COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12883

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12883

Stimulus Parameters Underlying Sound-Symbolic Mapping of Auditory Pseudowords to Visual Shapes

Simon Lacey,[a,b,c] Yaseen Jamal,[d] Sara M. List,[c,d] Kelly McCormick,[c,d] K. Sathian,[a,b,c,d,e] Lynne C. Nygaard[d]

[a]*Department of Neurology, Milton S. Hershey Medical Center, Penn State College of Medicine*
[b]*Department of Neural & Behavioral Sciences, Milton S. Hershey Medical Center, Penn State College of Medicine*
[c]*Department of Neurology, Emory University*
[d]*Department of Psychology, Emory University*
[e]*Department of Psychology, Milton S. Hershey Medical Center, Penn State College of Medicine*

Received 21 December 2018; received in revised form 6 June 2020; accepted 1 July 2020

Abstract

Sound symbolism refers to non-arbitrary mappings between the sounds of words and their meanings and is often studied by pairing auditory pseudowords such as “maluma” and “takete” with rounded and pointed visual shapes, respectively. However, it is unclear what auditory properties of pseudowords contribute to their perception as rounded or pointed. Here, we compared perceptual ratings of the roundedness/pointedness of large sets of pseudowords and shapes to their acoustic and visual properties using a novel application of representational similarity analysis (RSA). Representational dissimilarity matrices (RDMs) of the auditory and visual ratings of roundedness/pointedness were significantly correlated crossmodally. The auditory perceptual RDM correlated significantly with RDMs of spectral tilt, the temporal fast Fourier transform (FFT), and the speech envelope. Conventional correlational analyses showed that ratings of pseudowords transitioned from rounded to pointed as vocal roughness (as measured by the harmonics-to-noise ratio, pulse number, fraction of unvoiced frames, mean autocorrelation, shimmer, and jitter) increased. The visual perceptual RDM correlated significantly with RDMs of global indices of visual shape (the simple matching coefficient, image silhouette, image outlines, and Jaccard distance). Crossmodally, the RDMs of the auditory spectral parameters correlated weakly but significantly with those of the global indices of visual shape. Our work establishes the utility of RSA for analysis of large stimulus sets and offers novel insights into the stimulus parameters underlying sound symbolism, showing that sound-to-shape mapping is driven by acoustic properties of pseudowords and

Correspondence should be sent to Lynne C. Nygaard, Department of Psychology, College of Arts and Sciences, Emory University, Atlanta, GA 30322. E-mail: lnygaar@emory.edu; K. Sathian, Department of Neurology, Milton S. Hershey Medical Center, Penn State College of Medicine, Hershey, PA 17033-0859. E-mail: ksathian@pennstatehealth.psu.edu

suggesting audiovisual cross-modal correspondence as a basis for language users' sensitivity to this type of sound symbolism.

Keywords: Language; Multisensory; Representational similarity analysis; Sound symbolism

1. Introduction

It is commonly held that arbitrariness is a fundamental property of language, that is, that the sound structure of a word bears no relation to the thing it describes (de Saussure, 2011; but see Joseph, 2015). Whether this is always the case or whether natural relationships between sound and meaning exist in natural language has been debated since at least the Platonic dialog of *Cratylus* (Ademollo, 2011). One aspect of language that is non-arbitrary is sound symbolism (Perniss & Vigliocco, 2014), which includes a broad set of phenomena in which there is a perceived resemblance between speech sounds and their referents. An example is onomatopoeia, in which the sound of a word resembles the sound it represents (Catricalà & Guidi, 2015; Schmidtke, Conrad, & Jacobs, 2014), for example, "slap" or "splash," and mimetic words in Japanese, for example, "kirakira" (flickering light: Akita & Tsujimura, 2016). It is important to note, however, that while sound symbolism may be contrasted with arbitrariness, these are not mutually exclusive and may exist alongside one another in natural language (Lockwood & Dingemanse, 2015).

Sound symbolism often involves examples of crossmodal correspondence, that is, the near-universally experienced associations between seemingly arbitrary stimulus features in different senses (Spence, 2011). For example, high and low auditory pitch are consistently associated with small and large visual size (Evans & Treisman, 2010; Gallace & Spence, 2006), and with high and low visuospatial elevation, respectively (Ben-Artzi & Marks, 1995; Jamal, Lacey, Nygaard, & Sathian, 2017; Lacey, Martinez, McCormick, & Sathian, 2016). A well-known example of sound-symbolic crossmodal correspondence was first described by Köhler (1929, 1947) in which individuals consistently assigned the pseudoword "maluma" to a curvy, cloud-like shape and the pseudoword "takete" to an angular, star-like shape. Such crossmodal sound-symbolic associations occur not only for pseudowords but also for real words, for example, "balloon" and "spike" for rounded and pointed shapes (Sučević, Savić, Popović, Styles, & Ković, 2015).

Since Köhler's early work, sound symbolism has been demonstrated across different languages (Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016), with both similarities (e.g., Davis, 1961) and differences (e.g., Bremner et al., 2013; Rogers & Ross, 1975; Styles & Gawne, 2017) between Western and non-Western cultures. These studies show that the existence of sound symbolism in language is both prolific and robust. Furthermore, language users are sensitive to sound-symbolic associations in that they can correctly assign meaning to synonym–antonym pairs in an unfamiliar foreign language at above-chance levels (Nygaard, Cook, & Namy, 2009; Revill, Namy, Defife, & Nygaard, 2014; Tzeng, Nygaard, & Namy, 2016). Sound symbolism may also play a role in language processing and early word learning (Imai & Kita, 2014). For example, children of pre-reading age exhibit sensitivity to sound-symbolic crossmodal associations (Imai et al., 2015; Maurer, Pathman, & Mondloch, 2006; Ozturk, Krehm, & Vouloumanos, 2013), and recent studies have suggested that sound symbolism is important for specific word-to-meaning associations in young children with limited vocabularies (Gasser, 2004; Tzeng, Nygaard, & Namy, 2017). In adults, sound symbolism may offer linguistic processing advantages for categorization and word learning (Brand, Monaghan, & Walker, 2018; Gasser, 2004; Revill, Namy, & Nygaard, 2018), and for rehabilitation of patients with aphasia (Meteyard, Stoppard, Snudden, Cappa, & Vigliocco, 2015). More recently, neuroimaging studies have begun to reveal the neural correlates of sound symbolism (McCormick, Lacey, Stilla, Nygaard, & Sathian, 2018; Peiffer-Smadja & Cohen, 2019; Revill et al., 2014).

However, it remains an open question whether sound-symbolic correspondences are essentially based in auditory features of words and, if so, what auditory features are mapped onto which visual (or other) features of the referent that it sound-symbolically describes (and vice-versa). For sound-to-shape mapping, research has largely studied the rounded/pointed dimension, mostly concentrating on *phonological* features, for example, consonants versus vowels (Fort, Martin, & Peperkamp, 2015; Nielsen & Rendall, 2011), voiced versus unvoiced consonants (Cuskley, Simner, & Kirby, 2017; McCormick, Kim, List, & Nygaard, 2015), rounded versus unrounded vowels (Maurer et al., 2006; McCormick et al., 2015), obstruents versus sonorants (McCormick et al., 2015), or vowel formants[1] (Knoeferle, Li, Maggioni, & Spence, 2017). These phonemic feature differences are important: Styles and Gawne (2017) suggest that failures to replicate sound-to-shape mapping cross-culturally occur because the chosen pseudowords did not conform to the sound structure of the language spoken in the target culture. But while different phonemic categories, for example, consonants versus vowels or obstruents versus sonorants, have different acoustic properties, few studies have measured those properties directly to assess their contribution to sound symbolism. Of 21 studies of sound symbolism relating to roundedness/pointedness listed by Westbury, Hollis, Sidhu, and Pexman (2018, Table 1), only Monaghan, Mattock, and Walker (2012) and Ozturk et al. (2013) measured acoustic properties (of frequency, amplitude, and duration) and only to confirm whether word groups differed on these rather than to examine their contributions. Nonetheless, such acoustic differences may be important. A notable exception to the list in Westbury et al. (2018) is the study of Parise and Pavani (2011), who measured participants' vocalizations of a single vowel sound in response to the shape, luminance, or size of visual stimuli. These vocalizations were louder for complex (dodecahedron) compared to simple (triangle) shapes, and for brighter than darker stimuli, while the frequency of F3 was higher for triangles, a shape that is perhaps more obviously pointed than a dodecahedron. Subsequently, Knoeferle et al. (2017) showed that the frequencies of F2 and F3 are related to perceptual ratings of roundedness/pointedness. Pseudowords with lower/higher F2 were rated as more rounded/pointed, respectively, while roundedness ratings increased with higher values of F3, which reflects the amount of articulatory lip-rounding (Knoeferle et al., 2017). Note that there is a distinction to be made between how a speaker produces an utterance that is, any word can be spoken with higher or lower pitch, and the acoustic characteristics of particular phonemic elements that is, rounded vowels will always have a higher F3 compared to unrounded vowels.

Similarly, few studies have measured the properties of the visual shapes employed to study sound-symbolic crossmodal correspondences. To our knowledge, the sole exception is the cross-cultural study of Chen, Huang, Woods, and Spence (2016), who created shapes using radial frequency patterns: parametric sinusoidal modulations around the circumference of a circle that varied in frequency (number of modulations per unit length of the circumference), amplitude (magnitude of modulation), and “spikiness” (magnitude of a triangular wave function added to the sinusoid). Interestingly, while all three factors predicted sound-to-shape mapping regardless of culture, North Americans weighted amplitude more heavily than “spikiness,” but the reverse was true for Taiwanese participants, perhaps reflecting cultural preferences for analytic and holistic processing, respectively (Chen et al., 2016). Although some studies have addressed the visual shape effects of orthography (Cuskley et al., 2017) or typography (De Carolis, Marsico, Arnaud, & Coupé, 2018), these factors are obviously less relevant when pseudowords are presented auditorily.

Here, we investigate both acoustic and visual parameters of large sets of pseudowords (537) and visual shapes (90), respectively, in relation to perceptual ratings of their roundedness/pointedness. To do this, we employ a method novel to studies of sound symbolism: representational similarity analysis (RSA). RSA was originally developed as a method for analyzing functional magnetic resonance imaging (fMRI) data and has also been applied to various kinds of neurophysiological data (Kriegeskorte et al., 2008). In the context of fMRI, RSA compares the pairwise spatial distribution of activity for stimuli across voxels. This spatial pattern should be similar for stimulus pairs that are similar in some respect, for example, leopards and cheetahs, but dissimilar for stimulus pairs that are not, for example, leopards and polar bears (both mammalian quadrupeds but differing in size, appearance, taxonomy, and habitat). Computationally, the activity levels for each stimulus are vectorized and the first-order pairwise correlation is calculated: Similar pairs should be positively correlated and dissimilar pairs should be negatively correlated. Operationally, the results are displayed as a representational dissimilarity matrix (RDM) in which each cell value is $1 - r$: For very similar, highly positively correlated pairs, this value should approach 0 ($1 - 1$ = minimum dissimilarity); for very dissimilar, highly negatively correlated pairs, this value should approach 2 ($1 - (-1)$ = maximum dissimilarity). Such RDMs can then be compared, via second-order correlations, to reference RDMs based on, for example, (dis)similarity in habitat or taxonomy, as in our animal categorization example, or formal computational models, to test hypotheses about how information is organized in a particular brain region. Our approach here was to calculate RDMs for pseudowords based on ratings of their perceived roundedness/pointedness. If both members of a pair of pseudowords are considered rounded, these ratings will be more or less positively correlated; but for a pair containing a rounded and a pointed pseudoword, the ratings will be more or less negatively correlated, reflecting the degree of similarity or dissimilarity, respectively. Similarly, we could compute RDMs based on measurements of acoustic properties of the pseudowords (see below) and compare these to the perceptual ratings by way of second-order correlations between the perceptual and acoustic RDMs. To the extent that perception of the pseudowords as rounded/pointed correlates with an acoustic property, that property can be said to contribute to the sound-symbolic mapping of sound to shape. The same computations and principles apply to ratings and measurements of visual shapes. The advantages of the RSA approach over conventional correlational analyses are that, first, it allows analysis of stimulus properties that involve multiple measurements or samples per stimulus; second, RSA compares every item to every other item so that it analyzes similarity across and between all possible stimulus pairs. Conventional correlations only allow for an assessment of the association of a single acoustic measure with perceptual ratings and only considers pairs on a list-wise basis rather than examining all possible pairs. Thus, RSA allowed us to evaluate whether similarity across and between stimuli for a particular acoustic characteristic mirrored perceptual similarity across and between stimuli.

