ACOUSTIC PARAMETER COMBINATIONS UNDERLYING MAPPING OF AUDITORY PSEUDOWORD SOUNDS TO MULTIPLE DOMAINS OF MEANING: A MACHINE LEARNING APPROACH

G. Vinodh Kumar[1], Simon Lacey[1,2,3], Josh Dorsi[1], Lynne C. Nygaard[4] & K. Sathian[1,2,3]

Departments of Neurology[1], Neural & Behavioral Sciences[2], & Psychology[3], Penn State Health Milton S. Hershey Medical Center & Penn State College of Medicine, Hershey, PA 17033, USA.
Department of Psychology[4], Emory University, Atlanta, GA 30322, USA

Corresponding authors:

K. Sathian
Department of Neurology
Milton S. Hershey Medical Center
Penn State College of Medicine
Hershey, PA 17033-0859, USA
Tel: 717-531-1801
Fax: 717-531-0384
Email: ksathian@pennstatehealth.psu.edu

Lynne C. Nygaard
Department of Psychology
Emory University
College of Arts and Sciences
Atlanta, GA 30322, USA
Tel: 404-727-0766
Fax: 717-531-0384
Email: lnygaar@emory.edu

ABSTRACT

Sound symbolism, referring to the resemblance between the sound structure of words and their meaning, is commonly studied using auditory pseudowords. Companion studies across seven meaning domains demonstrated systematic relationships, varying by domain, between the perceptual ratings, phonetic features, and acoustic parameters of a set of 537 pseudowords (Lacey et al. 2024a, 2024b). Here we employed a k-nearest-neighbor (KNN) machine-learning algorithm to compare 4094 combinations of twelve acoustic parameters (3 spectro-temporal and 9 characterizing vocal quality) and identify the parameter combination that best predicted perceptual ratings in each domain. Using multiple regression, we then examined the relative contributions of the parameters comprising the best performing acoustic model for each domain. Finally, we used the KNN approach to generate sound-symbolic ratings, in the shape domain, for 160 real words and compared these predicted ratings with corresponding perceptual ratings. We found that sound-symbolic mappings rely on domain-specific combinations and weights of acoustic parameters. One spectro-temporal parameter, the fast Fourier transform, and one vocal parameter, the fraction of unvoiced frames, were both present in the best performing model for each meaning domain studied, indicating the general importance of these two parameters for sound-symbolic judgments. The predicted and perceptual ratings of the real words were strongly correlated, indicating the value of this approach to measure the degree of sound-symbolic mapping in natural languages, unconfounded by semantic bias. Our findings support the proposed relevance of sound symbolism to language.

1. INTRODUCTION

Interest in sound symbolism, that is, the non-arbitrary link between speech sounds and meaning, has been on the rise in the past two decades, challenging the traditionally held view of the 'arbitrariness of the sign' in natural languages (de Saussure, 1916/2009; Svantesson, 2017). Indeed, in addition to onomatopoeia, where the sound of the word imitates the sound it describes, sound-meaning mappings have been shown to extend to several non-auditory sensory domains like shape (Köhler, 1929; Ramachandran & Hubbard, 2001), size (Sapir, 1929), weight (Walker & Parameswaran, 2019), and brightness (Hirata et al., 2011); as well as to non-sensory, abstract, domains, such as valence (Uno et al., 2020), arousal (Lacey et al. 2024a), and social dominance (Auracher, 2017). Moreover, non-arbitrary correspondences between sound and meaning have been found at different developmental stages (Maurer et al., 2006; Asano et al., 2015), and in various linguistic, geographical and cultural contexts (Blasi et al., 2016). While converging lines of evidence establish the prevalence and robustness of sound-symbolism in natural languages, it remains unclear how various acoustic features combine to effectuate sound-meaning associations in different domains.

