, living in China to improvise non-linguistic vocalizations to distinguish between paired items contrasting in magnitude (e.g. a long vs. short string, a big vs. small ball). Then we played these vocalizations back to adult listeners in the United States and China to assess their ability to correctly guess the intended referents. We find that hearing and deaf producers both signalled greater magnitude items with longer and louder vocalizations and with smaller formant spacing. Only hearing producers systematically used fundamental frequency, communicating greater magnitude with higher f_o. The vocalizations of both groups were understandable to Chinese and American listeners, although accuracy was higher with vocalizations from older producers. American listeners relied on the same acoustic properties as Chinese listeners: both groups interpreted vocalizations with longer duration and greater intensity as referring to greater items; neither American nor Chinese listeners consistently used f_o or formant spacing as a cue. These findings show that the human ability to use vocalizations to communicate about the magnitude of external referents is highly robust, extending across listeners of disparate linguistic and cultural backgrounds, as well as across age and auditory experience.

Keywords: cross-cultural; nonverbal communication; vocalization; sound symbolism; iconicity

1. Introduction

1.1. Correspondence between vocalization and magnitude

The ability to appreciate magnitude in its various forms is ubiquitous in the lives of most animals. For example, the size of an animal's body, correlated with strength and fighting ability, is an important factor in competition for food, mates, and territory (Clutton-Brock et al., 1977; Clutton-Brock & Albon, 1979; Fitch & Hauser, 2003; Modig, 1996; Schuett, 1997). Those animals with vocal tracts – i.e., terrestrial vertebrates including amphibians, reptiles, birds, and mammals – typically provide acoustic cues to their body size in their vocalizations (Fitch & Hauser, 2003). Two such cues are the amplitude and duration of vocalizations. Larger animals tend to have greater lung capacity, allowing them to produce longer and louder sounds (Gillooly & Ophir, 2010). Animals that evolve mechanisms to increase the amplitude and duration of their vocalizations will sound bigger to other animals (Hewitt et al., 2002).

The frequencies of vocalizations provide another cue to size (Morton, 1977). The relationship between frequency and size – modeled by the source-filter theory of vocal production (Taylor & Reby, 2010) – arises from the physics of how sound waves interact with the vocal-production system. The fundamental frequency (f_o) of vocalization is determined by the length and mass of the vocal folds. Longer and heavier vocal folds vibrate at a slower frequency than smaller vocal folds. Therefore, larger animals, which typically possess larger larynxes and vocal tracts, will produce sounds with a lower f_o (Bowling et al., 2017; Morton, 1977; Taylor & Reby, 2010; Titze, 1994).

However, because vocal organs are soft and not tightly constrained by skeletal structure, studies of humans show that fundamental frequency is not reliably associated

with body size within age and sex classes (Künzel, 1989). A further complicating factor is that f_0 is also determined, not just by the length and mass of the vocal folds, but by sub-glottal air pressure and laryngeal muscle tension. A more reliable cue of body size is the resonant frequencies of vocalizations, which are determined by the length of the vocal tract – a property that is constrained by skeletal structures (Fitch, 2000). Longer vocal tracts, acting as a larger resonating chamber, will typically produce lower formant frequencies with reduced average distances between resonating clusters.

Like other vocalizing animals, humans also convey information about the size of their body through the sound of their voice. Similar to other animals, f_0 turns out not to be a very reliable cue to body size in humans either, once controlling for sex and age (adults compared to juveniles) (Pisanski et al., 2014; Pisanski & Bryant, 2019). Instead, the formant frequencies of vocalizations (e.g. speech) are a more reliable indicator (Pisanski et al., 2014; Rendall et al., 2005). Taller people typically have longer vocal tracts with lower resonant frequencies with reduced average distances between formants (Titze, 1994).

Listeners are sensitive to frequency information of vocalizations when making judgments about the size of the vocalizer. However, these judgments of speaker size are only moderately accurate (Pisanski & Bryant, 2019). Although f_0 is not actually a reliable cue, listeners nevertheless judge both male and female voices with a low f_0 as physically larger and stronger, as well as more dominant and more masculine (Pisanski & Bryant, 2019). Listeners are also sensitive to the formant frequencies of vocalizations – the more reliable cue to body size – associating vocalizations having lower, more closely spaced formant frequencies with greater speaker size.

This sensitivity to the correlation between vocalization frequency and body size is evident from an early age. Infants at just three months responded to the relationship between the size of an animal puppet and its call frequency, showing that they expected the larger puppet to produce a lower frequency sound (Pietraszewski et al., 2017). Interestingly, the ability to accurately estimate body size from vocal cues does not require visual experience. People who are blind are equally capable as sighted people of assessing body size from the f_0 and formant frequencies of vocalizations (Pisanski, Oleszkiewicz, et al., 2016; Pisanski et al., 2017).

On some theories, the function of vocalizations to communicate body size is thought to be so important that it figures into the evolution of speech itself. For example, some have argued that human ancestors evolved the flexible and fine motor control needed for speech in order to better exploit listeners' expectations about the relationship between qualities of the voice and body size (Fitch, 2000; Pisanski, Cartei, et al., 2016). It would have been advantageous to sound bigger in some contexts, and smaller in others. Indeed, this association between vocal quality and body size can be flexibly recruited by speakers. When speakers of different languages (Polish speakers, Cuban Spanish speakers, and Canadian English speakers) performed a vowel repetition task while pretending to be physically larger or smaller in size, they spontaneously modulated their f_0 and formant frequencies in correspondence with the size they intended to convey (Pisanski, Mora, et al., 2016).

1.2. *Vocalizing magnitude of external referents*

The work discussed so far highlights the deeply rooted ways that people associate vocal qualities, especially f_0 and formant frequencies, with the body size of the vocalizer,

and also how these associations might figure into the evolution of speech. But if we are to fully understand their significance in human communication, it is important to also examine the extent to which these associations can be transferred to external referents. Do people readily make use of these voice-size correspondences to communicate about different-sized referents more generally?

Experiments on sound symbolism show that people consistently use the frequency of speech sounds to inform their judgments of the size of an unknown referent. For example, an early experiment found that people are inclined to choose nonsense words containing high front vowels as labels for smaller objects, and those containing low back vowels for larger ones (Sapir, 1929), and more recent experiments have confirmed this finding (Thompson & Estes, 2011). This correspondence is thought to derive from the formant frequencies of the vowels: high front vowels (/i/) are characterized by a high second formant, and especially a large dispersion of the first and second formant frequencies (Ohala, 1983). Notably, this sound symbolism is evident across the vocabularies of spoken languages which show a prevalence of higher-frequency vowels in words denoting *small* (Blasi et al., 2016; Haynie et al., 2014; Ultan, 1978).

Several experiments suggest that speakers spontaneously and flexibly draw on correspondences between magnitude of external referents and vocalization frequency, amplitude and duration. In a study of infant directed speech (Nygaard et al., 2009), three adult English speakers produced the carrier phrase, “Can you get the *blicket* one?” while the meaning of “blicket” was varied in referring to several semantic dimensions, including *tall* versus *short* and *big* versus *small*. In line with vocal cues to body size, when referring to a larger compared to smaller referent, English speakers produced

"blicket" with lower f_0, as well as with higher intensity and greater duration. When referring to a taller vs. shorter referent, "blicket" had greater amplitude and greater duration. A subsequent experiment played the recorded sentences to listeners and found they could select the correct meaning from its dimensional opposite. In another study, participants were asked to read stories that contained either big or small elements (Perlman, Clark, et al., 2015). Speakers tended to pronounce size-related phrases of the story with a corresponding modulation in f_0, e.g., producing phrases like "a giant house" with a relatively lower f_0.

Recent studies have used a charades-like task to investigate people's ability to create non-linguistic vocalizations to communicate various kinds of meanings, including those relating to magnitude (Perlman, Dale, et al., 2015; Perlman & Cain, 2014). These studies have found that people consistently employ iconic signals, for example, expressing *big* with vocalizations that were longer, louder, and lower-pitched than those for *small*; expressing *long* with temporally longer sounds than for *short*; and expressing *many* compared to *few* with the quick repetition of syllables resulting in longer sounds. Using playback experiments, naïve listeners were found to be much higher than chance at inferring the intended meanings of these novel vocalizations (Perlman, Dale, et al., 2015; Perlman & Lupyan, 2018).

This research suggests the possibility that people who do not share a common language may nonetheless be able to communicate about the magnitude of external referents by utilizing universally understood vocal cues to size. Yet a major limitation of this previous work with non-linguistic vocalizations has been the homogenous backgrounds of participants, both vocalizers and listeners, who were all English speakers.

In the current study, we examine the vocalization-size correspondences across cultures, and test whether people are able to use these to support cross-cultural communication.