For the pseudowords, we chose acoustic parameters that would reflect the overall acoustic form of each word, capturing both the acoustic properties associated with phonemic content and aspects of the vocal characteristics of the speaker. We did so because we reasoned that the rating of a pseudoword as rounded or pointed could depend on the acoustic characteristics resulting from the phonemic content of the particular word (e.g., voicing or manner of articulation of a phonetic segment) and/or the vocal properties of the speaker's voice. For example, Tzeng et al. (2017) found that speakers produced pseudowords referring to bright colors with higher fundamental frequency and amplitude and shorter duration than those for darker colors and that listeners could reliably assign pseudowords to their target color using these prosodic cues. In other words, the acoustic-phonetic instantiation of spoken language depends on both *what* a speaker says and *how* they say it (Nygaard, Herold, & Namy, 2009). These two factors may not be easily separable but, in general terms depending on the measure, both contribute to the acoustic form of the speech signal. As such, we chose three parameters, speech envelope, spectral tilt, and the temporal fast Fourier transform (FFT), that captured the distribution of amplitude and frequency over time. In addition, we chose parameters that reflect the acoustic consequences of voicing, or the extent to which the vocal folds vibrate creating a voiced or periodic signal. Each measure reflected the amount and regularity of voicing as reflected in periodicity in the speech waveform: the fraction of unvoiced frames (FUF), mean autocorrelation, pulse number, jitter, shimmer, the standard deviation of the fundamental frequency, and the mean harmonics-to-noise ratio (HNR). Interestingly, recent work suggests that simple acoustic features such as the amplitude envelope are sufficient to decode cortical responses to speech (Daube, Ince, & Gross, 2019). Full details of the acoustic parameters are provided in Section 2.3.1. Briefly, we expected that parameters that captured low- and high-frequency information would reflect roundedness and pointedness, respectively, as has been demonstrated with low- and high-pitched auditory tones, that is, non-linguistic stimuli (e.g., Marks, 1987; Walker et al., 2010). In contrast, we expected that parameters capturing spectrotemporal aspects of the speech waveform would reflect roundedness and pointedness to the extent that the waveform reflected a speech pattern that was smooth and continuous as opposed to one that was uneven or contained abrupt transitions, as has also been demonstrated in non-linguistic contexts using sinusoidal and square waveforms (Parise & Spence, 2012) and by varying the "roughness" of electronically produced auditory noise (Liew, Lindborg, Rodrigues, & Styles, 2018).

In choosing visual parameters for the shapes, we were somewhat constrained by the fact that the shapes were all irregular; thus, it would not be possible to employ radial frequency measures (Wilkinson, Wilson, & Habak, 1998). We chose the Jaccard distance and the simple matching coefficient (SMC) which essentially measure pairwise shape similarity by the amount of overlap when the shapes are superimposed, together with image silhouette and outline which code object shape either taking account of area (silhouette) or independently of area (outline). Full details of the visual parameters are provided in Section 2.3.2; for all these measures we expected that, as the shapes transitioned from rounded to pointed, there would be a graded transition from positive to negative correlation as dissimilarity increased.

To assess whether an acoustic parameter contributed to perception of a pseudoword as rounded or pointed, we compared the RDM for each parameter to that for the auditory perceptual ratings: A significant correlation between the RDMs would indicate that the parameter influenced the mapping of sound to shape. We carried out the same analysis for visual parameters and perceptual ratings of the shapes. Although ratings of the shapes could not directly influence ratings of the pseudowords (auditory and visual ratings were provided by separate groups: see Methods), we also compared acoustic and visual parameters crossmodally. To the extent that an acoustic parameter was significantly related to roundedness/pointedness ratings of the pseudowords and *also*, crossmodally to a visual parameter that significantly captured visual roundedness/pointedness, this provided a supplementary confirmation of the acoustic parameter's relevance to sound-to-shape mapping. We follow this acoustic-visual-crossmodal sequence in Sections 2, 3, and 4.

2. Materials and methods

2.1. Perceptual ratings

The RSA described in Section 2.2 is based on perceptual ratings of auditory pseudowords collected by McCormick et al. (2015), who also created and recorded these stimuli, and ratings of visual shapes created and collected by McCormick and Nygaard (unpublished data). The rest of Section 2.1 summarizes the methods for the creation and rating of these two data sets.

2.1.1. Participants

A total of 61 Emory University students (28 males, 33 females; $M_{\text{age}} \pm SD$, 20 ± 4 years) gave informed consent and received course credit for their participation. In all, 30 participated in the rating task for visual shapes (14 males and 16 females) and a separate 31 participated in the rating task for auditory pseudowords (14 males and 17 females). All participants were native English (American) speakers and reported normal or corrected-to-normal vision and no known hearing, speech, or language disorders. All procedures were approved by the Emory University Institutional Review Board.

2.1.2. Auditory pseudowords

We used a set of 537 two-syllable pseudowords of the form "consonant, vowel, consonant, vowel" (CVCV) devised by McCormick et al. (2015). These were constructed using only phonemes and combinations of phonemes that occur in the English language, and items deemed to be homophones of real words (33 items out of an original array of 570) were removed. Consonants were sampled from sonorants, fricatives/affricates, and stops; of the obstruents, including fricatives/affricates and stops, half were voiced and half were unvoiced. Vowels were either front/rounded or back/unrounded. The pseudowords were recorded in random order by a female native speaker of American English (K.M.) with neutral intonation in a sound-attenuated room, using a Zoom 2 Cardioid microphone, and digitized at a 44.1 kHz sampling rate. Two independent judges listened to the recordings to assess whether each pseudoword was recorded with neutral intonation, sounded consistent with other recordings (e.g., the pseudoword was not spoken faster/slower or louder than others), and conformed to the target phonemic content. For those pseudowords where the judges agreed that the token did not conform on any aspect, that item was re-recorded and judged again. Items were also re-recorded if the two judges disagreed on any aspect. A total of 54 pseudowords were re-recorded and re-assessed, if necessary multiple times, before being considered acceptable. Each pseudoword was then down-sampled at 22.05 kHz, which is standard for speech, and amplitude-normalized using PRAAT speech analysis software (Boersma & Weenink, 2012). The pseudowords had a mean duration of 457 ± 62 ms. Briefly, McCormick et al. (2015) showed that judgments of "roundedness" for this set of stimuli were more associated with voiced (e.g., /b/, /d/) than unvoiced (e.g., /t/, /k/) consonants, and with back rounded vowels like /u/ or /o/. Judgments of "pointedness" were more associated with stops like /p/ and /t/ than sonorants like /m/ or /l/, and with front unrounded vowels like /i/ or /e/. It is important to note, however, that the graded nature of the ratings suggests that judgments were based on more than individual phonetic features and likely involved processing/analysis at the segment or even whole-word level (McCormick et al., 2015; see also Thompson & Estes, 2011). For a complete description of the stimulus set, see McCormick et al. (2015).

2.1.3. Visual shapes

We used 90 shapes, consisting of gray line drawings (RGB: 240, 240, 240) on a white background, created in Adobe Illustrator (Ventura, CA: McCormick et al., unpublished data; McCormick et al., 2018: see Fig. 1 for example) following a method similar to that of Monaghan et al. (2012). Shapes had four, five, or six protuberances and were constructed using a template of three concentric circles (25, 35, and 45 mm radii), the outer circle serving as a bounding border with protuberances, either rounded or pointed, extending to its perimeter. The two inner circles served to define the inward extent of each protuberance. Thinner protuberances (30 shapes) extended all the way to the innermost

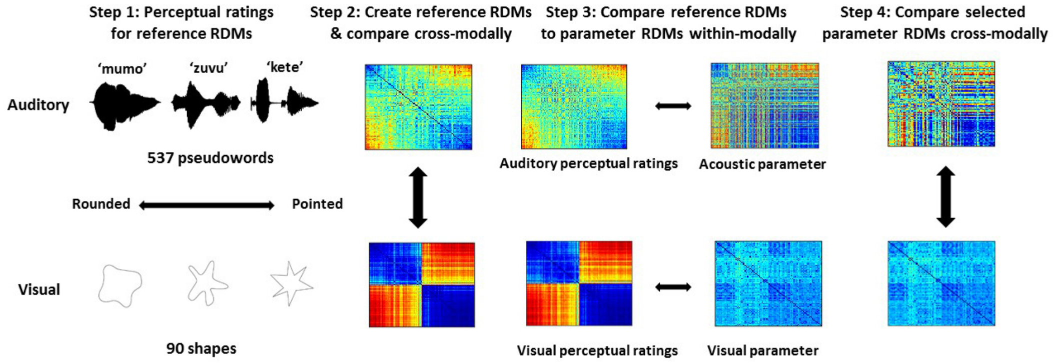
Step 1: Perceptual ratings for reference RDMs
Step 2: Create reference RDMs & compare cross-modally
Step 3: Compare reference RDMs to parameter RDMs within-modally
Step 4: Compare selected parameter RDMs cross-modally
Auditory
'mumo'
'zuvu'
'kete'
537 pseudowords
Rounded
Pointed
Visual
90 shapes
Auditory perceptual ratings
Acoustic parameter
Visual perceptual ratings
Visual parameter

Fig. 1. Analysis pipeline. Step 1: perceptual ratings of roundedness/pointedness for pseudowords and shapes were used to create reference representational dissimilarity matrices (RDMs). Step 2: crossmodal comparison of RDMs for perceptual ratings of pseudowords and shapes. Step 3: within-modal comparison of RDMs for perceptual ratings to those for acoustic and visual parameters of pseudowords and shapes, respectively. Step 4: crossmodal comparison of RDMs for selected acoustic and visual parameters.

circle; thicker protuberances (30 shapes) extended only to the middle circle; the remaining 30 shapes were constructed with a mix between thin and thick protuberances. For each shape of one category (rounded or pointed), there was a corresponding shape in the other category with the same outer and inner anchor points, resulting in 15 thick, 15 thin, and 15 mixed shapes in each category.

2.1.4. Perceptual rating tasks

Participants were randomly assigned to rate either pseudowords or shapes using one of two 7-point Likert-type scales. To avoid response bias, one of the scales rated roundedness from 1 (not rounded) to 7 (very rounded) and the other rated pointedness from 1 (not pointed) to 7 (very pointed). For pseudowords, 15 participants used the roundedness scale and 16 the pointedness scale (n = 31). To discourage participants from matching pseudowords with a specific word in the instructions (e.g., “teti” and “pointed”), the instructions included several related terms for the concepts of rounded and pointed. For the shapes, 17 participants used the roundedness scale and 13 the pointedness scale (n = 30).

The auditory pseudowords were presented over Beyerdynamic DT100 headphones at approximately 75 db SPL. The visual shapes were presented sequentially at the center of a desktop computer screen using E-Prime software Version 2.0.8.22 (Schneider, Eschman, & Zuccolotto, 2002). For both pseudowords and shapes, the 7-point rating scale appeared on the screen on each trial, either in the center of the screen for pseudowords or below each shape. The response keyboard always had 1–7 listed from left to right. All stimuli were presented only once and in random order.

2.2. Representational similarity analysis

We implemented RSA in MATLAB 2016a (The MathWorks, Natick, MA). In outline, we created reference RDMs for the pseudowords and shapes from the perceptual ratings of their roundedness and pointedness. We then compared these, via second-order correlations, both to each other and to RDMs derived from measurements of selected acoustic and visual parameters (see Section 2.3 for details of these) to assess how these parameters related to perception of roundedness and pointedness. We performed this latter step both within-modally (e.g., comparing perceptual ratings of the visual shapes to visual parameters) and crossmodally for selected parameters (i.e., comparing visual parameters to acoustic parameters). A schematic of the analysis pipeline is shown in Fig. 1, and we describe each step in more detail below.

As a first step, we created reference RDMs for pseudowords and shapes based on the perceptual ratings of their roundedness and pointedness. In these matrices, items were ordered left to right from the most rounded to the most pointed based on the mean rating for each item. To achieve this, one of the two rating scales was recoded so that 1 was equal to "not pointed" on one scale and "very rounded" on the other, and 7 was equal to "very pointed" on the first scale and "not rounded" on the second; that is, the "roundedness" scale was recoded to the "pointedness" scale. Since the two scales are in opposition to each other, they should be strongly negatively correlated and this was, in fact, the case (pseudowords $r_{535} = -.65$, $p < .001$; shapes $r_{88} = -.96$, $p < .001$: Fig. 2). Thus, the two scales were comparable and recoding to a single scale was justified. The correlation was stronger for shapes than for pseudowords, presumably reflecting that visual ratings can directly assess visual roundedness/pointedness, whereas the pseudowords were rated for a property that is primarily visual and therefore auditory roundedness/pointedness ratings access this indirectly. (With the exception of the pseudoword pointed scale, ratings were non-normally distributed, but testing these relationships with the non-parametric Spearman correlation produced the same pattern of results.)