A major focus of previous research has been on the contributions of phonetic and articulatory features to sound symbolism (Sapir, 1929; Nuckolls, 1999; Nielsen & Rendall, 2011, 2013; Sidhu & Pexman, 2018). For instance, pseudowords containing the phonemes /i/ and /e/ connote smallness (Sapir, 1929), brightness (Johansson et al., 2020), and sharpness (Sidhu & Pexman, 2018), compared to pseudowords containing the vowels /o/ and /u/ that has been associated to roundedness (McCormick et al., 2015), heaviness (Lacey et al., 2024a) and social dominance (Auracher, 2017). Similarly, the incidence of certain consonants versus vowels (Fort et al., 2015), voiced versus unvoiced consonants (McCormick et al., 2015; Cuskley et al., 2017), obstruents versus sonorants (McCormick et al., 2015), or vowel formants (Knoeferle et al., 2017) underlie inherent biases in assigning meaning to certain sounds (e.g., rounded versus pointed and dark versus bright). In the first of two companion studies, perceptual ratings were shown to have domain-specific phonetic associations across seven sound-symbolic meaning domains (Lacey et al. 2024a). These associations are important since the sound structure of any language is defined by such phonemic features.

Apart from the significance of phonetic features in sound symbolism, it is important to examine their acoustic characteristics. Parameters obtained by acoustic analysis have the advantage of defining speech more fundamentally, enabling the comparison of results from studies across different languages and cultures (Knoeferle et al., 2017; Lacey et al., 2020). More importantly, the multidimensional space provided by different acoustic parameters offers a detailed characterization of speech sounds, compared to the higher-level characterization of phonetic and articulatory spaces (Tsur, 2006). Additionally, nonlinguistic elements that influence sound symbolism, e.g., the prosody of speech, which refers to the intonation, rhythm, and relative loudness of an utterance, are best represented in the acoustic parameters. Acoustic analysis is also relevant from a neural standpoint because neuroimaging and electrophysiological evidence indicate that cerebral cortical responses are sensitive to the temporal evolution and frequency content of speech sounds (Kubanek et al., 2013; Mesgarani et al., 2014). Thus, defining sound

symbolism in terms of acoustic parameters offers an alternative and potentially more comprehensive picture of its underpinnings, complementing analyses of phonetic features alone.

Few studies have examined the role of different acoustic features in sound symbolism, with a focus on sound-to-meaning mapping in one or two domains, or employing only a small set of pseudowords (Knoeferle et al., 2017; Lacey et al., 2020). Consequently, the relative importance of different acoustic parameters in mediating sound-symbolic mapping across various meaning domains and the wide spectrum of acoustic signatures, corresponding to different phonemes, remains unknown. In the second companion study, different acoustic parameters, measuring both spectro-temporal properties and voice quality, were found to underpin the perceptual ratings of a large number of pseudowords across a number of sound-symbolic meaning domains (Lacey et al., 2024b). However, the precise combinations and weights of these parameters relevant for sound-symbolic mapping in different meaning domains remains unclear.

A primary motivation to understand the mechanism(s) underlying sound-symbolic association is to determine its role in natural languages. Since most studies have employed pseudowords to investigate the importance of different parameters in mediating sound-symbolic associations, it is uncertain to what extent the findings from such studies may apply to the vocabularies of natural languages. Moreover, research of this kind is characterized by a diversity of methods. One approach has been to analyze whether real words conform to the findings for pseudowords, e.g., whether real words for small and large objects contain the same small-associated (high front vowels like /i/) and large-associated phonemes (low-back vowels like /ɑ/), respectively (Thorndike, 1945), that have been found in studies using pseudowords. The results of this approach have been equivocal: some studies found systematic relationships between vowels in words and their meaning (e.g.,Thorndike, 1945), while others did not (e.g., Katz, 1986). Another approach is the best-worst judgement format developed by Hollis (2018), in which participants, on each trial, are presented with a subset of items (e.g., six items) from a list and are instructed to select the two extreme items on a given dimension (e.g., most rounded and most pointed or softest and hardest) in the presented subset. While this method provides a quantitative score of word-meaning association, it is based on ranking of items in particular lists and is thus of limited generalizability.