1.3. Current study

Do people share a universal sense of how to use their voice to express the magnitude of external referents? To find out, we conducted two complementary experiments: a vocal production game along with a playback experiment. To assess the robustness of cross-cultural communication of magnitude, our study focused on participants from different backgrounds of language, age, and auditory experience. In the production game, we asked both hearing children (aged 10-12 years) and deaf children and adolescents (8-20 years) raised in China to create nonlinguistic vocalizations that distinguished between paired items contrasting in magnitude, e.g., a big versus small ball; a long versus a short string. In the playback experiment, we presented the vocalizations back to both American and Chinese adults to test their ability to guess the intended referent. We ask (1) Do hearing and deaf Chinese participants, including younger and older participants, similarly use the duration, intensity, and frequency of their voice to signal magnitude? (2) Are Chinese vocalizers – hearing and deaf, young and old – able to communicate magnitude successfully to American as well as Chinese listeners? and (3) Do American listeners use the same vocal cues to magnitude as Chinese listeners?

2. Vocal Production Game

2.1. Methods

2.1.1. Participants

Our Chinese participants were drawn through a convenience sample obtained by contacts of the second author. We recruited 19 deaf children and adolescents from a special education boarding school in the Hubei Province of China. These participants comprised all the students at the school who were identified by their teachers and principal as having congenital deafness with severe or profound hearing loss, and otherwise normal cognitive functioning. They did not have cochlear implants or wear hearing aids. This information of their histories was obtained through oral communication between the school's teachers and the parents and/or guardians of the children. At the school, the students all received education in both Chinese Sign Language and spoken Mandarin. The mean age of deaf participants was 12.5 years (SD = 3.7 years, range 7-20 years).

Hearing participants were 16 Mandarin-speaking children with normal hearing 10.1 years (SD = 0.83 years, range = 9-12 years). The children attended a day school in the same region of China as the deaf participants. The participants included all the children in the classroom to which we had access. The research was approved by the Institutional Review Board of the University of Wisconsin-Madison under Social and Behavioral Sciences Protocols 2016-0632 and 2018-1415.

2.1.2. Materials

The experiment investigated four dimensions of magnitude, each instantiated by a pair of contrasting items (Figure 1): a *short* vs. a *long* string (*length*), a *small* vs. a *big* ball (*size*), a *little* vs. a *lot* of rice (*amount*), and a *few* (2) vs. *many* (5) marbles (*quantity*).

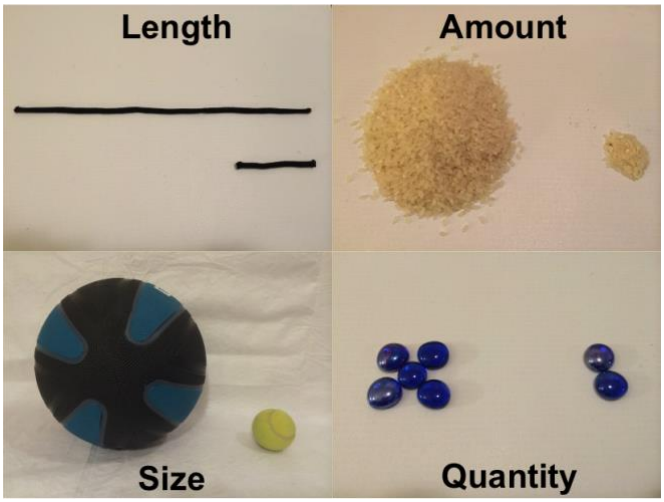

Figure 1. Items used in the vocal communication game. Each pair contrasts a different dimension of magnitude: long vs. short *length* of string, large vs. small *size* of ball, a lot vs. a little *amount* of rice, and a *quantity* of many vs. few marbles.

The items used to represent different dimensions of magnitude were selected opportunistically from materials available at the boarding school.

2.1.3. Design and Procedure

The deaf participants took part in the study at the boarding school they attended. The experiment was conducted by a native speaker of Mandarin Chinese and assisted by

a bilingual teacher at the school who spoke Mandarin and signed standard Chinese Sign Language. The teacher provided instructions in Chinese Sign Language.

Participants were first introduced to the experiment in their home classrooms while in a group. The assisting teacher placed the stimuli – the four contrasting pairs of items – on a desk in front of the class, with the two contrasting items of each pair placed side-by-side. She noted that the objects in each pair were different and invited the students to sign the difference. Their responses suggested that they readily identified the contrast for each contrasting stimulus pair (e.g., “big” and “small” size for the balls).

After going through the items, the teacher told the students that they would play a “guessing game”. She explained that the experimenter would point to one of the two items of each pair, and they would make a vocal sound to communicate the selected item to their teacher, whose back would be turned so that she could not see. The participants were encouraged to make a meaningful sound that they thought would help their teacher choose the correct item and were asked not to use Mandarin words.

The students participated individually in a quiet office at the school immediately following the classroom introduction. The participant was seated at a table beside the experimenter and the teacher sat with her back to the table. All of the items were placed in a row in front of them, with paired items placed next to each other. The instructions were repeated by a signing assistant as necessary. The participant was seated about 20 inches from the recorder, which was positioned between the participant and the experimenter.

On each trial, the experimenter first announced, in Mandarin, the label of the item pair that would be tested so that the (hearing) teacher knew which pair to select from. The

experimenter next gestured to indicate the selected pair of items to the participant and pointed specifically to the target item. The participant then produced a vocalization to communicate the target to the teacher. In response, the teacher turned toward the pair of items and pointed to indicate her guess of which one had been selected. No other feedback was provided. The session was recorded on an IPhone 6 plus, using Voice Recorder HD 9.2 software. Recordings were saved as .wav files sampled at 22.05 kHz.

Each item was tested once for a total of eight trials. First one item from all of the pairs was tested, and then the second item from the pairs was tested. The order of items, including the order of greater and lesser magnitude items, was randomized between participants.

This same basic procedure was used with the hearing children. The only notable difference was that the instructions and experiment were conducted in Mandarin Chinese.

2.1.4. Acoustic measurements of vocalizations

Acoustic measurements of the vocalizations were made with Praat phonetic analysis software version 6.0.34 (Boersma, 2001). All measurements were made blind to the intended referent of the vocalization. The onset and offset of each vocalization were marked by listening and visual inspection of the waveform and spectrogram. The fundamental frequency of each vocalization was measured using the Praat autocorrelation algorithm. When voicing was weak or inconsistent, resulting in what appeared to be inaccurate pitch tracking, the voicing threshold was increased to exclude spurious readings. The pitch range was set to the standard range of 75 Hz to 500 Hz, but was adjusted upwards as needed to accurately measure higher frequency vocalizations. Octave errors (spurious doubling or halving of f_0 tracking) were addressed by adjusting

the pitch range, and when spurious f_o was detected over quiet, voiceless segments, this was removed by increasing the voicing threshold before taking the f_o measurement. The average frequencies of the first, second, third, and fourth formants were extracted with the Praat 'Get … formant' functions using the standard settings. From these values, we computed formant spacing (ΔF) following the method described in (Reby & McComb, 2003): fitting a regression line through formants F_1-F_4 as predicted by $F_i=2_i-1$, i.e., 0.5 for F_1, 1.5 for F_2, etc. and with the intercept set to 0. The formant spacing is the slope (i.e., regression coefficient) (see also Pisanski et al., 2014) .

We used Praat script to measure the duration and intensity of the vocalizations in batch. Intensity, the root mean square amplitude measured in decibels, was obtained with the 'Get intensity' function, using mean energy as the averaging method, and subtracting mean pressure from the measurement. Because f_o and duration were right skewed, analyses were performed on log transformed values of these variables. (Notably, the log transformation of f_o corresponds with the perception of pitch from f_o, which is log based.)

2.1.5. Statistical analysis

All statistical analyses were performed in R version 3.4.3 (R Core Team, 2015). Analyses with mixed effects models were conducted using the lme4 package version 1.1-21 (Bates, Maechler, Bolker, & Walker, 2015). To establish statistical significance, we used *anova()* to compare a base model without the factor of interest to a model with the factor. We report the p-value corresponding the difference in deviance between the two models, based on a likelihood ratio test (Douglas Bates et al., 2014). The scripts and model specifications can be found on the Open Science framework at https://osf.io/rkjqs/.

We began by constructing linear mixed effects models to test whether magnitude affected the duration, intensity, and pitch of the produced vocalizations. The models included subject and item (ball, string, etc.) random intercepts and a magnitude random slope for participants (see OSF for full model specification).