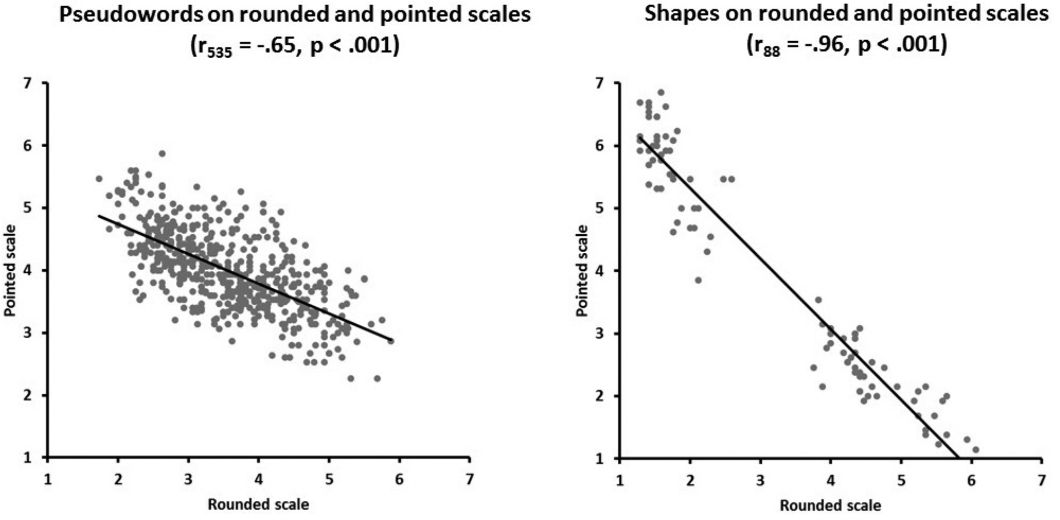

Fig. 2. The roundedness and pointedness scales were strongly negatively correlated, indicating that ratings of pseudowords (left) and shapes (right) were comparable across the two scales.

Once the pseudowords and shapes had been ordered in this way, the RDMs were constructed using the original un-recoded data since the RDMs reflected how the patterns of ratings were dissimilar across items regardless of the rating scale that any individual participant used. To create the reference RDMs (Fig. 1, Step 1), we calculated the first-order correlation (Pearson's r) between the perceptual ratings for each pair of pseudowords or shapes: Pairwise dissimilarity is given by $1 - r$ and this is the value entered in each cell of the RDM. Having created these reference RDMs, we could compare them to each other, via a second-order, non-parametric correlation (Spearman's r [r_s]: Fig. 1, Step 2),[2] to assess the extent to which the perceptual rating matrices were crossmodally consistent.

The next stage was to create RDMs reflecting the pairwise dissimilarity for acoustic parameters of the pseudowords and visual parameters of the shapes and to compare these to the RDMs for the auditory and visual perceptual ratings, respectively (Fig. 1, Step 3). These second-order correlations would enable us to see, for example, which acoustic parameters might contribute to perception of the pseudowords as rounded or pointed. Full details of the acoustic and visual parameters and the calculation of their RDMs are provided in Section 2.3.

Finally, to the extent that RDMs for the acoustic and visual parameters were significantly correlated with those for auditory and visual perceptual ratings, respectively, we could compare the RDMs of those parameters crossmodally (Fig. 1, Step 4). This comparison served two purposes. First, because the auditory and visual perceptual ratings were carried out by independent groups of participants, it was possible that each group judged pointedness and roundness on a different basis. If this were so, the RDMs of the acoustic and visual parameters would not necessarily be correlated with each other crossmodally. But if both groups were employing a common perceptual framework in the rating tasks regardless of modality, then the RDMs of parameters that correlated with the RDMs of perceptual ratings should also be crossmodally correlated. Second, and relevant to the study aims, in this data-driven approach a further test of whether an acoustic parameter is a likely candidate to drive sound-symbolic mapping of sound to shape would be that not only is its RDM correlated with the RDM for perceptual ratings of the pseudowords as rounded/pointed, but also crossmodally with the RDM for a visual parameter that predicts perceptual ratings of roundedness/pointedness for the shapes. Note that the reference RDM for pseudowords is a 537 × 537 matrix while that for shapes is 90 × 90. To perform the crossmodal second-order correlation the matrices must be the same size and therefore we down-sampled the pseudoword matrix by selecting every sixth word to create a 90 × 90 matrix (the number of samples per item remained unchanged, i.e., 31 rating scores).

2.3. Stimulus parameters

2.3.1. Acoustic parameters of pseudowords

As noted in the Introduction, we chose acoustic parameters that would reflect the overall acoustic form of each pseudoword, capturing the acoustic consequences of both their phonemic content and aspects of the vocal characteristics of the speaker. Therefore, we chose the speech envelope, spectral tilt, and temporal FFT, since these capture the distribution of both amplitude and frequency over time. Additionally, we chose parameters that reflect the proportion and regularity of voicing as reflected in acoustic periodicity in the speech waveform. The FUF, mean autocorrelation, and pulse number reflect the proportion of voiced segments, and the remaining parameters were chosen to reflect the regularity of voicing or voice quality during the production of each stimulus item: jitter, shimmer, the standard deviation of the fundamental frequency (the speech analysis software PRAAT [Boersma & Weenink, 2012] refers to this as “pitch standard deviation (PSD)” and we adopt this term here), and the mean HNR. Each parameter is described in detail below. Note that these parameters are not necessarily independent of each other (e.g., both FUF and pulse number reflect how often the vocal folds open and close). Additionally, although some parameters are most often studied in the context of voice pathology, the pseudowords were recorded by a speaker with a healthy voice and these parameters can certainly vary in a healthy voice (see Brockmann, Drinnan, Storck, & Carding, 2011). It should also be noted that, given the nature of each particular parameter or property, the number of measurements used to calculate the first-order correlation differed across the acoustic parameters.

For speech envelope, spectral tilt, and temporal FFT, we normalized the duration of all pseudowords to the mean of 457 ms, by removing and interpolating data points from longer and shorter items, respectively, using the resampling function in MATLAB. Although this was a necessary step to achieve common vector lengths for parameter estimation, it necessarily introduced some noise; however, since we resampled to the mean duration, the introduced noise would be proportional in magnitude to the standard deviation of the duration, which was small (standard deviation/mean = 62/457 ms, i.e., 13.5%). At the original sampling rate of 22,050 Hz (see Section 2.1.2), this gave a vector of 10,077 data points (22,050 × 0.457 = 10,077). This enabled us to obtain a common vector length for calculating these parameters across pseudowords that varied in duration and therefore equal numbers of data points per pseudoword for the pairwise correlations that form the RDMs. However, for these parameters, measurements were taken from that vector in different ways, for example, different window lengths, such that the number of measurements underlying the pairwise correlations for each of these parameter differed (see Supplementary Material). Speech envelope, spectral tilt, and temporal FFT were calculated in MATLAB 2016a while the remaining acoustic parameters were measured using the standard voice report settings in PRAAT (Boersma & Weenink, 2012); the RDMs were prepared using MATLAB (2016a). Speech envelope, spectral tilt, and the temporal FFT were all based on multiple samples for each pseudoword and therefore pairwise first-order correlations could be calculated at the item level resulting in a 537 × 537 matrix that could be compared directly to the 537 × 537 perceptual matrix. However, all the other acoustic parameters were expressed as a single value per pseudoword and, therefore, to compute the first-order correlations, these single values were binned into an 18 × 18 matrix with 30 pseudowords per cell (comparable to the 31 participants who provided the perceptual ratings). For the second-order correlations for these parameters, the RDM for the perceptual ratings was similarly created by binning the mean rating for each pseudoword into an 18 × 18 matrix, also with 30 pseudowords per cell.

Speech envelope: This is a measure of the amplitude profile across time which primarily reflects changes corresponding to phonemic properties and syllabic transitions (Aiken & Picton, 2008). A visual depiction of the speech envelope can capture the "shape" of the sound by showing these transitions. To the extent that transitions are abrupt, the amplitude profile will appear uneven or jagged, which should be associated with pointed pseudowords, and to the extent that they are more gradual, the profile will appear smoother and more continuous, which should be associated with rounded pseudowords. This expectation is similar to the study of Thoret, Aramaki, Kronland-Martinet, Velay, and Ystad (2014) which showed that participants could retrieve visual shape from the friction sounds produced when a shape was drawn; compare also, for example, the left and right panels of Fig. 5C, which display speech envelopes for rounded and pointed pseudowords, respectively.

Spectral tilt: This gives an estimate of the overall slope of the power spectrum sampling over the duration of the utterance. Spectral tilt occurs because high frequencies typically have less power than low frequencies and therefore the power spectrum slopes downward from low to high frequencies. Flattening spectral tilt, that is, migrating power to high-frequency bands, improves the intelligibility of speech in noise (Lu & Cooke, 2009). Spectral tilt may relate to roundedness/pointedness in that a steep slope, in which power is concentrated in the low-frequency bands, is more likely to reflect sonorants and back rounded vowels that are associated with roundedness (McCormick et al., 2015). However, the slope should flatten out for pseudowords containing obstruents and/or front unrounded vowels associated with pointedness (McCormick et al., 2015) as power migrates to the higher frequencies associated with these phonemic properties.

Temporal FFT: The FFT converts temporal or spatial signals into the corresponding frequency domain. The FFT analysis of temporal data, such as the acoustic speech signal in our pseudowords, derives the frequency components of that signal, some with more energy than others, and can be calculated over the duration of the sound signal. Thus, this parameter reflects the power spectrum of the frequency composition across time (Singh, 2015). To the extent that there is more power at the lower/higher frequencies, associated with roundedness/pointedness, respectively, the temporal FFT should reflect the shape associations of the pseudowords.

Fraction of unvoiced frames: This is a measurement of voice stability over time, with the number of unvoiced elements expressed as the percentage of measurement windows that do not engage the vocal folds (Boersma & Weenink, 2012). FUF depends on the phonemic content of an utterance, particularly when measured across the duration of the pseudoword, and will obviously increase for utterances that include unvoiced elements like obstruents and decrease for those containing voiced elements, typically long vowels

(Mezzedimi, di Francesco, Livi, Spinosi, & De Felice, 2017). Since auditory “roundedness” is more associated with voiced than unvoiced elements (McCormick et al., 2015), we would expect FUF to increase as ratings of pseudoword transition from rounded to pointed.

Mean autocorrelation: This a measure of the similarity, or correlation, between a sound and a delayed copy of itself. As such, it is a measure of the periodicity of a signal wherein 0 is a white noise signal and 1 is a perfectly periodic signal (Boersma & Weenink, 2012). When a single phoneme is sustained, for example a long vowel like “ooo” or consonant like “mmm,” each successive segment should sound very similar to the one before; that is, they should be highly correlated. Higher autocorrelation values indicate a smoother voice pattern and/or more voiced segments, which should be reflected in roundedness ratings, while lower values indicate an uneven pattern and/or fewer voiced or periodic segments, which should be reflected in ratings of pointedness.

Pulse number: This is the number of glottal pulses, that is, opening and closing of the vocal folds, during production of vowels or voiced consonants measured across the whole utterance (Boersma & Weenink, 2012). To understand how this manifests in the voice, it is necessary to consider an extreme form of phonation, known as pulse register phonation, in which rapid glottal pulses are followed by a long closed phase (Hollien, Girard, & Coleman, 1977; Whitehead, Metz, & Whitehead, 1984). The auditory perception of this vocal register has been described as a “creaky voice” (Ishi, Sakakibara, Ishiguro, & Hagita, 2008) or—onomatopoeically—as a “glottal rattle” (Hornibrook, Ormond, & Maclagan, 2018). As such, it is a measure of vocal roughness or unevenness; lower pulse numbers indicate a rougher, more uneven voice pattern, and/or fewer voiced segments, which should be associated with pointed pseudowords, while higher pulse numbers indicate a smoother voice pattern, and/or more voiced segments, which should be associated with rounded pseudowords.

Jitter: This is a measure of voice quality that indexes variation in the vibration of the vocal cords (Teixeira & Fernandes, 2014). Jitter is defined as the frequency variation between consecutive periods expressed as a percentage; here, we calculated local jitter, the mean absolute difference in frequency between consecutive periods of the speech waveform divided by the mean difference over all periods of the speech waveform and expressed as a percentage (Boersma & Weenink, 2012). Perceptually, high values of jitter manifest as a “breaking” or rough voice, that is, one that varies in the consistency and length of each period of the waveform corresponding to each opening and closing of the vocal cords. Jitter is typically measured for long vowel sounds, where little frequency variation would be expected, and therefore high levels of jitter indicate voice pathology (Teixeira & Fernandes, 2014). Jitter and shimmer (see below) have also been associated with changes in emotion and stress in speech (Van Puyvelde, Neyt, McGlone, & Pattyn, 2018), suggesting that this acoustic measure can convey non-linguistic information. In the production of the pseudowords, cycle-to-cycle frequency variation or jitter should increase from rounded to pointed pseudowords, reflecting increased vocal instability or variation and perceived roughness.