The purpose of this study is threefold: first, to find the combination of acoustic parameters that carries the most relevant information for sound-symbolic associations in each meaning domain; second, to examine the relative weights of these parameters in mediating the associations; and third, to generate sound-symbolic ratings for a set of real words based solely on their acoustic properties, unconfounded by semantic bias. To these ends, we employed a *k*-nearest-neighbor (KNN) machine-learning approach to identify the combinations of acoustic parameters that best predicted the perceptual ratings obtained on seven meaning domains in the companion study of Lacey et al. (2024a). Our approach is based on the principle that, if the acoustic properties of pseudowords vary systematically in accordance to their degree of sound-symbolic association in a given meaning domain, e.g., shape, then similarly rated pseudowords should form clusters in a multidimensional acoustic parameter space. We used this premise to iteratively explore all possible multidimensional acoustic parameter spaces based on the parameters we studied, and

find the combinations, in each domain, for which similarly rated pseudowords formed the most coherent clusters, i.e., with minimal distances between such pseudowords. Finally, we extended this approach to generate predicted ratings of real words, assessing sound symbolism without the confounds of semantic processing.

2. METHODS

2.1. Participants

The current study used data from participants in a companion study (Lacey et al., 2024a). 389 native speakers of American English (171 males, 208 females, 6 nonbinary, 2 agender, 1 gender fluid and 1 unknown; mean age 27 years 1 month; SD 8 months) were recruited and compensated online via the Prolific participant pool (https://prolific.co). The Gorilla platform (https://gorilla.sc; Anwyl-Irvine et al., 2020) was used to obtain informed consent from participants and to administer the experimental procedures. All procedures were approved by the Emory University Institutional Review Board. For full details on recruitment and data collection, see Lacey et al. (2024a). Additionally, data from 24 monolingual speakers of American English (20 females, mean age 30 years, SD 7 years) recruited for a different experiment (Dorsi et al., 2023) were used in this study. These participants were emailed a link to the experiment on Pavlovia.org after they gave consent to participate over the telephone; this study was approved by the Penn State University Institutional Review Board.

2.2 Stimuli

We used two sets of audio recordings comprising (1) pseudowords, and (2) real words. 537 bisyllabic pseudowords of the form “consonant, vowel, consonant, vowel” (CVCV), generated using a subset of phonemes that occur in English and spoken by a female, native speaker of American English (McCormick et al., 2015), as described previously (Lacey et al., 2020; 2024a) were used. The mean duration of the pseudowords was 457 ms (SD ± 62 ms). We also chose 160 real words from Sidhu et al. (2021) who assessed the meanings of ~1700 English nouns on a rounded-to-pointed shape dimension. The words were chosen based on their meaning scores such that half came from the two extremes (40 each from the most rounded and the most pointed meaning scores) and the other half from around the median. These words were recorded by a different female native speaker of American English. The real words had a mean duration of 569 ms (SD ± 117 ms: Dorsi et al., 2023).

2.3. Perceptual rating tasks

As detailed by Lacey et al. (2024a), participants were randomly assigned to rate 537 pseudowords on one of two 7-point Likert type scales representing categorial opposites across seven meaning domains: shape (rounded/pointed, N =30/30 respectively), size (small/big, N = 32/31), brightness (bright/dark, N = 24/26), texture (hard/soft, N = 27/27), weight (light/heavy, N = 29/26), arousal (calming/exciting, N = 30/28), and valence (good/bad, N = 26/23). Here, we converted these categorically opposing scales in each meaning domain to a single scale by subtracting the ratings given to pseudowords on the roundedness, smallness, brightness,

hardness, lightness, excitement and goodness scales from 8. This operation rescaled the order of the ratings on these scales, i.e., ratings 1 to 7 became 7 to 1, concordant with the order on their corresponding opposing scales. This also resulted in one rescaled rating per meaning domain, in which the two ends of the 7-point Likert scale represented the extremes of the respective categorical opposites, e.g., for shape, the rescaled rating 1 denoted 'most rounded' and 7 denoted 'most pointed'.