2.2. Results

There were 277 vocalizations produced by the 35 participants. One hearing participant recorded vocalizations on only six of eight trials and a second hearing participant recorded vocalizations for only seven trials. Table 1 displays the descriptive statistics for the vocalizations. Listening through the vocalizations confirmed that the participants followed the instructions to avoid producing real Chinese words. The vocalizations varied widely in their acoustic properties, with some sounding more speech-like while others sounded more like affective vocalizations. Most, but not all, of the sounds were voiced, and many exceeded the typical pitch and amplitude range normally associated with speech.

Table 1. Acoustic measurements of vocalizations.

Contrast	Group	Duration (s)	Intensity (dB)	f_0 (Hz)	ΔF (Hz)
Greater (all)	Hearing	0.65 (0.31)	64.8 (5.8)	299.7 (86.2)	1168 (82.6)
Lesser (all)	Hearing	0.51 (0.22)	59.8 (4.3)	267.8 (58.6)	1186 (86.6)

Greater (all)	Deaf (older)	0.62 (0.12)	61.4 (6.2)	310.0 (151.9)	1105 (41.4)
Lesser (all)	Deaf (older)	0.52 (0.17)	56.0 (5.4)	280.1 (109.2)	1138 (52.7)
Greater (all)	Deaf (younger)	0.83 (0.30)	61.1 (5.5)	404.2 (281.6)	1130 (40.5)
Lesser (all)	Deaf (younger)	0.77 (0.30)	60.2 (6.2)	410.0 (290.9)	1130 (45.3)
Long (string)	Hearing	0.84 (0.46)	65.6 (5.9)	288.1 (86.4)	1168 (113)
Short (string)	Hearing	0.57 (0.19)	60.1 (4.7)	286.8 (72.2)	1186 (103)
Long (string)	Deaf (older)	0.62 (0.15)	61.8 (5.9)	303.3 (161.4)	1103 (36.9)
Short (string)	Deaf (older)	0.49 (0.17)	55.7 (3.7)	278.3 (111.7)	1131 (42.9)
Long (string)	Deaf (younger)	0.78 (0.16)	62.2 (5.2)	398.6 (279.6)	1135 (45.3)
Short (string)	Deaf (younger)	0.70 (0.12)	59.8 (6.6)	403.9 (299.0)	1123 (60.1)
Big (ball)	Hearing	0.50 (0.18)	65.0 (5.7)	298.5 (88.2)	1178 (80)

Small (ball)	Hearing	0.43 (0.24)	59.1 (4.5)	260.3 (64.4)	1180 (90.1)
Big (ball)	Deaf (older)	0.64 (0.09)	62.5 (5.8)	333.4 (142.3)	1104 (35.0)
Small (ball)	Deaf (older)	0.50 (0.19)	56.4 (5.9)	266.5 (117.3)	1138 (49.3)
Big (ball)	Deaf (younger)	0.81 (0.17)	61.9 (5.9)	444.9 (377.6)	1133 (37.1)
Small (ball)	Deaf (younger)	0.87 (0.22)	60.3 (7.3)	397.4 (271.2)	1136 (42.3)
Lot (rice)	Hearing	0.63 (0.19)	64.2 (5.9)	302.5 (81.8)	1170 (68.1)
Little (rice)	Hearing	0.57 (0.19)	59.6 (4.5)	247.4 (53.1)	1207 (90.8)
Lot (rice)	Deaf (older)	0.60 (0.13)	60.8 (5.6)	307.1 (162.2)	1095 (49.2)
Little (rice)	Deaf (older)	0.55 (0.14)	55.5 (6.8)	279.1 (118.0)	1130 (59.0)
Lot (rice)	Deaf (younger)	0.87 (0.43)	60.0 (5.9)	363.4 (155.3)	1136 (43.6)
Little (rice)	Deaf (younger)	0.78 (0.21)	61.6 (7.0)	426.8 (350.8)	1139 (37.8)

Many (marbles)	Hearing	0.61 (0.25)	63.8 (6.1)	310.6 (96.7)	1157 (66.3)
Few (marbles)	Hearing	0.47 (0.22)	60.3 (3.6)	278.1 (40.6)	1172 (59.5)
Many (marbles)	Deaf (older)	0.62 (0.12)	60.6 (8.1)	296.0 (165.6)	1119 (46.2)
Few (marbles)	Deaf (older)	0.52 (0.20)	56.3 (5.5)	296.4 (107.3)	1153 (63.3)
Many (marbles)	Deaf (younger)	0.87 (0.39)	60.4 (5.6)	409.8 (308.3)	1116 (38.4)
Few (marbles)	Deaf (younger)	0.75 (0.16)	59.1 (4.4)	411.8 (283.1)	1123 (42.7)

Note. Means are shown with standard deviations in parentheses.

Figure 2 displays results from the vocal production game. For display purposes only, we median-split the deaf participants into a younger group, M_{age}=9.5, and an older group, M_{age}=16.8. Hearing participants produced vocalizations that distinguished the greater magnitude items with longer duration *β = .48, 95% CI = [.19, .76], t = 3.24, p = .003*, and higher intensity, β = .88, 95% CI = [.67, 1.08], t = 8.26, p = < .001. Vocalizations for greater magnitude items also had a *higher f_o, β = .41, 95% CI = [.04, .77], t = 2.19, p* = .048. However, when intensity and duration were added as covariates, this effect was attenuated, β = .36, 95% CI = [-.07, .79], t = 1.64, p = .113. There was no reliable effect of magnitude on the formant spacing of vocalizations, β = -.22, 95% CI = [-.51, .07], t = -1.50, p = .137.

As can be seen in Figure 2, the effects on duration were strongest within the *length* and *size* dimensions, whereas the effects on intensity were robust across each dimension. The effect on pitch was less consistent across dimensions, only being reliably used when signaling differences in *amount*, and marginally, *size*.

Deaf participants communicated items of greater magnitude using vocalizations with longer duration, $\beta = .39$, 95% CI = [.10, .68], $t = 2.67$, $p = .014$. An interaction between magnitude and age, $\beta = .08$, 95% CI = [.01, .15], $t = 2.17$, $p = .040$, indicated that this effect was stronger with older participants. Vocalizations for greater magnitude items were also louder, $\beta = .49$, 95% CI = [.22, .77], $t = 3.50$, $p = .002$, an effect also more prominent with older participants, $\beta = .10$, 95% CI = [.04, .16], $t = 3.23$, $p < .001$. The f_0 of their vocalizations did not vary by magnitude, $\beta = .06$, 95% CI = [-.05, .18], $t = 1.08$, $p = .287$, nor did it interact with age, $\beta = .02$, 95% CI = [-.01, .05], $t = 1.51$, $p = .143$. The main effect of magnitude was still not significant when adding duration and intensity as covariates, $\beta = -.03$, 95% CI = [-.15, .09], $t = -.50$, $p = .620$. Vocalizations for greater magnitude items had smaller formant spacing, $\beta = -.34$, 95% CI = [-.64, -.04], $t = -2.24$, $p = .035$. There was a significant interaction between magnitude and age on formant spacing, $\beta = -.11$, 95% CI = [-.18, -.04], $t = -2.96$, $p = .006$.

As shown in Figure 2, the vocalizations of older deaf participants were patterned similarly to hearing children: the strongest effects on duration were seen in the *length* and *size* dimensions, while the effect on intensity was more consistent across the dimensions. Older deaf participants did not reliably use pitch to signal magnitude, except in the limited case of *size*. The vocalizations of younger deaf participants generally patterned like those of hearing and older deaf participants with respect to duration and intensity, but

this only reached significance for the effect of *length* on duration. Younger deaf participants did not show evidence of the use of pitch in any dimension.

We next directly compared deaf and hearing participants. Across both groups, participants signaled greater magnitude with vocalizations of longer duration, $\beta = .42$, 95% CI = [.23, .60], $t = 4.33$, $p < .001$, greater intensity, $\beta = .65$, 95% CI = [.47, .83], $t = 7.00$, $p < .001$, and higher *f*o, $\beta = .15$, 95% CI = [.02, .28], $t = 2.32$, $p = .027$. However, the effect of *f*o was no longer significant after adding duration and intensity as covariates, $\beta = .05$, 95% CI = [-.08, .19], $t = .74$, $p = .460$. Both groups produced vocalizations with smaller formant spacing for greater magnitude items, $\beta = -.24$, 95% CI = [-.43, -.06], $t = -2.54$, $p = .013$. There was no statistical difference between deaf and hearing participants in the use of duration, $\beta = .21$, 95% CI = [-.16, .58], $t = 1.13$, $p = .264$, intensity, $\beta = .32$, 95% CI = [-.03, .67], $t = 1.79$, $p = .080$, *f*o, $\beta = .18$, 95% CI = [-.07, .42], $t = 1.42$, $p = .166$, or in formant spacing, $\beta = -.04$, 95% CI = [-.42, .33], $t = -.22$, $p = .823$.