Shimmer: In contrast to jitter, shimmer is a measure of voice quality that indexes period-to-period variation in amplitude (Brockmann et al., 2011). While minor variations in amplitude are normal, substantial variability can indicate voice pathology stemming from glottal resistance, that is, stiffness of the vocal cords, which manifests as breathiness or hoarseness (Brockmann et al., 2011; Teixeira & Fernandes, 2014). Since shimmer reflects vocal instability, low shimmer manifests in a smooth speech pattern, whereas high shimmer results in an uneven speech pattern that should be associated with roundedness and pointedness, respectively. Here, we measured local shimmer, defined as the mean absolute difference in amplitude between consecutive periods divided by the mean amplitude and expressed as a percentage (Boersma & Weenink, 2012).

Pitch standard deviation: The PSD indicates the variation in the fundamental frequency present in the speech signal. This is a measure of vocal inflection with low PSD manifesting as a level, monotone voice and high PSD as a "lively" voice (Kliper, Portuguese, & Weinshall, 2016). As such, PSD can reflect an individual's emotional state (Kliper et al., 2016), suggesting that this acoustic measure is also capable of conveying non-linguistic information. For present purposes, low and high PSD/vocal inflection should indicate roundedness and pointedness, respectively.

Mean harmonics-to-noise ratio: This parameter measures the ratio between the dominant periodic, or harmonic, element of the speech signal and the aperiodic, or noise, element, thus providing an estimate of the overall periodicity of the sound expressed in dB (Teixeira & Fernandes, 2014). The noise element arises from turbulent airflow at the glottis when the vocal cords do not close properly (Ferrand, 2002). As the noise element increases, and therefore, mean HNR decreases, the voice becomes increasingly hoarse or quavery (Ferrand, 2002). In other words, as mean HNR decreases, the speech pattern becomes progressively less smooth and more uneven, reflecting a transition from roundedness to pointedness.

2.3.2. Visual parameters of shapes

As mentioned in the Introduction, the choice of visual parameters of the shapes was constrained by the fact that the shapes were all irregular; thus, we were unable to employ radial frequency measures (Wilkinson et al., 1998). For irregular shapes, one option would be to adopt the particle morphology measure of "roundness" used in geology to classify grain shape by curve fitting (Boggs, 2009; Folk, 1965).[3] However, while this would be possible here for the rounded shapes, it would not be meaningful for the pointed shapes because their protuberances end in a single point; that is, curvature is zero. In practice, geologists generally classify particles by reference to visual analog scales (Folk, 1965, p. 10), highly similar to the approach we took here for perceptual ratings. More recently, particle morphology has been assessed using Fourier analyses (Boggs, 2009) similar to the spatial FFT described below. As a first step in calculating the visual parameters, we removed excessive background from the images to arrive at the smallest area that contained all the shapes overlaid on one another. This area was 200 × 200 pixels, giving 40,000 data points for each shape for all visual parameters (see Figs. S1 and S2). Note that the visual parameters are indices of global shape, whereas the acoustic parameters reflected specific acoustic aspects.

Jaccard distance

The Jaccard distance is a measure of dissimilarity between two sets or items that uses a present/absent coefficient (Ricotta & Pavoine, 2015) and has been used as a measure of shape similarity (e.g., Davico et al., 2019; Devaprakash et al., 2019). The Jaccard similarity coefficient, J (Jaccard, 1901), can be interpreted as the intersection of the two shapes divided by their union; that is, the more the two shapes overlap when superimposed on each other, the larger their intersection and the greater the similarity coefficient. The starting point is to designate pixels in the shape as 1 and pixels in the background as 0 (e.g., Devereux, Clarke, Marouchos, & Tyler, 2013) and to calculate J for each pair of shapes. The coefficient J is given by $a/(a + b + c)$, where a = pixels present in both shapes, b = pixels present in the first shape but not the second, and c = pixels present in the second shape but not the first. The Jaccard distance is then $1 - J$ and in constructing the RDM this pairwise measure replaces $1 - r$.

Simple matching coefficient

The SMC also reflects shape similarity, being calculated in the same way as the Jaccard similarity coefficient except that it includes an additional term, d, representing pixels that are absent from both shapes in the particular pair under consideration but present in other shapes in the set (Ricotta & Pavoine, 2015). This term appears in in both numerator and denominator such that the SMC is given by $(a + d)/(a + b + c + d)$. The RDM is constructed by replacing $1 - r$ with the pairwise $1 -$ SMC as a measure of dissimilarity. By taking account of pixels that are present in other shapes in the set, the SMC provides a measure of the similarity of particular shapes not only to each other but also in relation to the remaining shapes in the set.

Image silhouette

Image silhouettes enable us to compare shapes on the basis of low-level visual feature information (e.g., Devereux et al., 2013; Kriegeskorte et al., 2008). Images are binarized such that pixels in the shape = 1 and pixels in the background = 0 (Devereux et al., 2013), that is, essentially separating figure from ground. Thus, this parameter explicitly codes roundedness and pointedness and, by including pixels within the shape, also accounts for area. The resulting 2D information is reduced to a single vector for each image (e.g., the vectors 1,0,1,0,1 … n and 1,1,1,1,0 … n describe different shapes by recording the presence/absence of shape pixels at specific positions in the vector) and the pairwise correlation of these vectors forms the basis of $1 - r$ in the RDM.

Image outlines

In this case, perimeter pixels forming the shape outline = 1 and all other pixels = 0. The results are again vectorized for each image and the pairwise correlation of these vectors forms the basis of $1 - r$ in the RDM. In contrast to the image silhouette parameter, the image outline provides an index of roundedness and pointedness independent of area by focusing only on the perimeter pixels.

Spatial FFT

The spatial FFT is based on grayscale value variations at each point in space, that is, at each pixel across the whole image, and captures how often these variations repeat per unit of distance, that is, their spatial frequency. Thus, analogously to the temporal FFT, this parameter reflects the power spectrum of frequency distribution across space: Concentrations of power at low or high frequencies should reflect roundedness or pointedness, respectively.

Note that the RDMs for image silhouette, outlines, and the spatial FFT are calculated using the pairwise first-order Pearson correlation and therefore $1 - r$ can range from 0 to 2, while the simple matching coefficient and Jaccard distance RDMs are calculated with coefficients that cannot exceed 1 and therefore $1 - J$ and $1 -$ SMC range from 0 to 1.

3. Results

3.1. Crossmodal comparison of visual and auditory perceptual ratings

Examination of the RDMs for the perceptual ratings of the pseudowords and shapes suggests that roundedness/pointedness ratings for the pseudowords were relatively graded: While some were rated as very rounded or pointed, leading to high similarity values at either end of the diagonal and a cluster of high dissimilarity values at the corners, there was a wide range of pseudowords rated at intermediate points on the scale, leading to a range of dissimilarity values (Fig. 3, left panel). By contrast, roundedness/pointedness ratings for the shapes were essentially binary (Fig. 3, right panel): Participants largely rated the shapes as either highly rounded or highly pointed with few shapes considered intermediate, leading to dissimilarity values that tended to be uniformly high or low. Despite this apparent qualitative difference, there was a significant, positive, second-order correlation between the RDMs for the auditory and visual perceptual ratings ($r_{s\ 4003} = .64$, $p < .0001$), indicating that the ratings were crossmodally consistent even though they were made by independent groups of participants. This is important because the auditory pseudowords were rated for a property that is primarily defined visually and therefore crossmodal consistency was not guaranteed a priori. In the present context, this crossmodal consistency serves as a verification of sound-symbolic associations between the auditory pseudowords and visual shapes used.

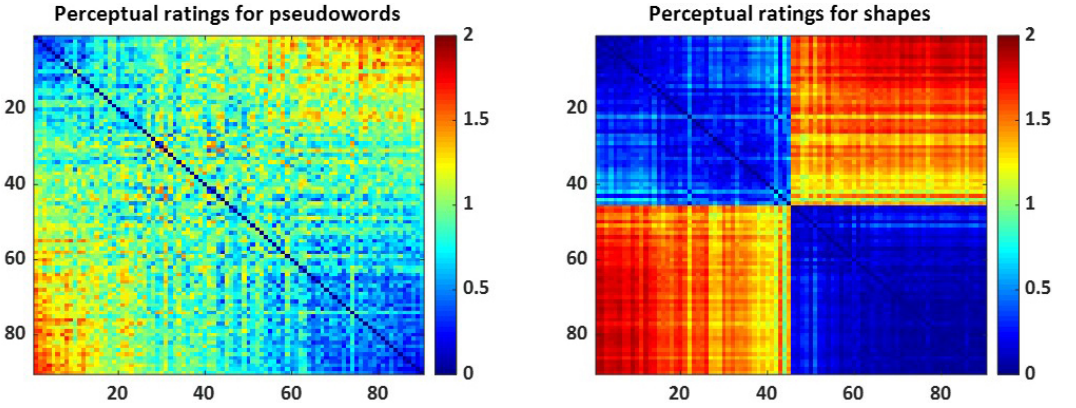

Fig. 3. Representational dissimilarity matrices (RDMs) for perceptual ratings of roundedness/pointedness for auditory pseudowords (left) and visual shapes (right). Items in each RDM are ordered left to right from most rounded to most pointed. Color bar shows pairwise dissimilarity, where 0 = zero dissimilarity (items are identical) and 2 = maximum dissimilarity (items are completely different). Auditory and visual RDMs were significantly positively correlated ($r_{s\ 4003} = .64$, $p < 0.0001$).

3.2. Comparison of auditory perceptual ratings to acoustic parameters of pseudowords

To determine which of the selected acoustic parameters were related to the auditory perceptual ratings of roundedness/pointedness, we computed the second-order correlation between the RDM for auditory perceptual ratings and that for each of the acoustic parameters (Fig. 1, Step 3). After correction for multiple comparisons (Bonferroni-corrected α for 10 tests = 0.005), there were significant positive correlations for spectral tilt ($r_{s\ 143914} = .43$, $p < .0001$), the temporal FFT ($r_{s\ 143914} = .25$, $p < .0001$), and speech envelope ($r_{s\ 143914} = .14$, $p < .0001$): Fig. 4 shows the RDMs for these, illustrating the 537 × 537 matrices for the entire pseudoword set. The mean autocorrelation ($r_{s\ 151} = -.2$, $p = .02$) and mean HNR ($r_{s\ 151} = -.16$, $p = .04$) were also correlated (negatively), but these correlations did not survive Bonferroni correction (Fig. S3 shows the down-sampled 18 × 18 RDMs for these). The remaining acoustic parameters were uncorrelated ($r_{s\ 151} = -.05$ to -0.1, all $p > .1$: Fig. S3). As the Bonferroni correction method is relatively conservative, minimizing Type 1 error, we also tested for significance with the less restrictive modified Bonferroni correction suggested by Holm (1979), but the pattern of results was unchanged, indicating that Type 2 error was unlikely.

Notwithstanding the different numbers of samples per pseudoword, the acoustic parameters whose RDMs were correlated with the RDM for perceptual ratings of the pseudowords are likely to be important for auditory perception of pointedness/roundedness. The spectral tilt, temporal FFT, and speech envelope parameters all included multiple measurements per pseudoword, thus preventing a simple correlation between these and the single rating value for each pseudoword. Therefore, we illustrate their relationships with perceptions of roundedness/pointedness qualitatively by providing examples that show how these parameters vary between a more rounded pseudoword (mumo:

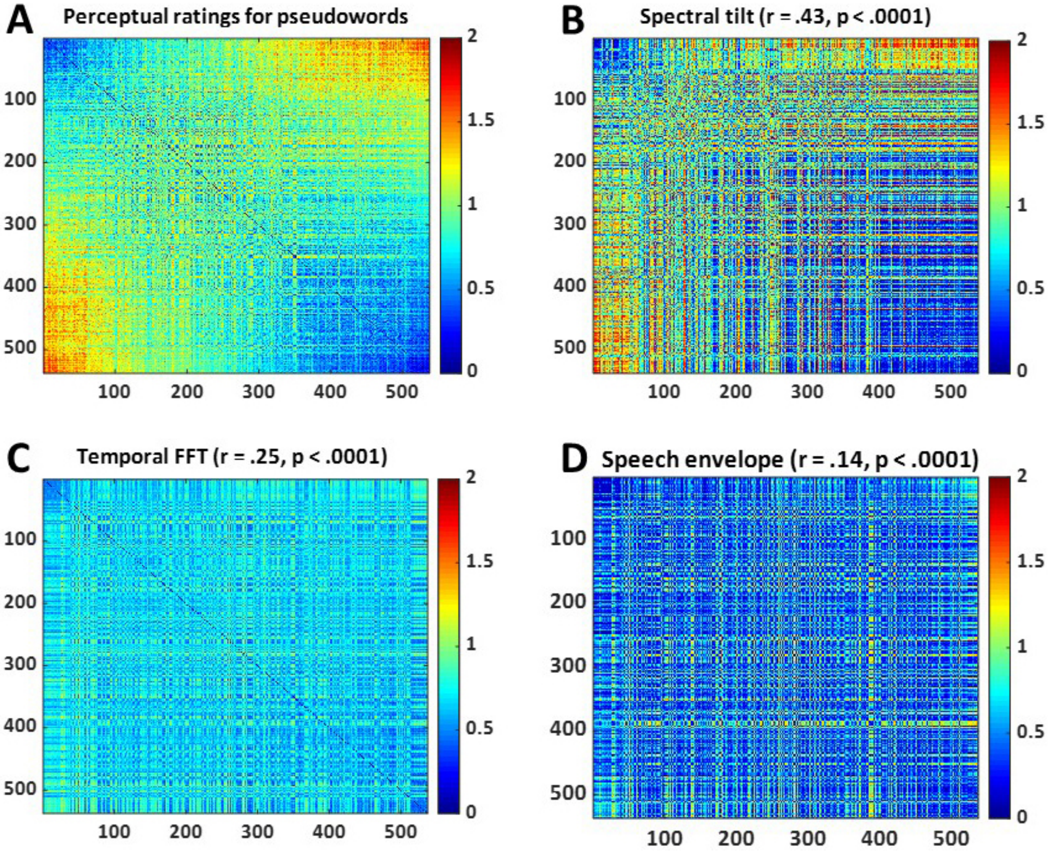

Fig. 4. Representational dissimilarity matrices (RDMs) for (A) auditory perceptual ratings of roundedness/pointedness for the pseudowords; (B) spectral tilt; (C) temporal FFT; (D) speech envelope. Items in each RDM are ordered left to right from most rounded to most pointed. Interpretation of the color bar as for Fig. 3; r = Spearman correlation coefficient for B–D versus A; df = 143,914 in all cases.