For rating the 160 real words, 24 participants were randomly assigned to rate their sounds using one of two 7-point Likert type scales; either a 1 (very rounded) to 7 (very pointed) scale or the reverse. Thus, 12 participants provided ratings on each scale. For both tasks, each word was rated once and the order of word presentation was randomized for each participant.

2.4. Acoustic parameters

As described by Lacey et al. (2024b), we selected three spectro-temporal acoustic parameters representing the overall acoustic forms of the pseudowords: the fast Fourier transform (FFT), spectral tilt, and speech envelope, measuring the distributions of amplitude and frequency over time (Aiken & Picton, 2008; Singh, 2015); and nine parameters capturing the degree and regularity of voicing. These included the fraction of unvoiced frames (FUF), mean autocorrelation, and pulse number, indexing the proportion and characteristics of voiced segments (Boersma & Weenink, 2012); jitter, shimmer, pitch standard deviation (PSD) and harmonics-to-noise ratio (HNR), measuring the regularity of voicing or voice quality (Teixeira & Fernandes, 2014); mean pitch, i.e. the fundamental frequency of the speech sound (Kliper et al., 2016); and item duration. Detailed descriptions of the parameters are provided in Lacey et al. (2024b).

2.5. k-nearest neighbor classifier

A *k*-nearest neighbor (KNN) machine-learning algorithm was employed to find the set of acoustic parameters that best predicted the rescaled ratings for the 537 pseudowords in each of the seven meaning domains studied. The algorithm comprised five steps as described below.

2.5.1. Step 1: Data preparation

Each parameter extracted for a particular pseudoword served as a separate dimension on which pairwise distances were computed between pseudowords. The three spectro-temporal parameters comprised an array of values for each pseudoword; therefore, as described by Lacey et al. (2024b), pairwise distances between pseudowords were computed for these parameters by subtracting the pairwise correlations (r) from 1, resulting in dissimilarity indices ($1 - r$). This computation rescaled correlations, ranging from +1 to -1, to dissimilarity values, ranging from 0 to 2. The FFT and speech envelope for each pseudoword were represented by 10077 and 20154 values respectively; therefore, we used Pearson's correlations to compute pairwise distances for these parameters. However, the spectral tilt for each pseudoword was represented by only 8 data points; therefore, pairwise distances were computed using non-parametric Spearman's

correlations (Lacey et al., 2024b). Finally, the dissimilarity values for the spectro-temporal parameters were linearly rescaled to lie between 0 and 1. For the nine vocal parameters, each pseudoword was represented by a single value. Therefore, pairwise distances between pseudowords were computed by linearly rescaling the parameter values to lie between 0 and 1 and then taking the absolute difference between the corresponding values of each pseudoword pair.

2.5.2. Step 2: Selection of parameter sets for testing

Since the primary goal of the analyses was to identify the set of acoustic parameters that best predicted the ratings of the pseudowords, we varied the number of parameters from 1-12 and iteratively performed the subsequent steps with all 4094 possible parameter combinations.

2.5.3. Step 3: k-nearest neighbors

The rating for every pseudoword was iteratively predicted in each of the multidimensional parameter spaces defined by a particular combination of acoustic parameters. In each of these spaces, the predicted rating for a given pseudoword depends on the perceptual ratings of its nearest neighboring pseudowords. The distances between pseudowords were determined by Euclidean distances while neighborhood size was taken as the square root of the total number of pseudowords: $k = \sqrt{537} \approx 23$ (Duda et al., 2000). In each parameter space, the rating for a given pseudoword was predicted based on the modal ratings of its 23 nearest neighbors (for instance, if a given pseudoword's neighbors were most frequently rated as 3, then the given pseudoword was assigned the predicted rating 3). The procedure was repeated iteratively for every pseudoword using a leave-one-out procedure. It is important to note that, in each parameter space, the 23 nearest neighbors for a given pseudoword are invariant because the Euclidean distances between pseudowords are based on their acoustic features, whereas the ratings of these neighbors vary on a participant-by-participant basis. Thus, the rating for each pseudoword was predicted on a participant-wise basis in all 4094 parameter spaces (thus, for shape, N = 60; for size, 63; brightness, 50; texture, 54; weight, 55; arousal, 58; valence, 49).