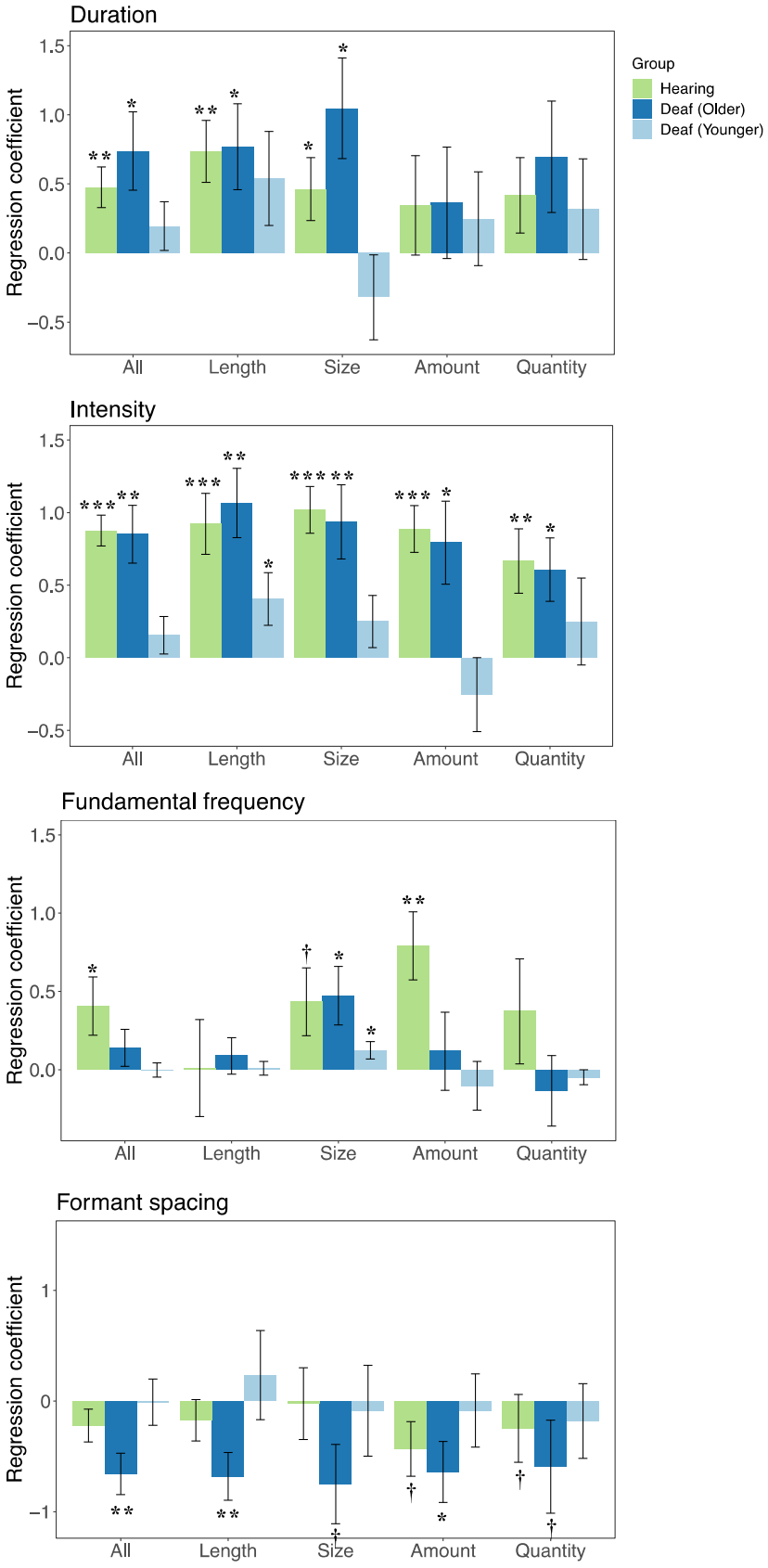
Duration
Intensity
Fundamental frequency
Formant spacing
Regression coefficient
Group
Hearing
Deaf (Older)
Deaf (Younger)
All
Length
Size
Amount
Quantity

Figure 2. Panels show regression coefficients with standard errors from the analyses of the duration, intensity, fundamental frequency, and formant spacing of vocalizations. Results for deaf participants are shown for ten younger participants aged 7 to 12 years and nine older participants aged 12 to 20 years. Positive coefficients indicate that vocalizations for the greater magnitude item had larger values for the given acoustic property. For example, in the top panel, the positive values of the length bars indicate that vocalizations for the long item were longer in duration than vocalizations for the short item. The complete statistical models and results are in the OSF repository. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$

3. Playback Experiments

We next examined whether the vocalizations produced in the vocal production game were understandable to naïve listeners, and whether their comprehension was affected by language and culture. Do Chinese listeners have an easier time understanding the vocalizations compared to American listeners? Are the vocalizations produced by deaf participants as understandable as those from hearing participants? Finding that the vocalizations are understandable by American and Chinese listeners alike would suggest that there is something about how people encode magnitude using their voice that extends across cultural and linguistic experience. Finding that the vocalizations of deaf participants are understandable to both groups of listeners would suggest that these magnitude-vocalization associations develop robustly across large differences in experience with sound.

3.1. Methods

3.1.1. American participants

We recruited 393 American participants from Amazon Mechanical Turk. The number of participants was chosen with the aim that each vocalization was heard by at least 10 listeners after excluding participants who did not pass the screening criteria. Of the 393 American participants, we excluded 8 (2.0%) because they did not respond accurately to the initial test trial to ensure their computer speakers were correctly functioning. An additional 19 (4.8%) participants were excluded because they failed to accurately identify at least one of the two 'clap' trials, which served as a check that they were attending properly to the task. We also excluded 30 participants (7.6%) because they matched a vocalization to the 'clap' icon on at least two trials. In total, we excluded the data of 57 (14.5%) participants. The remaining 336 participants included 186 males, 149 females, and 1 participant who did not disclose their gender. Their mean age was 34.6 years (SD = 10.7 years). Finally, we excluded any individual trials in which participants matched a vocalization to the 'clap' icon, which amounted to 57 out of the remaining 2676 target trials (2.1%).

3.2.2. Chinese participants

We recruited 162 participants living in China. As with American listeners, we aimed to ensure each vocalization was heard by at least 10 participants after filtering those who did not pass the screening criteria. Participants were recruited through the Chinese social media 'WeChat' and 'QQ,' and by word of mouth. J. Paul sent the

individual URLs via social media to eligible Chinese participants in her contacts in China, who were asked to complete a URL and also to look for additional participants.

All participants were native Mandarin Chinese speakers living in China at the time of data collection. Most participants were monolingual speakers who had no or little knowledge of a foreign language, and a small percentage were college students who studied English as a required subject in school. We purposely avoided participants whom we knew to have extensive experience with a foreign language or culture. Most participants were recruited from Hubei Province, and the rest from the cities of Beijing and Tianjin, and the provinces of Guangdong and Shandong.

Chinese listeners were screened as much as possible according to the same criteria as the American listeners. One difference was that participants in this version were provided a 'no sound' response to indicate that they did not hear a sound play in the trial. Thus, in addition to the other criteria, we excluded 5 participants (3.1%) who indicated that no sound played on 25% or more of trials. Another difference was that the computer speakers test question was played at the start of each block for Chinese participants, and therefore these exclusions were implemented on a block basis. This led to the removal of 46 blocks (15.4%) from 25 individuals because the participant failed to correctly respond to the speaker trial. Additionally, 22 participants (13.6%) were excluded because they missed half (four) or more clap test trials. Finally, 2 of the 162 participants (1.2%) were excluded for guessing 'clap' in response to a quarter or more (≥ 8) vocalization trials. In total, 33 of 162 participants (20.4%) were excluded from analysis because of difficulties with the task. The remaining 129 participants included 76 females and 53 males, with a mean age 31.2 years (sd = 10.7).

From these data, we screened the remaining 25 (0.7%) ‘no sound’ responses of the remaining 3681 target trials. Finally, we excluded trials in which the participant guessed ‘clap’ in response to a vocalization, which were 69 (1.9%) of the remaining total of 3664 trials.

3.1.3. Design and procedure with American participants

The stimuli used in the playback experiments were the 277 vocalizations produced in the vocal production game, which included 152 vocalizations by the 19 deaf children and adolescents and 125 vocalizations by the hearing children. Participants were presented with the vocalizations in a Qualtrics survey that was accessed through Amazon Mechanical Turk. Participants listened through their own speakers (e.g. headphones, ear buds, computer speakers), and were free to adjust their volume however they preferred.