"moo-moh") and a more pointed pseudoword (kete: "keh-teh"). For each pseudoword, Fig. 5A illustrates the spectral tilt, the overall slope of the spectrum from low to high frequencies. As predicted, the slope is steeper for the rounded pseudoword "mumo" (Fig. 5A, left panel), because power is concentrated in the low-frequency bands associated with the sonorants and back rounded vowels that reflect roundedness (McCormick et al., 2015). By contrast, the slope is flatter for the pointed pseudoword "kete" (Fig. 5A, right panel) because more power is present at the higher frequencies of the obstruents and/or front unrounded vowels associated with pointedness (McCormick et al., 2015). Fig. 5B shows the waveforms and the spectrograms resulting from the temporal FFT; these both clearly distinguish "mumo," with its voiced segments, from "kete," which contains unvoiced consonants. The spectrogram (Fig. 5B, bottom panels) plots amplitude (the degree of dark shading) for multiple frequencies (y-axis) across time (x-axis) and shows smoother changes for “mumo” compared to “kete” where these are more abrupt, especially at high frequencies. Fig. 5C shows the speech envelope for each pseudoword and that, as predicted, the envelope is continuous and smoother for “mumo” compared to “kete,” which has an envelope that is discontinuous and uneven. Note that the speech envelope is related to the waveform, where this continuous-smooth/discontinuous-uneven relationship for rounded versus pointed words can also be seen (Fig. 5B, top panels: Compare also the waveforms for “mumo” and “kete” to the intermediate pseudoword “zuvu” in Fig. 1, Step 1).

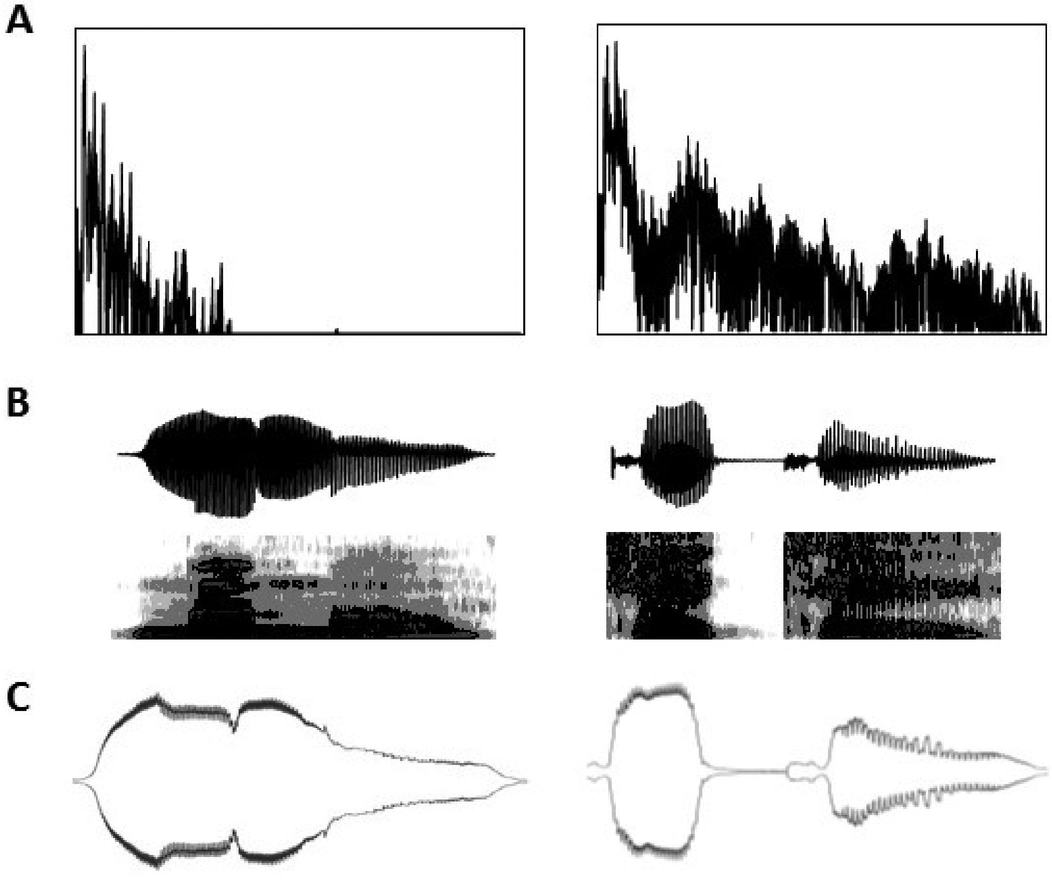

Fig. 5. (A) Spectral tilt, the overall slope of power spectral density, for the rounded pseudoword “mumo” (left) and the pointed pseudoword “kete” (right). Spectral tilt is steeper for “mumo” where power is concentrated in low frequency bands, but flatter for “kete” as power migrates to high frequency bands. (B) The waveform (top panels) and spectrogram (bottom panels) illustrate aspects of the temporal FFT; the spectrogram captures more abrupt changes in power, especially at higher frequencies, for “kete” compared to “mumo.” (C) Speech envelope for “mumo” is continuous and smoother compared to “kete,” which is discontinuous and uneven (similar to the waveform for these pseudowords in B: top panels). All examples produced using PRAAT speech analysis software (Boersma & Weenink, 2012).

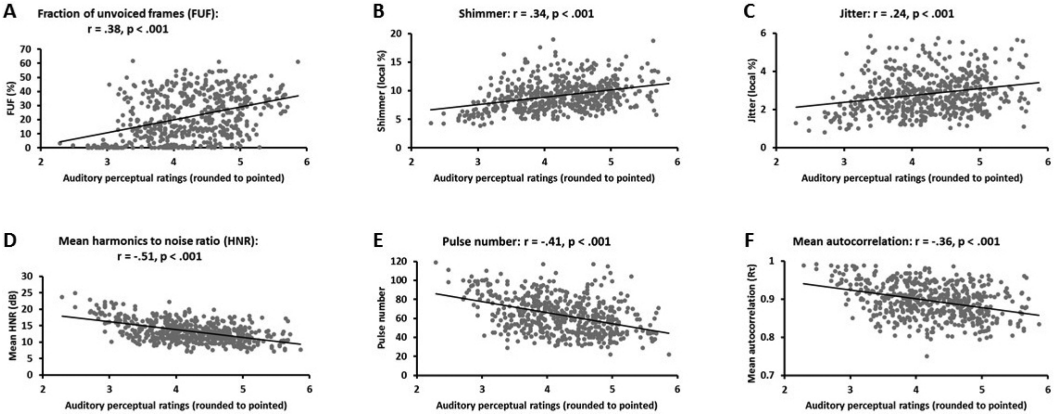

Fig. 6. Correlation between auditory perceptual ratings of the pseudowords and acoustic parameters of voice quality (note that the rating scale is truncated in all panels because there are no values <2 or >6; the mean autocorrelation scale (F) is truncated because there are no values below 0.7). df = 535 in all cases.

Although RSA showed that none of the acoustic voice quality parameters was significantly related to auditory perceptual ratings of the pseudowords, the power of these analyses was limited by the smaller matrix size (18 × 18) for the RDMs of these parameters. Therefore, we supplemented RSA with conventional correlation analyses between the parameter values (one per pseudoword) and their perceptual ratings. These analyses showed that many of these parameters were related to auditory perception of roundedness/pointedness, demonstrating the relationships predicted in Section 2.3.1. After correction for multiple comparisons (Bonferroni-corrected α for 7 tests = .007), the FUF, shimmer, and jitter were all significantly positively correlated with perceptual ratings (Fig. 6A–C, respectively) reflecting increasing variation in aspects of voice quality as the pseudowords transitioned from rounded to pointed. Mean HNR, pulse number, and mean autocorrelation were all significantly negatively correlated with perceptual ratings (Fig. 6D–F); for these parameters, *lower* values indicate greater vocal variability or roughness in voice quality as well as the presence or absence of voiced segments in each pseudoword and were associated with higher ratings indicating pointedness. However, the correlation between perceptual ratings and PSD was relatively weak (r_{535} = .1) and did not pass the Bonferroni-corrected α (see Fig. S4). With the exception of PSD and the mean autocorrelation, all the voice parameter measurements were non-normally distributed; therefore, we also tested these relationships with the non-parametric Spearman test, but, although the correlations were slightly weaker, we obtained the same pattern of results. Neither set of results was changed by reference to the modified Bonferroni-correction (Holm, 1979).

3.3. *Comparison of visual perceptual ratings to visual shape parameters*

We carried out within-modal second-order correlations between the RDM of the visual ratings of roundedness/pointedness and that for each visual parameter of the shapes

(Fig. 1, Step 3). After correction for multiple comparisons (Bonferroni-corrected α for 5 tests = 0.01), there were significant positive correlations for the SMC ($r_{\text{s}\ 4003} = .28$, $p < .0001$), silhouette ($r_{\text{s}\ 4003} = .14$, $p < .0001$), image outlines ($r_{\text{s}\ 4003} = .13$, $p < .0001$), and Jaccard distance ($r_{\text{s}\ 4003} = .1$, $p < .0001$: Fig. 7) but not for the spatial FFT ($r_{\text{s}\ 4003} = .01$, $p = .6$: Fig. S5). This pattern of results was unchanged by reference to the modified Bonferroni-correction (Holm, 1979). Note that the RDM for image outlines (Fig. 7D) indicates relatively higher dissimilarity between shapes than that for other visual parameters, and that dissimilarity values ($1 - r$) were fairly uniform around 1, indicating that the majority of shapes were only weakly correlated with each other, whether positively or negatively. The reason for this can be seen in Fig. S1, which shows all 90 image outlines overlaid on one another with darker/lighter areas, indicating the intersection of more/fewer outlines. There are relatively few very dark intersections, indicating that shape outlines rarely overlapped by much with other shapes and that therefore all the shapes were different to a large degree.

3.4. Comparison of acoustic and visual parameters

Finally, to the extent that they were significantly correlated with their within-modal perceptual ratings, we compared RDMs of the acoustic and visual parameters to each other crossmodally (Fig. 1, Step 4). We compared RDMs of the three most strongly

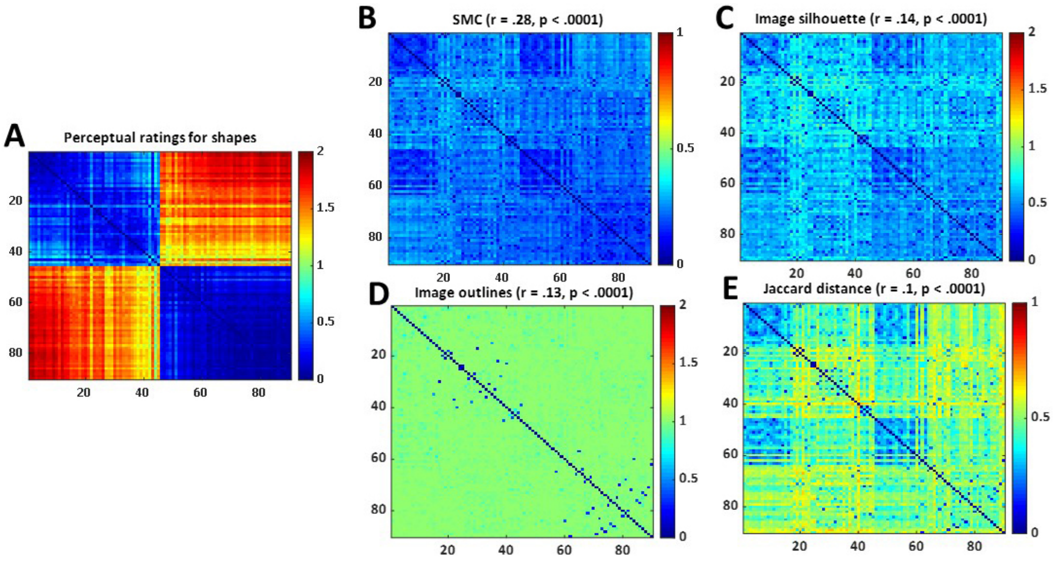

Fig. 7. Representational dissimilarity matrices (RDMs) for (A) perceptual ratings of roundedness/pointedness for visual shapes; (B) SMC; (C) image silhouette; (D) image outlines; and (E) Jaccard distance. Items in each RDM are ordered left to right from most rounded to most pointed. Interpretation of the color bar as for Fig. 3; r = Spearman correlation coefficient for B–E versus A; df = 4,003 in all cases. Note that the maximum dissimilarity value for the simple matching coefficient (B) and Jaccard distance (E) is 1 while that for image silhouette (C) and outlines (D) is 2 (see Section 2.3.2).