2.5.4. Steps 4 and 5: Model evaluation and selection

In each meaning domain, participant-wise predicted ratings for 537 pseudowords were concatenated into a vector for each combination of acoustic parameters and correlated with the corresponding vectorized perceptual ratings. Because of the ordinal nature of the ratings, Spearman's correlation (ρ) was used. The combination of parameters yielding the highest ρ value for a given meaning domain was chosen as the optimal combination for predicting pseudoword ratings in that domain.

[INSERT FIGURE 1 ABOUT HERE]

2.6. Multiple regression

For each meaning domain, the acoustic parameters comprising the optimal combination in the KNN model were entered into multiple regression equations to quantify their relative contributions in predicting the ratings. Since the spectro-temporal parameters for each

pseudoword were represented by multiple values, they could not be entered directly as predictors in the regression equations. Therefore, we indirectly examined the influence of each parameter, both spectro-temporal and vocal, by predicting the pairwise dissimilarity of the predicted ratings using the corresponding pairwise dissimilarity index of the acoustic parameters as shown below:

$$Predicted_{ij} = \beta_1 . AP_{1_{ij}} + \beta_2 . AP_{2_{ij}} + \beta_3 . AP_{3_{ij}} \ \ldots \ + I + \epsilon \qquad \mathbf{1}$$

where $Predicted_{ij}$ represents the pairwise dissimilarity of predicted ratings between pseudowords *i* and *j*, generated by converting the correlation (Pearson's *r*) between participant-wise predicted ratings for each pair of pseudowords to a dissimilarity score (1-*r*), and then rescaling to lie between 0 and 1; $AP_{n_{ij}}$ represents the pairwise dissimilarity between acoustic parameters of the corresponding pseudowords (computed as described in section 2.5.1); β_n denotes the corresponding coefficients; and I and ϵ represent the interaction and error terms in the multiple regression model.

Prior to performing the regressions, we examined whether the pairwise dissimilarity values for each acoustic parameter (independent variables [IV]) entering the multiple regression analyses met the assumptions of collinearity. The variance inflation factor (VIF) for every acoustic parameter, computed by regressing each IV against every other IV in the multiple regression model, quantifies the extent to which the contribution of each acoustic parameter is inflated by its correlation with other parameters. Generally, VIF = 1 implies no correlation, 1<VIF<5 indicates moderate correlation, and VIF>5 suggests high correlation (Dodge, 2008; Everitt & Skrondal, 2010) between the IVs.

2.7 Cross-domain dissimilarity

We examined the dissimilarity between different domains of meaning based on the optimal combinations of acoustic parameters and their relative contributions in mediating the corresponding sound-symbolic associations. For each domain, we projected the relative weights of different parameters from the multiple regression onto a twelve-dimensional acoustic parameter space, resulting in a multidimensional vector for each of the seven meaning domains. We evaluated the dissimilarity between these vectors by converting the cosine similarity between these vectors to a dissimilarity value. In a multidimensional vector space, cosine similarity is based on the angular separation between the vectors and their magnitude. Cosine similarity between any two vectors was computed using the following formula:

$$cosine\ similarity = \frac{\sum_i^n \boldsymbol{A_i B_i}}{\sqrt{\sum_i^n \boldsymbol{A_i^2}}\sqrt{\sum_i^n \boldsymbol{B_i^2}}} \qquad \mathbf{2}$$

where A_i and B_i are the relative weights (β-coefficients) of different acoustic parameters of vectors $\overline{\boldsymbol{A}}$ and $\overline{\boldsymbol{B}}$.

Finally, the cosine similarity was converted to a cosine dissimilarity index using:

$$cosine\ dissimilarity = 1 - cosine\ similarity \qquad \mathbf{3}$$

Since cosine similarity ranges from +1 to -1, the operation in equation 3 rescales the dissimilarity to range between 0 and 2, where 0 refers to the most similar, and 2 to the most dissimilar vectors.