Participants then listened to eight vocalizations in a session, which comprised the complete set of vocalizations from a single producer in the production game. The vocalizations were presented in a randomized order for each session. On each trial, participants listened to a vocalization as they viewed pictures of the two relevant items contrasting in magnitude. Each picture showed both contrasting items with the target item circled, thereby highlighting the difference in magnitude between them. Participants made their selection by ticking the box corresponding to their chosen picture. They could listen to each vocalization as many times as they wished. As an attention check, the trials included two clapping sounds for which they needed to select a picture of clapping hands. These attention checks confirmed whether participants were paying attention and could actually hear the sounds.

A soundcheck at the beginning of the session confirmed that participants could hear the vocalizations through their speakers. This consisted of a concatenated sequence of all eight of the vocalizations on which participants would subsequently be interpreting. They were asked to listen to this sequence, count the number of sounds they heard, and enter the number into a blank. In addition to serving as a soundcheck, this also familiarized participants with the range of vocalizations, allowing them to adapt to the producer's voice.

After the test trial, participants were presented with the following instructions before beginning the test trials.

> Which item do you think the person was referring to with each vocalization? Select the picture with the circled item that you think the person was trying to communicate. The pictures include a long string, a short string, a big ball, a small ball, a lot of rice, a little rice, 5 stones, 2 stones, and a picture of clapping hands.
>
> Listen to each sound as many times as you need. There will be a total of 10 sounds. A few of the sounds will be the sound of two claps. Choose the picture of the clapping hands for this sound.

3.1.4. Design and procedure with Chinese participants

The playback experiment with Chinese participants was designed to be as similar as possible to the experiment with American listeners. However, because we did not have access to as many Chinese participants, each participant listened to the vocalizations of four producers (instead of one). These were divided into four blocks, each consisting of the eight vocalizations from a single producer. Each block also contained two 'clap' screening trials. A given survey presented vocalizations from only hearing or deaf

producers. For each set of producers, two versions of the survey were created, one the reverse order of the second. (One pair of surveys only contained three blocks because there was an odd number of producers.) As in the experiment with American listeners, Chinese listeners were also exposed to the full set of vocalizations prior to the experimental trials. In this case, listeners began each of the four blocks with a soundcheck in which they counted a concatenated sequence of all eight vocalizations on which they would be tested in the block. Thus, they were similarly exposed to the full set of vocalizations from a given producer prior to guessing their meanings in the test trials. Instructions were adapted from the English instructions for the four-block format, translated into Mandarin, and presented in Chinese script. Additionally, a 'sound did not play' response was added to the survey.

3.1.5. Analysis

As in the prior experiment, statistical analyses were conducted in R version 3.4.3 (R Core Team, 2015), and mixed effects models analyses were performed with the lme4 package version 1.1-21 (Bates, Maechler, Bolker, & Walker, 2015). Analysis scripts and full model results are at https://osf.io/rkjqs/.

We used mixed-effects logistic regression to determine whether listeners were more accurate than chance at selecting the correct referents of the vocalizations, first testing each listener group (American, Chinese) separately. Chance performance was set to 50% (excluding attention checks) by offsetting the intercept by the log odds of 0.5. The models included random intercepts by listener, producer, and the intended referent

(big, few, small, etc.). For example, the model predicting accuracy from producer group (deaf vs. hearing) for American listeners was:

```
glmer(is_correct ~ offset(logit(chance)) + hearing_vs_deaf +
(1|listener_id) + (1|producer_id) + (1|intended_referent),
data=filter(data,exp==”american”), family=binomial)
```

We also used mixed-effects logistic regression to examine how the selections listeners made were influenced by the acoustic properties of the vocalizations. The models included each of the four acoustic variables as predictors. They included random intercepts for listener, producer, and intended referent. For example:

```
glmer(greaterResp ~ scale(log(duration)) +scale(intensity)+
scale(log(frequency)) + scale(form_slope) + (1|responseId) + (1|producer)+
(1|meaning), data=filter(data, exp==”chinese"), family=binomial)
```

3.2. Results

3.2.1. American listeners

Overall guessing rates by production group and by item pair are shown in Figure 3. American listeners were 63% accurate at guessing the intended magnitude of vocalizations produced by hearing children, a rate significantly higher than chance, b_0 = .58, 95% CI = [.25, .90], z = 3.49, p < .001. They were 60% accurate for vocalizations produced by deaf children, also significantly higher than chance, b_0 = .48, 95% CI = [.17, .79], z = 3.07, p = .002. Guessing rates of vocalizations produced by the hearing and deaf

participants were not significantly different, b_0 = .12, 95% CI = [-.25, .48], z = .63, p = .530. Owing to the large age difference among the deaf vocalizers, we examined whether guessing rates were different for vocalizations made by older vs. younger participants, finding that vocalizations produced by older deaf participants were marginally more accurate b = .06, 95% CI = [-.01, .13], z = 1.71, p = .087. As shown in Figure 3, there was no notable difference in guessing accuracy between item pairs.

We next assessed how listeners' choices of the greater- or lesser-magnitude referent were influenced by the duration, intensity and frequency of the vocalizations. Figure 4 visualizes these results, showing the degree to which particular vocal characteristics led listeners to select one item over the other. The results are shown for all pairs of items together and for each stimulus pair separately. Across all pairs of items, the results showed that listeners were more likely to select the greater-magnitude item in response to longer, b = .60, 95% CI = [.46, .74], z = 8.41, p = < .001, and more intense, b = .43, 95% CI = [.23, .62], z = 4.26, p = < .001, vocalizations. f_0 was not a significant predictor, b = -.05, 95% CI = [-.21, .11], z = -.59, p = .558, nor was formant spacing, b = .00, 95% CI = [-.13, .13], z = .01, p = .989.

3.2.2. Chinese listeners

Overall guessing rates by production group and by item pair are shown in Figure 3. Listeners were 65% accurate at guessing the correct referent of vocalizations produced by hearing children, significantly higher than chance, b_0 = .63, 95% CI = [.31, .94], z = 3.84, p < .001. They were 59% accurate at guessing the meanings of vocalizations by deaf participants, also higher than chance, b_0 = .33, 95% CI = [.13, .52], z = 3.33, p =

.001. Guessing rates were significantly more accurate for vocalizations produced by older deaf vocalizers, b = .07, 95% CI = [.03, .12], z = 3.10, p = .002. As can be seen in Figure 3C, guessing accuracy was similar across items pair, with slightly elevated accuracy for *length*.

Analyses of which acoustic cues influenced Chinese listeners' selections (Figure 4) showed that listeners were more likely to select the greater-magnitude item of the pair in response to vocalizations having longer duration, b = .90, 95% CI = [.78, 1.03], z = 14.18, p = < .001, and higher intensity, b = .72, 95% CI = [.55, .90], z = 8.19, p = < .001. Again, f_0 was not a significant predictor, b = -.11, 95% CI = [-.29, .07], z = -1.22, p = .223. Formant spacing was a marginally significant predictor, with listeners more likely to select greater-magnitude items in response to vocalizations with larger formant spacing, b = .11, 95% CI = [-.02, .24], z = 1.69, p = .092.

3.2.2. Comparing American and Chinese listeners

Combining the data from American and Chinese listeners revealed that both were about equally accurate in guessing the intended magnitudes of the vocalizations, $z<1$, and, taken together, the groups were slightly more accurate at guessing the intended magnitude from vocalizations produced by hearing compared to deaf vocalizers, b = .32, 95% CI = [.03, .60], z = 2.16, p = .031. The combined dataset also revealed a significant effect of age, b = .07, 95% CI = [.02, .12], z = 2.87, p = .004. No other effects or interactions were reliable.

Finally, we tested whether Chinese and American listeners relied on different acoustic cues in selecting the magnitude of referent. Although both Chinese and

American listeners relied on duration and intensity, Chinese listeners were more affected by duration, b = .25, 95% CI = [.11, .38], z = 3.58, p < .001, and intensity b = .19, 95% CI = [.06, .32], z = 2.93, p = .003. The analogous interaction for f_0 was not significant, b = -.07, 95% CI = [-.20, .05], z = -1.14, p = .253. There was a marginally significant interaction such that Chinese listeners were more influenced by formant spacing, b = .13, 95% CI = [-.01, .28], z = 1.78, p = .075.