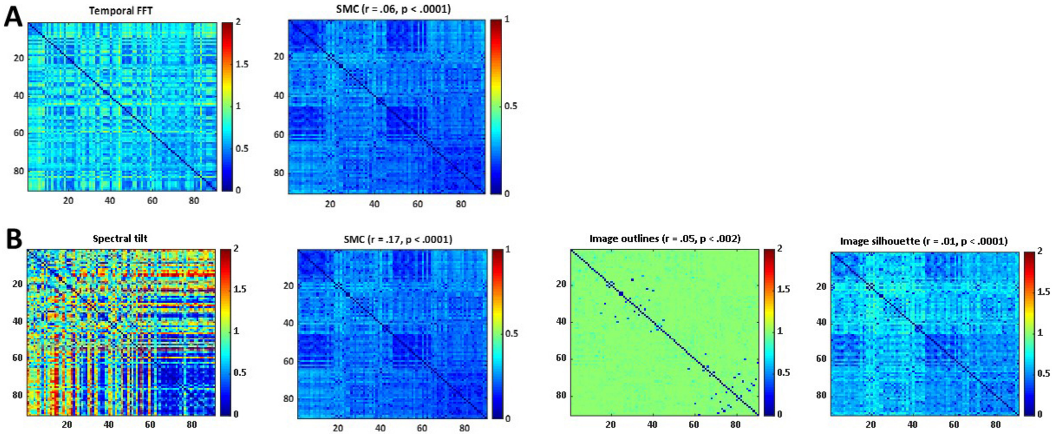

Fig. 8. (A) Representational dissimilarity matrices (RDMs) for temporal FFT and the simple matching coefficient were correlated; (B) the RDM for spectral tilt was correlated with the SMC, silhouette, and image outlines. Interpretation of the color bar as for Fig. 3; r = Spearman correlation coefficient for the RDMs of the visual parameters relative to the RDM of the acoustic parameter at the left of each row; df = 4,003 in all cases.

correlated visual parameters: the SMC, silhouette, and image outlines, and RDMs of the three most strongly correlated acoustic parameters: spectral tilt, temporal FFT, and speech envelope, down-sampling the acoustic RDMs (by selecting every sixth word from rounded to pointed) to maintain a consistent matrix size of 90 × 90.

After correction for multiple comparisons (Bonferroni-corrected α for 9 tests = 0.0056), the auditory temporal FFT was only significantly correlated with the visual SMC ($r_{s\ 4003} = .06$, $p < .0001$), but auditory spectral tilt was significantly correlated with the visual parameters of SMC ($r_{s\ 4003} = .17$, $p < .0001$), silhouette ($r_{s\ 4003} = .01$, $p < .0001$), and image outlines ($r_{s\ 4003} = .05$, $p = .002$: Fig. 8); all other correlations were nonsignificant ($r_{s\ 4003} = -.002$ to .02, all $p > .1$). This pattern of results was unchanged by reference to the modified Bonferroni-correction (Holm, 1979). As laid out in Section 2.2, these significant crossmodal correlations (albeit weak) between parameters provide some comfort that the independent groups involved in the ratings exercise were likely employing a common perceptual framework for pointedness/roundedness ratings regardless of modality. More specifically, the temporal FFT and spectral tilt may be capturing aspects of the speech signal that are related to the spatial aspects of shape captured by the SMC, image outlines, and silhouette.

4. Discussion

4.1. Value of RSA

The present study is the first to examine how acoustic and visual parameters contribute to sound-to-shape mapping in the same experimental paradigm. Previous studies have only examined these separately (acoustic—Knoeferle et al., 2017; Parise & Pavani, 2011; visual—Chen et al., 2016). In addition, the number of pseudowords used here, and thus the sampling of the potential phonemic space, is much more comprehensive than earlier investigations of pseudoword-shape mapping: 537 versus 100 in Knoeferle et al. (2017), the largest previous set to our knowledge. While the method of creating rounded/pointed shapes was very different, our set size of 90 is comparable to that of 72 in Chen et al. (2016), again the largest that we are aware of in earlier work. We also demonstrate the utility of a novel application of RSA (Kriegeskorte et al., 2008) for assessing the relevance of a particular acoustic parameter to the perception of an auditory pseudoword as rounded or pointed, and facilitating crossmodal comparison with visual parameters. The use of RSA allows very different stimulus sets and physical parameters to be compared on an equal footing, that of their pairwise dissimilarity (Kriegeskorte et al., 2008). The advantages of RSA are two-fold: First, it allows comparison of parameters that involve multiple measurements/samples of pseudowords, as in the case of a number of the acoustic and visual parameters studied here. For these parameters, a conventional correlational analysis is not appropriate to assess their relationship to perceptual ratings, or for crossmodal comparisons. The RSA approach enables such comparisons by constructing dissimilarity matrices in which each cell contains a single value representing the pairwise dissimilarity between stimuli, effectively compressing the large number of samples per stimulus. The resulting matrices are then simply compared using non-parametric (Spearman) correlation. Second, regardless of the parameter, RSA compares each item to every other item; thus, all possible pairs enter the analysis as opposed to conventional correlational analyses which only treat items as a group. This RSA approach, combined with substantially larger set of pseudowords than in many previous studies, allowed us to investigate sound-to-shape mapping not only at a more granular level but also more comprehensively. A potential drawback of RSA, however, is that for parameters that are expressed as a single value for each item, the items must be binned into sets of stimuli to allow computation of dissimilarity; since the number of such item sets is only a fraction of the total number of items, this results in a loss of sensitivity. For such parameters, more conventional correlational analyses may remain appropriate, as discussed further below.

The RSA approach could prove fruitful in investigating other sound-symbolic mappings. For example, in sound-to-size mapping (e.g., “mil” is small and “mal” is large: Sapir, 1929), we would expect that spectral tilt would be flatter for “small” words, since these would involve the higher frequencies associated with small size, and steeper for “large” words since these would involve lower frequencies, a prediction that can be made not only from the findings of Knoeferle et al. (2017) but also from the well-known crossmodal correspondence between auditory pitch and visual size (Evans & Treisman, 2010; Gallace & Spence, 2006). In fact, non-linguistic crossmodal correspondences might be a good source of predictions about sound-symbolic mapping for pseudowords: For example, we could expect pseudowords reflecting the brightness/darkness dimension to be modulated by their amplitude and/or pitch (see Spence, 2011). Indeed, if these non-linguistic correspondences are found to effectively predict sound-symbolic linguistic correspondences, it would suggest that natural languages incorporate these general perceptual or cognitive constraints into the mapping of sound to meaning (Blasi et al., 2016; Namy & Nygaard, 2008; Revill et al., 2014). Additionally, other types of associations, for example between roundedness/pointedness and female/male first names, respectively, previously demonstrated by shape-matching and phonemic analysis (Sidhu & Pexman, 2015), could also be reflected in acoustic analyses of those names.

4.2. Relationships between acoustic parameters and pseudoword ratings

The present study used a large set of pseudowords rather than real words. This had the advantage of being a highly constrained set of stimuli, controlling for, and systematically sampling variation in, vowel quality, consonant voicing, manner and place of articulation, and syllable structure (McCormick et al., 2015), thus enabling us to assess the roundedness/pointedness of a wide range of speech sounds that were as free of semantic associations as possible. Perception of a pseudoword as either rounded or pointed may depend on the acoustic consequences of the phonological content of the word itself and/or the vocal properties of the speaker's voice. RSA showed that there were relatively strong correlations between the RDM of auditory perceptual ratings and the RDMs for spectral tilt, temporal FFT, and speech envelope, parameters that likely primarily reflect phonetic content. Spectral tilt was steeper for rounded pseudowords where power is concentrated at the lower frequencies associated with sonorants and back rounded vowels that reflect roundedness, but flatter for pointed pseudowords as spectral power migrated to the higher frequencies for obstruents and/or front unrounded vowels associated with pointedness (McCormick et al., 2015). Underlying the temporal FFT relationship, changes in the distribution of spectral power over time were smoother and occurred at lower frequencies for rounded pseudowords while for pointed pseudowords these transitions were more abrupt and occurred at higher frequencies, consistent with previous work showing that formant frequencies are higher for more pointed words (Knoeferle et al., 2017) and for vocalizations produced in response to more pointed shapes (Parise & Pavani, 2011). The speech envelope was smoother and more continuous for rounded words, whereas for pointed words it was more uneven and discontinuous.

However, RSA did not show significant relationships between perceptual ratings and any of the measures of voice quality, which were all based on a single measure per pseudoword; thus, to create RDMs, the data had to be combined across multiple contiguous stimuli (we chose 30) in the larger matrix, so as to allow computation of dissimilarity between *sets* of stimuli, rather than between individual stimuli as was possible for spectral tilt, temporal FFT, and speech envelope. Since the resulting matrices (18 × 18) were substantially smaller than the full matrices (537 × 537), the statistical power of RSA was necessarily limited for comparisons based on these variables. To overcome this limitation, we also conducted conventional correlational analyses for these parameters. These correlational analyses showed that several measures of vocal quality (FUF, jitter, shimmer, mean HNR, pulse number, and mean autocorrelation) were predictors of perceptual ratings of the pseudowords. As the variability of voice parameters increased or voice quality changed, as reflected in higher FUF, jitter, and shimmer values and lower mean HNR, pulse number and mean autocorrelation values, these increases in the noisy or rough quality of the speech pattern became less associated with roundedness and more associated with pointedness. Although these parameters are typically used to distinguish characteristics of individual speakers and to characterize and assess properties of voice quality, it is notable that in the present context, variation along these dimensions relates to differences among pseudoword productions by a single speaker (see Brockmann et al., 2011). The relationship between parameters of voice quality and sound-to-shape mapping is novel and raises the question whether different vocal registers can be manipulated to influence this mapping; for example, whether rounded words spoken with the "glottal rattle" of the pulse register (Hornibrook et al., 2018) would be perceived as less rounded than when spoken in the modal register of normal speech (Nygaard, Herold, et al., 2009; Tzeng, Duan, Namy, & Nygaard, 2018). However, because we compared across many different pseudowords sampling an array of phonetic features and assessed these voice measures across the entire utterance, the effects may not have exclusively reflected changes in vocal quality since, as noted in Section 2.3.1, these parameters can be influenced by phonemic content as well. In this context, we also observed no significant relationship between perceptual ratings of the pseudowords and variation in fundamental frequency or PSD (whether assessed via conventional correlational analyses or RSA). Since the speaker deliberately recorded the words with minimal inflection, it is possible that pitch did not vary enough for an effect of PSD to be detected. Thus, while the other voice parameters may reflect both the acoustic correlates of the speech sounds that each pseudoword contains and voice quality differences, PSD may in this case may have primarily reflected how these speech sounds were produced.

In using continuous measurements of acoustic parameters and perceptual ratings, the present results extend previous work which relied on categorical linguistic contrasts, for example between consonants and vowels (Fort et al., 2015; Nielsen & Rendall, 2011) or voiced and unvoiced phonemes (McCormick et al., 2015). McCormick et al. (2015) suggested that perceivers make rounded/pointed judgments of a pseudoword by reference to both its specific individual phonemic components and to the overall inventory of features (acoustic, linguistic, or articulatory) within the utterance. Measuring acoustic parameters of the entire speech signal allowed us to provide some evidence in support of the idea that perceivers based their shape judgments on a global auditory assessment of each pseudoword. Spectral tilt, the temporal FFT, and the speech envelope are all complex measures of the complete pseudoword and cannot be reduced to a single value. Consistent with such holistic processing of the speech signal, these parameters were all related to roundedness/pointedness ratings, as demonstrated using RSA. Styles and Gawne (2017) reported a failure to replicate sound-to-shape mapping across languages and suggested that this was because the pseudowords employed, "kiki" and "bubu," did not conform to the phonological structure of the target language. However, failures to replicate invariably involved categorical responses (Bremner et al., 2013; Rogers & Ross, 1975; Styles & Gawne, 2017), so this may only be a partial explanation. This could be explored further using RSA for larger stimulus sets and continuous measurements of acoustic parameters.

4.3. Relationships between visual parameters and shape ratings

Since the visual shapes employed here were asymmetric, we were limited in our measurement choices. Nonetheless, RSA showed that the SMC, silhouette, image outlines, and Jaccard distance were related to perception of the shapes as rounded or pointed. The SMC and Jaccard distance are pairwise measures of global shape matching, while the silhouette and image outlines are vectorized measures of the shapes that lend themselves to computation of pairwise dissimilarity. Thus, the RDMs based on these measures were all constructed from estimates of the pairwise dissimilarity of global shape. However, it should be noted that, aside from the SMC, the relationships between these RDMs and the RDM for visual perceptual ratings were modest, possibly reflecting the limited degrees of freedom used to generate the shapes: They were compositionally very similar, all consisting of gray outlines on a white background so that the grayscale contrast was identical across all shapes, and they all lacked internal patterns. Interestingly, the RDMs for the visual ratings and the spatial FFT were uncorrelated, suggesting that spatial frequency is not a critical parameter underlying sound symbolism, at least for the visual shapes used here.

4.4. Crossmodal relationships

Of the crossmodal comparisons, the RDM for acoustic spectral tilt was significantly correlated with the RDMs for the visual SMC, silhouette, and image outlines; the RDMs for the acoustic temporal FFT and the visual SMC were also correlated. Although these relationships were fairly weak, it is worth noting that spectral tilt (the parameter most strongly correlated with ratings of auditory roundedness/pointedness) was correlated with three visual indices, indicating that it may indeed be related to some aspects of visual shape and thus relevant to the kind of sound-symbolic crossmodal correspondence studied here. The crossmodal relationship between the SMC and the temporal FFT (another auditory spectral parameter) may also be tapping into this sound-symbolic correspondence. It is interesting that the spectral parameters of the pseudowords were related to their auditory ratings on the rounded-to-pointed dimension, and that both the spectral parameters we tested were related to global indices of visual shape, which were themselves related to the visual ratings of the shapes on the rounded-to-pointed dimension. We propose that these relationships may be particularly relevant for sound symbolism. However, further work is needed to confirm that auditory spectral parameters and global indices of visual shape, but not the spatial frequency spectrum of visual shapes, underpin sound-symbolic crossmodal correspondences. Further work should also examine why these crossmodal relationships are fairly weak: It may be that the underlying relationships are not linear, or it may simply be that the auditory parameters are a proxy for roundedness/pointedness while the visual parameters measure this more explicitly. Although a number of voice quality measures were related to auditory perceptual ratings, we did not attempt to directly connect them to the visual shape measures that were related to the visual perceptual ratings, since the former relationships were based on conventional correlation and the latter on RSA.

An interesting point is that the sensory modality most associated with a particular word can change or be added to over time[4] (Marks, 1978). For example, in Old English the word "sharp" originally applied primarily to the sense of touch before becoming associated with taste during the eleventh century, and visual shape and audition during the fourteenth century in Middle English (Marks, 1978). Relatedly, it is well known that the "hierarchy of the senses" has changed over time (see Kambaskovic & Wolfe, 2016); for example, both touch and hearing have been considered more primary than vision at different times. While the timescales involved probably preclude empirical enquiry, it may be worth considering whether a set of pseudowords have stronger connections to sound-symbolic mappings in one modality over another, and whether this follows the current sensory hierarchy.

4.5. Limitations and future directions

An obvious limitation of the present study is that, inevitably, we did not test all possible acoustic and visual parameters of the pseudowords and shapes; we may therefore have omitted parameters that turn out to be equally, or more, important. However, to the extent that the parameters examined here were not significantly related to the perceptual ratings or crossmodally, either using RSA or conventional correlations, our results help focus the search space for future studies. Also, we only tested the roundedness/pointedness dimension; acoustic and visual parameters might be differently weighted for other dimensions in other domains relevant to sound symbolism (see Knoeferle et al., 2017, for different weightings for shape and size). For instance, acoustic parameters that do not contribute to perception of roundedness/pointedness might still be important for onomatopoeic words, like "bang," "splash," or "slap," that reflect auditory rather than visual properties. Alternatively, it may also be the case that some parameters do not contribute to sound-symbolic mapping across a range of target domains. Testing across different domains might help to explain why this may be so and thus even non-relevant parameters could further our understanding of sound symbolism, albeit in a negative sense.

It might be objected that measures of voice quality were correlated with perceptual ratings because the speaker who recorded the pseudowords pronounced them differently according to her expectations of their roundedness/pointedness, researchers not being immune to, or unaware of, sound-symbolic mappings. We think this unlikely for several reasons. First, the speaker made a conscious effort to speak with neutral intonation and sound files were selected (from multiple takes) by two independent judges on the basis that they sounded both neutral and consistent with the other recordings. Since Parise and Pavani (2011) showed that people spontaneously vocalize differently to different stimulus attributes, the requirement to employ a neutral intonation may actually have reduced the true effect. Second, unlike other acoustic parameters such as amplitude or pitch, it would be hard to consciously modulate complex parameters such as shimmer or mean HNR. Even if this could be achieved, it is unlikely that it could be sustained over a set of more than 500 items in such a way as to produce the correlations seen in Fig. 6A, particularly when the items were recorded in random order rather than a fixed order along the rounded-to-pointed scales.

A drawback of our use of pseudowords is that they are not part of actual language although they were sampled from linguistic segments and conformed to the phonological constraints of standard American English (McCormick et al., 2015). Disadvantages of using real words, for example, mimetics or onomatopoeic words, include the loss of the control that we were able to command in using a carefully constructed stimulus set, or very small set sizes: For example, if one were to control for word length by choosing only two-syllable onomatopoeic words, set size would likely be diminished still further if one wanted a set of such words that all relate, as here, to a single dimension. However, the present study is exploratory, demonstrating the viability of RSA as a method and the importance of some acoustic and visual parameters but not others. Future work could proceed to examine these parameters in relation to real words indicating roundedness/pointedness, for example, "spike" versus "balloon" (Sučević et al., 2015), or to other kinds of shapes. Since the smaller set sizes for such words would entail smaller sets of shapes, the perceptual ratings of words and shapes could be carried out as a within-participant factor rather than, as here, a between-participant factor. This design aspect is a further limitation of the current study since it means that the pseudowords were never explicitly assigned to an actual shape. Thus, while we can reach some conclusions about sound symbolism, it is less easy to draw conclusions about the sound-shape crossmodal correspondence since the pseudowords and shapes were never explicitly compared by participants. This might be not problematic if the association between visual roundedness/pointedness and auditory pseudowords was, as seems likely, relative rather than absolute, as with the crossmodal correspondence between auditory pitch and visuospatial elevation (Spence, 2019). But the effect of these acoustic and visual parameters on the crossmodal correspondence could certainly now be tested further with smaller stimulus sets since the relationships between acoustic spectral parameters and global indices of visual shape may potentially underlie sound-to-shape mapping (see also Daube et al., 2019). In future work, we could more closely examine the relationship of the present work to sound-symbolic crossmodal correspondences by having people assign words to shapes and then examining the relationship between acoustic parameters and the visual properties of the shapes that people choose.

A final limitation is that, although we report the effects of the parameters individually, some parameters are likely interdependent either conceptually (e.g., both the pulse number and FUF reflect how often the vocal folds open and close or relative amount of voicing in an utterance) or computationally (e.g., the simple matching coefficient is a variation of the formula for the Jaccard distance—and both of these might be related to image silhouette even though the computation of the latter is different). Where different parameters are not independent of each other, it is hard to assess their unique contribution; however, it is unlikely that a single parameter is determinative of either auditory or visual roundedness/pointedness.

5. Conclusions

Our novel application of RSA on large sets of 537 auditory pseudowords and 90 visual shapes that were previously constructed and rated on the rounded-to-pointed dimension led to the following conclusions: (a) The auditory and visual ratings were closely interrelated, in keeping with the well-known crossmodal correspondence between auditory pseudowords and the roundedness or pointedness of visual shapes. (b) Global acoustic measures of the pseudowords, the speech envelope and spectral measures (spectral tilt and the temporal FFT), were related to the auditory ratings. For rounded compared to pointed pseudowords, the speech envelope and spectral power changes over the pseudoword were smoother, and spectral tilt was steeper with greater concentration in lower frequencies. (c) Multiple global indices of visual shape (the SMC, silhouette, image outlines, and Jaccard distance), but not their spatial FFT, were related to the visual ratings. (d) Among these acoustic and visual parameters that were related to the corresponding perceptual ratings, the acoustic spectral measures were crossmodally related to the global indices of visual shape. (e) While voice quality measures were not found to be related to the auditory ratings using RSA, many of them (the HNR, pulse number, FUF, mean autocorrelation, shimmer, and jitter) were shown by conventional analyses to be correlated with the auditory ratings; however, their potential relationship to relevant visual measures was not undertaken here. Overall, our findings extend those of previous studies (Chen et al., 2016; Knoeferle et al., 2017; Parise & Pavani, 2011) by providing new insights into the stimulus features that may mediate sound-symbolic crossmodal correspondences. Here, we show for the first time that the sound-symbolic mapping of sound to shape is related to acoustic properties of pseudowords. Further research is required to establish whether these factors contribute consistently across a range of sound-symbolic mappings or whether they are differently weighted across different mappings, and to understand their neural basis.

Acknowledgments

This work was supported by grants to K.S. and L.C.N. from the National Eye Institute at the NIH (R01EY025978) and the Emory University Research Council. Support to K.S. from the Veterans Administration and to S.M.L. from the Laney Graduate School, Emory University is also acknowledged. We thank Jee Young Kim and Valentin Lazar for their advice and assistance. An earlier version of this manuscript was released as a preprint at bioRxiv.org, doi: 10.1101.515581.

Authors’ contributions

S.M.L., K.M., S.L., K.S., and L.C.N. designed the research; K.M. created all the stimuli; S.M.L. and K.M. performed the research; S.M.L., Y.J., and S.L. analyzed the data; and S.L., S.M.L., K.M., Y.J., K.S., and L.C.N. wrote the paper.

Open Research badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/ekpgh and https://osf.io/y9zjc/.

Notes

1. Broadly speaking, vowels can be identified by their fundamental frequency (F0) and the relative frequencies of their formants—the resonance frequencies of the vocal tract when producing the vowel sound. The first three formants, F1–F3, are the most informative about vowel identity with higher formants contributing to speaker identity (Knoeferle et al., 2017).
2. Note that in order to avoid artificially inflating the degrees of freedom (df), the second-order correlations between matrices were calculated using one half of the off-diagonal data rather than the entire matrix: df is therefore given by $((n^2 - n)/2) - 2$, where n^2 gives the size of the matrix, $-n$ removes the diagonal cells, and dividing by 2 removes the redundant half of the cells, the matrices being symmetric across the diagonal.
3. We thank an anonymous reviewer for drawing this possibility to our attention. In this system, roundness is measured as the mean radius of the curvatures best fitting each of the outward "corners" divided by the radius of the largest inscribed circle, that is, the circle best fitting all the inward "corners" (Boggs, 2009; Folk, 1965).
4. We thank an anonymous reviewer for drawing this to our attention.

References

Ademollo, F. (2011). *The Cratylus of Plato: A commentary*. Cambridge, UK: Cambridge University Press.

Aiken, S. J., & Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing*, *29*, 139–157.

Akita, K., & Tsujimura, N. (2016). Mimetics. In T. Kageyama & H. Kishimoto (Eds.), *Handbook of Japanese lexicon & word formation* (pp. 133–160). Boston, MA: Walter de Gruyter Inc.

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics*, *57*, 1151–1162.

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 of the United States of America*, *113*, 10818–10823.

Boersma, P., & Weenink, D. (2012). PRAAT: Doing phonetics by computer. Accessed at: http://www.praat.org/. Accessed March 15, 2020.

Boggs Jr., S. (2009). *Petrology of sedimentary rocks* (2nd ed.). New York: Cambridge University Press.

Brand, J., Monaghan, P., & Walker, P. (2018). The changing role of sound symbolism for small versus large vocabularies. *Cognitive Science*, *42*(Suppl 2), 578–590.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape-sound matches, but different shape-taste matches to Westerners. *Cognition*, *126*, 165–172.

Brockmann, M., Drinnan, M. J., Storck, C., & Carding, P. N. (2011). Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *Journal of Voice*, *25*, 44–53.

Catricalà, M., & Guidi, A. (2015). Onomatopoeias: A new perspective around space, image schemas, and phoneme clusters. *Cognitive Processing*, *16*(Suppl 1), S175–S178.