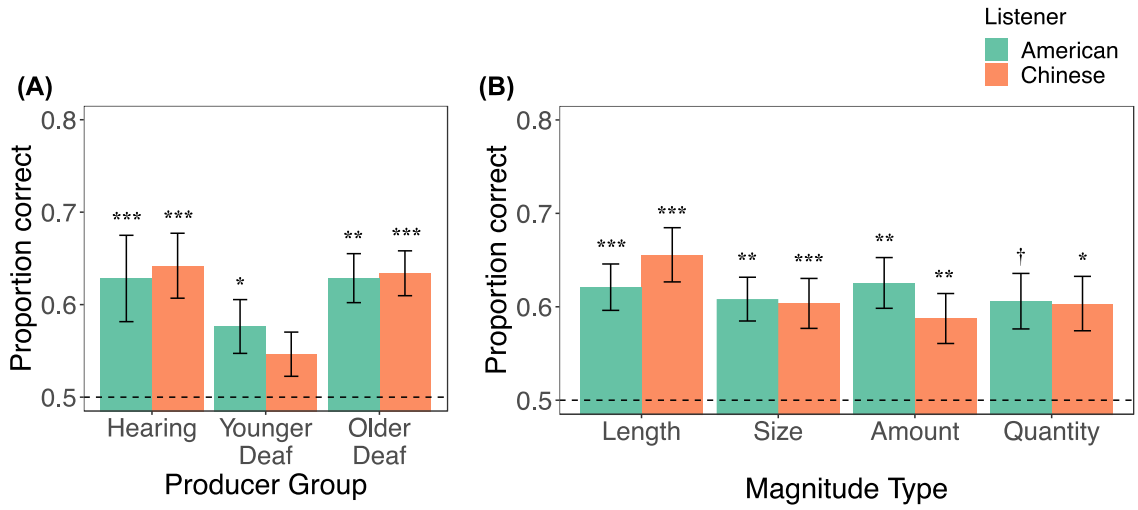

Figure 3. Results from the two playback listening experiments, with American listeners in green and Chinese listeners in orange. (A) shows listener accuracy with vocalizations from hearing, young deaf, and older deaf producers. (B) shows accuracy by item pair. Chance is indicated by the horizontal dashed line. Error bars depict the standard error. Stars indicate significant accuracy compared to chance. ***p<0.001, **p<0.01, *p<0.05

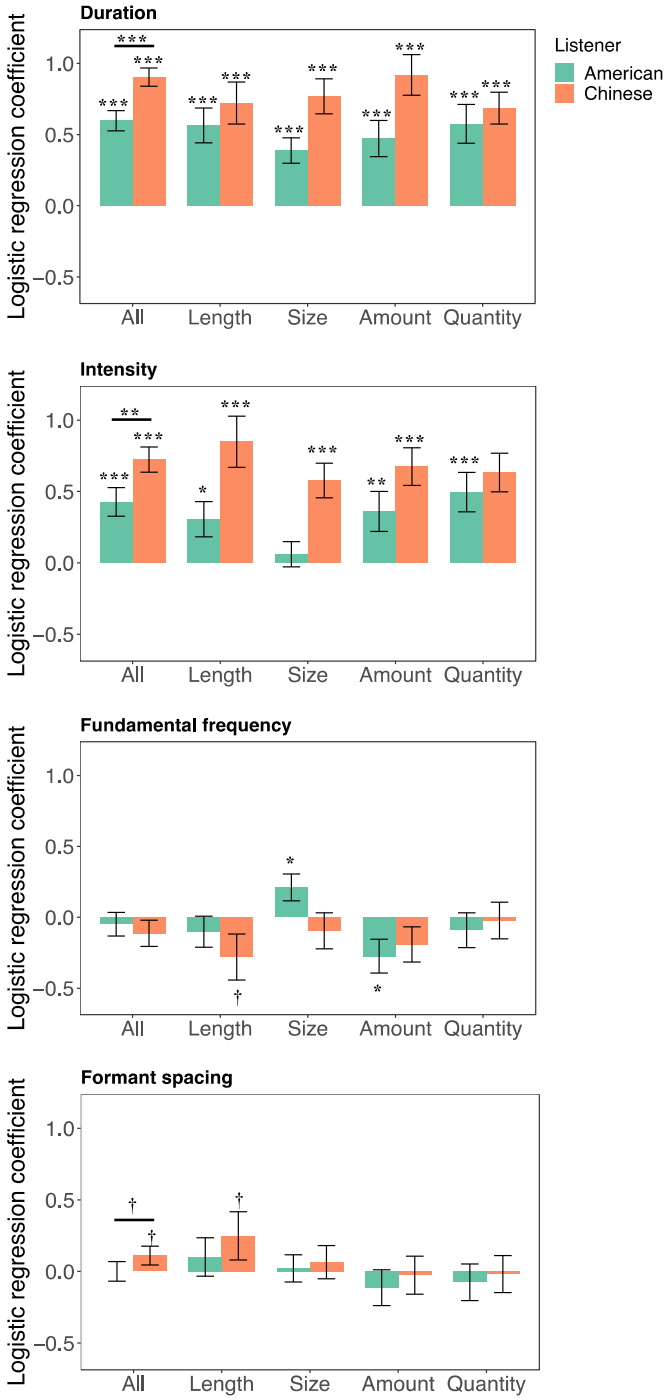

Figure 4. Plots show how strongly listeners were influenced by the acoustic cues of duration, intensity, fundamental frequency, and formant spacing in selecting referents. American listeners are in green and Chinese listeners in orange. The results for all item pairs and each pair separately are indicated on the x-axis of each plot. Logistic regression

coefficients are on the y-axis. Positive coefficients indicate that larger values of the variable were associated with selection of the greater magnitude item. Error bars show the standard errors of the coefficients. Variables were normalized before being entered into the model. Stars plus a line over All items indicates the level of significance for the interaction between that variable and the listener group. ***p<0.001, **p<0.01, *p<0.05

4. General Discussion

Like other vocalizing vertebrates, humans convey information about their body size through the sound of their voice (Fitch & Hauser, 2003). The vocalizations produced by larger animals are typically longer in duration and louder in intensity (Gillooly & Ophir, 2010), as well as lower in fundamental frequency, with lower, more closely spaced formant frequencies (Bowling et al., 2017; Morton, 1977; Pisanski et al., 2014; Rendall et al., 2005). Here, we investigated the extent to which people are able to extend the use of voice-size correspondences to communicate about the magnitude of external referents. First, in a vocal production game, we asked hearing children, as well as deaf children and adolescents living in central China to improvise non-linguistic vocalizations to distinguish between paired items contrasting in magnitude (e.g. a long vs. short string, a big vs. small ball). We then played these vocalizations back to adult listeners in the United States and China to assess the ability of listeners to correctly guess their intended referents. This setup allowed us to assess the role of shared language and culture, and of experience with hearing, in people's ability to use their voices to convey magnitude information about external referents.

All participants, hearing and deaf alike, distinguished the magnitude of items using vocalizations with similar acoustic properties. Deaf children and adolescents

consistently produced vocalizations with longer duration and greater intensity to signal items with greater magnitude, similar to the children in the hearing group, and to American adults in previous studies (Perlman & Cain, 2014; Perlman, et al., 2015). These patterns were somewhat pronounced in the productions of older deaf participants compared to younger deaf participants. Both hearing and deaf participants systematically modulated the formants of their vocalizations with respect to the magnitude of the referent, reflected in a smaller spacing between the formants in reference to greater items. But while hearing participants tended to signal greater magnitude with higher f_0, deaf participants did not modulate f_0, which was the only evident way the vocalizations differed between the groups. Notably, the relationship between magnitude and f_0 became nonsignificant when duration and intensity were added to the regression.

Our sample size of deaf participants was limited, and our analyses are likely to be underpowered as a result. It is unclear how a larger sample might change conclusions about the use by deaf participants of vocal cues to magnitude, but it could reveal further differences between reliance on frequency/amplitude/duration that are too small to detect with our current power. A larger number of deaf vocalizers would also be helpful in assessing how the variability of vocalizations by deaf participants compared to the variability of vocalizations produced by the hearing participants—an analysis that we currently don't have enough power to carry out.

The playback experiments showed that Chinese and American listeners were both able to determine the intended referents of the vocalizations produced in the production game at rates considerably above chance. This success included vocalizations made by both hearing and deaf producers, and it generalized across the different item pairs.

Indeed, American listeners were as accurate as Chinese listeners at guessing the intended magnitude from the vocalizations. Both listener groups were also similar in being more accurate with vocalizations produced by older participants. Moreover, American listeners relied on the same acoustic properties – duration and intensity – as Chinese listeners, although to a somewhat lesser degree. Both groups showed a strong proclivity to interpret vocalizations with longer duration and greater intensity as referring to the greater item. In contrast, neither American nor Chinese listeners were consistent in their use of fundamental frequency or formant spacing as a cue to magnitude.

Taken together, these findings demonstrate that people from different cultural and linguistic backgrounds have a similar sense of how the duration and intensity of vocalizations correspond with the magnitude of external objects. These correspondences are evident both in the innovation of vocalizations to refer to these objects, as well as in the comprehension of those vocalizations. Moreover, the correspondences are highly robust: deafness does not prevent people from effectively using their voices to communicate magnitude information, even to people from a distant culture.

Our finding that older deaf participants used vocal cues more systematically than younger deaf participants suggests that the association of vocal duration/intensity and magnitude is strengthened over developmental time, even in the absence of hearing experience. The precise experience that is required is at present unclear. It is possible participants may have developed their intuition to map vocal duration and intensity to magnitude through their kinesthetic experiences relating magnitude to the force and extension of their actions. With age, they may also gain experience controlling their vocal tracts, as well as visually observing the vocalizations of others, including clues derived

from lip reading. More generally, through education and life experience, they may have improved their ability to reason about the task, resulting in the more systematic use of vocal cues.

Along with the robust correspondence we observed between duration/intensity and magnitude, we found that the correspondences with vocalization frequency were more variable. Both hearing and participants produced vocalizations for lesser magnitude referents with more widely spaced formants compared to vocalizations for greater magnitude referents. For deaf participants, this correspondence strengthened with age. This points to modulations of the length of the vocal tract, including oral features such as tongue position and lip rounding/spreading. For example, in producing vocalizations for greater magnitude items, participants might have increased the length of the vocal tract and the size of their oral cavity by protruding and rounding their lips.

In contrast, only hearing, and not deaf, participants consistently used fundamental frequency to distinguish magnitude. This result is unsurprising considering the fine motor control required for f_o modulation (Fitch, 2010), which may not be as finely tuned for deaf participants without access to auditory feedback (e.g., Boothroyd, 1973). Perhaps more surprising is that hearing participants reliably signaled greater magnitude items with *higher* f_o. As this relationship was secondary to that between magnitude and duration/intensity, it may reflect a contribution of articulatory effort, which is positively associated with f_o in addition to intensity and duration (Gussenhoven, 2002). As muscular tension goes up, f_o rises (Titze, 1994). This suggests that hearing participants might have signaled greater magnitude with higher f_o because it is associated with more forceful vocalizations.

Whatever the explanation, these findings indicate that f_0 was not a primary property of vocalizations used to communicate magnitude. This conclusion contradicts the size frequency code which posits that the connection between low-frequency vocalizations and larger body size is rooted in our vertebrate phylogeny (Ohala, 1994). It also contradicts perceptual experiments showing that young children associate high-pitched/low-pitched tones with small/large objects (e.g., Mondloch & Maurer, 2004b), and similarly, that blind individuals make this same cross-modal association in an auditory-tactile task (Hamilton-Fletcher et al., 2018). However, considering that formant spacing reliably corresponded with magnitude, our findings are consistent with research showing that formants provide a more robust cue to body size than fundamental frequency (Pisanski et al., 2014; Rendall et al., 2005).

A further challenge to the size-frequency code comes from the comprehension experiment: while both American and Chinese listeners reliably used intensity and duration as guides to magnitude, neither group relied in any consistent way on frequency information. Their matching was not consistently influenced by f_0 or formant spacing. These findings do not fit with the strong claim of an innate and automatic sense of correspondence between referent size and frequency of vocalizations. They are, however, consistent with accounts in which the use of frequency is more variable and context-specific, which could lead to contradictory findings depending on task demands. It is possible, for example, that the low-f_0 – large size correspondence is specific to animate referents and absent for the kinds of inanimate referents we used (but see Evans & Treisman, 2010; Gallace & Spence, 2006, though these involve RT differences in speeded matching tasks with tones, rather than vocal production).

In addition to our current data, a few recent studies have also found variability in the use of f_0 to represent magnitude and related meanings. For instance, although one experiment found that American adults produced lower-pitched vocalizations to convey the concept of 'big' (Perlman & Cain, 2014), participants in a subsequent experiment did not (Perlman, et al., 2015). In contrast, participants in both experiments were highly consistent in using duration and intensity. Another recent study finding evidence contrary to the size-frequency code investigated the expression of politeness in Korean (Idemaru et al., 2019). Politeness is thought to be, in part, rooted in dominance relationships that depend on body size – people may raise their fundamental frequency to appear smaller and more submissive – and thus it has been proposed as a prime example of the manifestation of the size-frequency code in spoken languages (Ohala, 1994). Idemaru et al. (2019) found that while Korean listeners reliably associated lower intensity speech with more politeness (i.e. making oneself sound smaller), they were variable in their interpretation of pitch. An analysis of participants found that listeners were split: some thought high-pitched utterances sounded more deferent, while others thought this of low-pitched utterances. The variability found in these studies could reflect competing motivations for the association of pitch with magnitude. Higher pitch can reflect smaller size or more forceful vocalization.

However, while the current findings related to vocalization frequency are consistent with these previous studies, some caution is warranted in interpreting them. While all the vocalizations could be distinctly characterized by their duration and intensity, the vocalizations did not always have a distinct, stable, and clearly evident fundamental frequency (i.e. periodic vibration of the vocal folds). In addition, frequency

information of the vocalizations – fundamental frequency and formant frequencies – might have been more difficult to detect in the playback across different listeners' computer speakers and headphones. Thus, under noisy conditions, listeners might have focused more on duration and loudness as reliable cues.

Taken together, our findings provide evidence for the hypothesis that different prothetic (i.e. more-or-less) properties including length, size, quantity, duration, and loudness – but not sound frequency – correspond via a generalized mental magnitude system, which represents the dimensions according to a common, amodal metric (Mondloch & Maurer, 2004a; Stevens, 1957; Walsh, 2003; Winter et al., 2015). Bigger and longer objects, larger quantities, and louder and longer sounds may correspond with each other because they are all at the "more" end of this scale. In contrast, the correspondence between magnitude and vocal frequency may derive from a different type of association, which could be subject to greater variability across contexts and cultures.

Our findings also have some implications for understanding the evolution of language. Some theorists have proposed that humans evolved the fine motor control needed for speech in order to better exploit listeners' expectations about the relationship between qualities of the voice and body size (Fitch, 2000; Pisanski, Cartei, et al., 2016). In the current experiments, we show that some of the same vocal qualities thought to underlie perceptions of body size – duration and intensity – are readily extended to refer to the magnitude of objects. Indeed, this communication was found to be effective across cultures and linguistic backgrounds and even sensory experience. Previous research has shown that nonlinguistic vocalizations can communicate information across disparate cultures, including, for example, basic emotions (Sauter et al., 2010) and the social

relationships between co-laughers (Bryant et al., 2016). Here, we show that non-linguistic vocalizations can serve effectively for communicating about inanimate referents. Thus, our ancestors may have been able to use these nearly universally recognized vocalization-magnitude correspondences to bootstrap the formation of mutually understood vocal conventions prior to the advent of spoken language.

5. Conclusion

The human ability to use vocalizations to communicate about the magnitude of objects is highly robust, extending across listeners of disparate linguistic and cultural backgrounds, as well as across age and even auditory experience. This ability is based on what appears to be a widely shared, deeply rooted sense of how to translate dimensions of magnitude into meaningful vocalizations. Our findings thus point to a robust common ground by which humans, faced with a challenge to communicate without a shared language, can generate meaningful vocalizations to communicate about magnitude. We find this to be a compelling demonstration of the human capacity to imbue vocalizations with meaning.

Authors' Contributions: M. Perlman, J. Paul and G. Lupyan devised the experiments. J. Paul conducted the vocal communication game experiment, and M. Perlman conducted the playback experiments. M. Perlman analyzed the data. M. Perlman, G. Lupyan, and J. Paul wrote the manuscript.

Competing Interests: We have no competing interests.

Funding: This work was funded in part by NSF-INSPIRE 1344279 and NSF-PAC 1734260 to GL.

Acknowledgments: We would like to thank all of the participants for taking part in the study. We thank Haiying Wen and Xulun Chen for their help with translation between spoken Mandarin and Standard Chinese Sign Language in data collection. Finally, we are grateful to the editor Julie Van Dyke and to the reviewers Greg Bryant and Katarzyna Pisanski for their helpful comments on the manuscript.

Author note: A preliminary analysis of the acoustic properties of the vocalizations in the production game was reported in the Proceedings of the 37th Annual Conference of the Cognitive Science Society. This report included analysis of the duration, intensity, and f_0 of vocalizations produced by deaf and hearing participants, but did not examine effects of age or conduct detailed formant analyses.