Chen, Y.-C., Huang, P.-C., Woods, A., & Spence, C. (2016). When “bouba” equals “kiki”: Cultural commonalities and cultural differences in sound-shape correspondences. *Scientific Reports*, *6*, 26681. https://doi.org/10.1038/srep26681

Cuskley, C., Simner, J., & Kirby, S. (2017). Phonological and orthographic influences in the bouba–kiki effect. *Psychological Research*, *81*, 119–130.

Daube, C., Ince, R. A. A., & Gross, J. (2019). Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Current Biology*, *29*, 1924–1937.

Davico, G., Pizzolato, C., Killen, B. A., Barzan, M., Suwarganda, E. K., Lloyd, D. G., & Carty, C. P. (2019). Best methods and data to reconstruct paediatric lower limb bones for musculoskeletal modelling. *Biomechanics and Modeling in Mechanobiology*, in press. https://doi.org/10.1007/s10237-019-01245-y

Davis, R. (1961). The fitness of names to drawings: A cross-cultural study in Tanganyika. *British Journal of Psychology*, *52*, 259–268.

De Carolis, L., Marsico, E., Arnaud, V., & Coupé, C. (2018). Assessing sound symbolism: Investigating phonetic forms, visual shapes, and letter fonts in an implicit bouba-kiki experimental paradigm. *PLoS One*, *13*, e0208874. https://doi.org/10.1371/journal.pone.0208874

de Saussure, F. D. (2011). General principles: Nature of the linguistic sign. In I. P. Meisel & H. Saussy (Eds.), *Course in general linguistics* (W. Baskin, Trans.; pp. 65–70). New York: Columbia University Press.

Devaprakash, D., Lloyd, D. G., Barrett, R. S., Obst, S. J., Kennedy, B., Adams, K. L., Hunter, A., Vlahovich, N., Pease, D. L., & Pizzolato, C. (2019). Magnetic resonance imaging and freehand 3-D ultrasound provide similar estimates of free Achilles tendon shape and 3-D geometry. *Ultrasound in Medicine & Biology*, *45*, 2898–2905.

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *Journal of Neuroscience*, *33*, 18906–18916.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*, 6. https://doi.org/10.1167/10.1.6

Ferrand, C. T. (2002). Harmonics-to-noise ratio: An index of vocal aging. *Journal of Voice*, *16*, 480–487.

Folk, R. L. (1965). *Petrology of sedimentary rocks*. Austin, TX: Hemphill. Retrieved from the Walter Geology Library, University of Texas. https://web.archive.org/web/20060214063526/http://www.lib.utexas.edu/geo/folkready/folkprefrev.html

Fort, M., Martin, A., & Peperkamp, S. (2015). Consonants are more important than vowels in the bouba-kiki effect. *Language & Speech*, *58*, 247–266.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, *68*, 1191–1203.

Gasser, M. (2004). The origins of arbitrariness in language. *Proceedings of the Annual Meeting of the Cognitive Science Society Conference*, *26*, 434–439.

Hollien, H., Girard, G. T., & Coleman, R. F. (1977). Vocal fold vibratory patterns of pulse register phonation. *Folia Phoniatrica et Logopaedica*, *29*, 200–205.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Hornibrook, J., Ormond, T., & Maclagan, M. (2018). Creaky voice or extreme vocal fry in young women. *New Zealand Medical Journal*, *131*, 36–40.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130298.

Imai, M., Miyazaki, M., Yeung, H. H., Hidaka, S., Kantartzis, K., Okada, H., & Kita, S. (2015). Sound symbolism facilitates word learning in 14-month-olds. *PLoS One*, *10*, e0116494.

Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., & Hagita, N. (2008). A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, & Language Processing*, *16*, 47–56.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *37*, 547–579.

Jamal, Y., Lacey, S., Nygaard, L., & Sathian, K. (2017). Interactions between auditory elevation, auditory pitch, and visual elevation during multisensory perception. *Multisensory Research*, *30*, 287–306.

Joseph, J. E. (2015). Iconicity in Saussure's linguistic work, and why it does not contradict the arbitrariness of the sign. *Historiographia Linguistica*, *42*, 85–105.

Kambaskovic, D., & Wolfe, C. T. (2016). The senses in philosophy and science: From the nobility of sight to the materialism of touch. In H. Roodenburg (Ed.), *A cultural history of the senses in the Renaissance* (pp. 107–125). London: Bloomsbury Press.

Kliper, R., Portuguese, S., Weinshall, D. (2016). Prosodic analysis of speech and the underlying mental state. In S. Serino (Ed.), *Pervasive computing paradigms for mental health: MindCare 2015 selected papers* (pp. 52–62). Cham, Switzerland: Springer.

Knoeferle, K., Li, J., Maggioni, E., & Spence, C. (2017). What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports*, *7*, 5562. https://doi.org/10.1038/s41598-017-05965-y.

Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.

Köhler, W. (1947). *Gestalt psychology: An introduction to new concepts in modern psychology*. New York: Liveright.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*, 1126–1141.

Lacey, S., Martinez, M. O., McCormick, K., & Sathian, K. (2016). Synesthesia strengthens sound-symbolic cross-modal correspondences. *European Journal of Neuroscience*, *44*, 2716–2721.

Liew, K., Lindborg, P., Rodrigues, R., & Styles, S. J. (2018). Cross-modal perception of noise-in-music: Audiences generate spiky shapes in response to auditory roughness in a novel electroacoustic concert setting. *Frontiers in Psychology*, *9*, 178. https://doi.org/10.3389/fpsyg.2018.00178.

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in Psychology*, *6*, 1246. https://doi.org/10.3389/fpsyg.2015.01246

Lu, Y., & Cooke, M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, *51*, 1253–1262.

Marks, L. E. (1978). *The unity of the senses: Interrelations among the modalities*. New York: Academic Press.

Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 384–394.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental Science*, *9*, 316–322.

McCormick, K., Kim, J. Y., List, S., & Nygaard, L. C. (2015). Sound to meaning mappings in the bouba-kiki effect. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings 37th Annual Meeting Cognitive Science Society* (pp. 1565–1570). Austin TX, USA: Cognitive Science Society.

McCormick, K., Lacey, S., Stilla, R., Nygaard, L. C., & Sathian, K. (2018). Neural basis of the sound-symbolic crossmodal correspondence between auditory pseudowords and visual shapes. *bioRxiv*, https://doi.org/10.1101/478347

Meteyard, L., Stoppard, E., Snudden, D., Cappa, S. F., & Vigliocco, G. (2015). When semantics aids phonology: A processing advantage for iconic word forms in aphasia. *Neuropsychologia*, *76*, 264–275.

Mezzedimi, C., di Francesco, M., Livi, W., Spinosi, M. C., & De Felice, C. (2017). Objective evaluation of presbyphonia: Spectroacoustic study on 142 patients with *Praat*. *Journal of Voice*, *31*, 257.e25–257.e32.

Monaghan, P., Mattock, K., & Walker, P. (2012). The role of sound symbolism in language learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *38*, 1152–1164.

Namy, L. L., & Nygaard, L. C. (2008). Perceptual-motor constraints on sound to meaning correspondence in language. *Behavioral and Brain Sciences*, *31*, 528–529.

Nielsen, A., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology*, *65*, 115–124.

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition*, *112*, 181–186.

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*, 127–146.

Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, *114*, 173–186.

Parise, C. V., & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research*, *214*, 373–380.

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: A study using the implicit association test. *Experimental Brain Research*, *220*, 319–333.

Peiffer-Smadja, N., & Cohen, L. (2019). The cerebral bases of the bouba-kiki-effect. *NeuroImage*, *186*, 679–689.

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: From a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130300.

Revill, K. P., Namy, L. L., Defife, L. C., & Nygaard, L. C. (2014). Cross-linguistic sound symbolism and crossmodal correspondence: Evidence from fMRI and DTI. *Brain and Language*, *128*, 18–24.

Revill, K. P., Namy, L. L., & Nygaard, L. C. (2018). Eye movements reveal persistent sensitivity to sound symbolism during word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 680–698.

Ricotta, C., & Pavoine, S. (2015). Measuring similarity among plots including similarity among species: An extension of traditional approaches. *Journal of Vegetation Science*, *26*, 1061–1067.

Rogers, S., & Ross, A. (1975). A cross-cultural test of the maluma–takete phenomenon. *Perception*, *5*, 105–106.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, *12*, 225–239.

Schmidtke, D. S., Conrad, M., & Jacobs, A. M. (2014). Phonological iconicity. *Frontiers in Psychology*, *5*, 80. https://doi.org/10.3389/fpsyg.2014.00080

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh: Psychology Software Tools Inc.

Sidhu, D. M., & Pexman, P. M. (2015). What's in a name? Sound symbolism and gender in first names. *PLoS ONE*, *10*, e0126809. https://doi.org/10.1371/journal.pone.0126809

Singh, L. (2015). Speech signal analysis using FFT and LPC. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *4*, 1658–1660.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, and Psychophysics*, *73*, 971–995.

Spence, C. (2019). On the relative nature of (pitch-based) crossmodal correspondences. *Multisensory Research*, *32*, 235–265.

Styles, S. J., & Gawne, L. (2017). When does Maluma/Takete fail? Two key failures and a meta-analysis suggest that phonology and phonotactics matter. *iPerception*, *8*, 204166951772480. https://doi.org/10.1177/2041669517724807

Sučević, J., Savić, A. M., Popović, M. B., Styles, S. J., & Ković, V. (2015). Balloons and bavoons versus spikes and shikes: ERPs reveal shared neural processes for shape-sound-meaning congruence in words, and shape-sound congruence in pseudowords. *Brain & Language*, *145*(146), 11–22.

Teixeira, J. P., & Fernandes, P. O. (2014). Jitter, shimmer and HNR classification within gender, tones and vowels in healthy voices. *Procedia Technology*, *16*, 1228–1237.

Thompson, P. D., & Estes, Z. (2011). Sound symbolic naming of novel objects is a graded function. *Quarterly Journal of Experimental Psychology*, *64*, 2392–2404.

Thoret, E., Aramaki, M., Kronland-Martinet, R., Velay, J.-L., & Ystad, S. (2014). From sound to shape: Auditory perception of drawing movements. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 983–994.

Tzeng, C. Y., Duan, J., Namy, L. L., & Nygaard, L. C. (2018). Prosody in speech as a source of referential information. *Language, Cognition and Neuroscience*, *33*, 512–526.

Tzeng, C. Y., Nygaard, L. C., & Namy, L. L. (2016). The specificity of sound symbolic correspondences in spoken language. *Cognitive Science*, *41*, 2191–2220.

Tzeng, C. Y., Nygaard, L. C., & Namy, L. L. (2017). Developmental change in children's sensitivity to sound symbolism. *Journal of Experimental Child Psychology*, *160*, 107–118.

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice stress analysis: A new framework for voice and effort in human performance. *Frontiers in Psychology*, *9*, 1994. https://doi.org/10.3389/fpsyg.2018.01994

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synesthetic cross-modality correspondences. *Psychological Science*, *21*, 21–25.

Westbury, C., Hollis, G., Sidhu, D. M., & Pexman, P. M. (2018). Weighing up the evidence for sound symbolism: Distributional properties predict cue strength. *Journal of Memory & Language*, *99*, 125–150.

Whitehead, R. L., Metz, D. E., & Whitehead, B. H. (1984). Vibratory patterns of the vocal folds during pulse register phonation. *Journal of the Acoustical Society of America*, *75*, 1293–1297.

Wilkinson, F., Wilson, H. R., & Habak, C. (1998). Detection and recognition of radial frequency patterns. *Vision Research*, *38*, 3555–3568.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Fig. S1. Shape images were cropped to 200 × 200 pxiels to remove excessive background and arrive at the smallest area that contained all the shapes. The figure shows all 90 image outlines overlaid on one another; darker/lighter areas indicate the intersection of more/fewer outlines. There are few very dark intersections indicating that shape outlines rarely overlapped by much with other shapes and that therefore all the shapes were different to a large degree. This explains the high dissimilarity values displayed in the image outlines RDM in Fig. 7D in the main text.

Fig. S2. All 90 image silhouettes overlaid on one another; darker/lighter areas indicate more/less overlap between shapes. The dark central area indicates pixels that were common to all shapes.

Fig. S3. RDMs for auditory perceptual ratings of roundedness/pointedness (A) and for the acoustic parameters that were not significantly correlated with perceptual ratings after Bonferroni-correction (B–H); r = Spearman correlation coefficient, df = 151 in all cases.

Fig. S4. Pitch standard deviation was not significantly correlated with auditory perceptual ratings of roundedness/pointedness after Bonferroni-correction.

Fig. S5. RDMs for visual perceptual ratings of roundedness/pointedness (left) and for the spatial FFT (right) were not significantly correlated after Bonferroni-correction; *r* = Spearman correlation coefficient, df = 4,003.