Bibliography

Bates, D, Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7. http://CRAN.R-project.org/package=lme4

Bates, Douglas, Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *ArXiv:1406.5823 [Stat]*. http://arxiv.org/abs/1406.5823

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818–10823. https://doi.org/10.1073/pnas.1605782113

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

Boothroyd, A. (1973). Some experiments on the control of voice in the profoundly deaf using a pitch extractor and storage oscilloscope display. *IEEE Transactions on Audio and Electroacoustics*, *21*(3), 274–278. https://doi.org/10.1109/TAU.1973.1162465

Bowling, D. L., Garcia, M., Dunn, J. C., Ruprecht, R., Stewart, A., Frommolt, K.-H., & Fitch, W. T. (2017). Body size and vocalization in primates and carnivores. *Scientific Reports*, *7*, 41070. https://doi.org/10.1038/srep41070

Bryant, G. A., Fessler, D. M. T., Fusaroli, R., Clint, E., Aarøe, L., Apicella, C. L., Petersen, M. B., Bickham, S. T., Bolyanatz, A., Chavez, B., Smet, D. D., Díaz, C., Fančovičová, J., Fux, M., Giraldo-Perez, P., Hu, A., Kamble, S. V., Kameda, T., Li, N. P., … Zhou, Y. (2016). Detecting affiliation in colaughter across 24 societies. *Proceedings of the National Academy of Sciences*, *113*(17), 4682–4687. https://doi.org/10.1073/pnas.1524993113

Clutton-Brock, T. H., & Albon, S. D. (1979). The Roaring of Red Deer and the Evolution of Honest Advertisement. *Behaviour*, *69*(3), 145–170. https://doi.org/10.1163/156853979X00449

Clutton-Brock, T. H., Harvey, P. H., & Rudder, B. (1977). Sexual dimorphism, socionomic sex ratio and body weight in primates. *Nature*, *269*(5631), 797–800. https://doi.org/10.1038/269797a0

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 6.1-12. https://doi.org/10.1167/10.1.6

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Sciences*, *4*(7), 258–267. https://doi.org/10.1016/S1364-6613(00)01494-7

Fitch, W. T. (2010). *The evolution of language*. Cambridge University Press.

Fitch, W. T., & Hauser, M. D. (2003). Unpacking "Honesty": Vertebrate Vocal Production and the Evolution of Acoustic Signals. In A. M. Simmons, R. R. Fay, & A. N. Popper (Eds.), *Acoustic Communication* (pp. 65–137). Springer New York. https://doi.org/10.1007/0-387-22762-8_3

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, *68*(7), 1191–1203. https://doi.org/10.3758/BF03193720

Gillooly, J. F., & Ophir, A. G. (2010). The energetic basis of acoustic communication. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20092134. https://doi.org/10.1098/rspb.2009.2134

Gussenhoven, C. (2002). Intonation and Interpretation: Phonetics and Phonology. In *Speech Prosody 2002: Proceedings of the First International Conference on Speech Prosody.* (pp. 47–57).

Hamilton-Fletcher, G., Pisanski, K., Reby, D., Stefańczyk, M., Ward, J., & Sorokowska, A. (2018). The role of visual experience in the emergence of cross-modal correspondences. *Cognition*, *175*, 114–121. https://doi.org/10.1016/j.cognition.2018.02.023

Haynie, H., Bowern, C., & LaPalombara, H. (2014). Sound Symbolism in the Languages of Australia. *PLOS ONE*, *9*(4), e92852. https://doi.org/10.1371/journal.pone.0092852

Hewitt, G., MacLarnon, A., & Jones, K. E. (2002). The functions of laryngeal air sacs in primates: A new hypothesis. *Folia Primatologica; International Journal of Primatology*, *73*(2–3), 70–94. https://doi.org/10.1159/000064786

Idemaru, K., Winter, B., Brown, L., & Oh, G. E. (2019). Loudness Trumps Pitch in Politeness Judgments: Evidence from Korean Deferential Speech. *Language and Speech*, 002383091882434. https://doi.org/10.1177/0023830918824344

Künzel, H. J. (1989). How Well Does Average Fundamental Frequency Correlate with Speaker Height and Weight? *Phonetica*, *46*(1–3), 117–125. https://doi.org/10.1159/000261832

Modig, A. O. (1996). Effects of body size and harem size on male reproductive behaviour in the southern elephant seal. *Animal Behaviour*, *51*(6), 1295–1306. https://doi.org/10.1006/anbe.1996.0134

Mondloch, C. J., & Maurer, D. (2004a). Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(2), 133–136.

Mondloch, C. J., & Maurer, D. (2004b). Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(2), 133–136. https://doi.org/10.3758/CABN.4.2.133

Morton, E. S. (1977). On the occurence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, *111*(981), 855–869.

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The Semantics of Prosody: Acoustic and Perceptual Evidence of Prosodic Correlates to Word Meaning. *Cognitive Science*, *33*(1), 127–146. https://doi.org/10.1111/j.1551-6709.2008.01007.x

Ohala, J. J. (1983). Cross-Language Use of Pitch: An Ethological View. *Phonetica*, *40*(1), 1–18. https://doi.org/10.1159/000261678

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In *Sound Symbolism* (pp. 325–347). Cambridge University Press.

Perlman, M., & Cain, A. A. (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture*, *14*(3), 320–350. https://doi.org/10.1075/gest.14.3.03per

Perlman, M., Clark, N., & Johansson Falck, M. (2015). Iconic Prosody in Story Reading. *Cognitive Science*, *39*(6), 1348–1368. https://doi.org/10.1111/cogs.12190

Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society Open Science*, *2*(8), 150152. https://doi.org/10.1098/rsos.150152

Perlman, M., & Lupyan, G. (2018). People Can Create Iconic Vocalizations to Communicate Various Meanings to Naïve Listeners. *Scientific Reports*, *8*(1), 2634. https://doi.org/10.1038/s41598-018-20961-6

Pietraszewski, D., Wertz, A. E., Bryant, G. A., & Wynn, K. (2017). Three-month-old human infants use vocal cues of body size. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1856), 20170656. https://doi.org/10.1098/rspb.2017.0656

Pisanski, K., & Bryant, G. A. (2019). The evolution of voice perception. In N. S. Eidsheim & K. L. Meizel (Eds.), *The Oxford Handbook of Voice Studies* (pp. 269–300). Oxford University Press.

Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice Modulation: A Window into the Origins of Human Vocal Control? *Trends in Cognitive Sciences*, *20*(4), 304–318. https://doi.org/10.1016/j.tics.2016.01.002

Pisanski, K., Feinberg, D., Oleszkiewicz, A., & Sorokowska, A. (2017). Voice cues are used in a similar way by blind and sighted adults when assessing women's body size. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-10470-3

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., Fink, B., DeBruine, L. M., Jones, B. C., & Feinberg, D. R. (2014). Vocal indicators

of body size in men and women: A meta-analysis. *Animal Behaviour*, *95*, 89–99. https://doi.org/10.1016/j.anbehav.2014.06.011

Pisanski, K., Mora, E. C., Pisanski, A., Reby, D., Sorokowski, P., Frackowiak, T., & Feinberg, D. R. (2016). Volitional exaggeration of body size through fundamental and formant frequency modulation in humans. *Scientific Reports*, *6*, 34389. https://doi.org/10.1038/srep34389

Pisanski, K., Oleszkiewicz, A., & Sorokowska, A. (2016). Can blind persons accurately assess body size from the voice? *Biology Letters*, *12*(4). https://doi.org/10.1098/rsbl.2016.0063

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Reby, D., & McComb, K. (2003). Anatomical constraints generate honesty: Acoustic cues to age and weight in the roars of red deer stags. *Animal Behaviour*, *65*(3), 519–530. https://doi.org/10.1006/anbe.2003.2078

Rendall, D., Kollias, S., Ney, C., & Lloyd, P. (2005). Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: The role of vocalizer body size and voice-acoustic allometry. *The Journal of the Acoustical Society of America*, *117*(2), 944–955. https://doi.org/10.1121/1.1848011

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, *12*(3), 225–239. https://doi.org/10.1037/h0070931

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of*

the National Academy of Sciences, *107*(6), 2408–2412. https://doi.org/10.1073/pnas.0908239106

Schuett, G. W. (1997). Body size and agonistic experience affect dominance and mating success in male copperheads. *Animal Behaviour*, *54*(1), 213–224. https://doi.org/10.1006/anbe.1996.0417

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Taylor, A. M., & Reby, D. (2010). The contribution of source–filter theory to mammal vocal communication research. *Journal of Zoology*, *280*(3), 221–236. https://doi.org/10.1111/j.1469-7998.2009.00661.x

Thompson, P. D., & Estes, Z. (2011). Sound symbolic naming of novel objects is a graded function. *The Quarterly Journal of Experimental Psychology*, *64*(12), 2392–2404. https://doi.org/10.1080/17470218.2011.605898

Titze, I. R. (1994). *Principles of Voice Production* (Facsimile edition). Allyn & Bacon.

Ultan, R. (1978). Size-sound symbolism. In *Universals of Human Language*.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488. https://doi.org/10.1016/j.tics.2003.09.002

Winter, B., Marghetis, T., & Matlock, T. (2015). Of magnitudes and metaphors: Explaining cognitive interactions between space, time, and number. *Cortex*, *64*, 209–224. https://doi.org/10.1016/j.cortex.2014.10.015