2.8. Shape ratings of real words

As a proof of concept, we sought to apply our approach to predict the shape ratings of real words without the influence of semantic knowledge. To achieve this, we used the perceptual ratings of pseudowords in the acoustic neighborhood of each real word. By applying the KNN approach (detailed in section 2.5), we identified the optimal parameter combination for predicting the shape ratings of pseudowords and determined the 23 pseudowords ($k = \sqrt{538} \approx 23$; 537 pseudowords and 1 real word) in the acoustic neighborhood of each real word.

The parameter combination defined the multidimensional space in which we iteratively computed the Euclidean distance of each real word from all 537 pseudowords. The nearest pseudoword neighbors of a given real word were then determined based on the Euclidean distance, and the modal rating of these nearest neighbors was taken as the predicted rating for that real word. Prior to computing the Euclidean distance of each real word from every pseudoword, corresponding dissimilarity indices were computed following the procedure elaborated in section 2.5.1. Focusing on the parameter combination that best predicted the shape ratings of pseudowords, we computed dissimilarity indices only for those parameters. Finally, we computed the mean predicted ratings across participants and examined the Pearson correlation between these mean predicted ratings and the mean perceptual ratings across items.

3. RESULTS

3.1. Combinations of acoustic parameters mediating sound-symbolic associations

Different combinations of acoustic parameters best predicted the pseudoword ratings for different domains of meaning. Overall, the FFT and FUF featured in all the seven combinations, followed by spectral tilt and duration which were present in the best combinations for six meaning domains. The speech envelope occurred in five combinations, while mean pitch and HNR featured in two meaning domains each. Finally, PSD featured only once - in the best parameter combination for valence (**Figure 2**).

Acoustic parameters comprising the optimal combination for each meaning domain were entered into multiple regression equations to quantify their relative contributions in predicting the ratings. Prior to that, we ensured that these acoustic parameters (IV) met the assumptions of collinearity (see section 2.6). Diagnostic tests indicated that multicollinearity was not a concern for any of the seven domains. Following Menard (2002), tolerance values were all above 0.2, with observed values 0.45-0.99 (**Supplementary Tables 1A-G: Collinearity diagnostics**). The variance inflation factors (VIF) of all the parameters across the seven meaning domains (1.05-2.07) were well below the threshold value of 10 recommended by Myers (1986). Furthermore, the mean VIFs across meaning domains (1.10-1.45) were all close to 1, as recommended by Bowerman & O'Connell (1990) (**Supplementary Tables 1A-G: Collinearity diagnostics**).

The following sections outline, for each domain, the optimal combination defined by our KNN model and their relative influence in predicting the ratings. The full regression results are reported in Supplementary tables 1A-G.

[INSERT FIGURE 2 ABOUT HERE]

3.1.1. Shape

A combination of six parameters including all three spectro-temporal parameters: FFT, spectral tilt, and speech envelope; and three vocal parameters: FUF, shimmer, and duration; comprised the optimal model that best predicted (r_s = 0.59) participants' (N = 60) perceptual ratings of the 537 pseudowords in the shape domain (**Figure 2**).

The multiple regression analysis indicated that the regression equation explained 44.2% of the variance ($F_{(6,143909)}$ =18995.85, p<0.001) (**Supplementary** T**able 1A**), where each of the six parameters significantly predicted the shape ratings in the order: spectral tilt, FFT, duration, speech envelope, FUF, and shimmer (**Figure 3A**).

3.1.2. Size

Participants' (N = 63) perceptual ratings of the pseudowords on the size domain were best predicted by a combination of five parameters (r_s = 0.45) comprising all three spectro-temporal parameters: FFT, spectral tilt, and speech envelope; and two voice parameters: FUF and duration (**Figure 2**).

When these five parameters were entered into the regression, the coefficients of spectral tilt, FFT, FUF, and duration were significant, in this order, explaining 18.1% of the variance

($F_{(5,143910)}$ = 6339.87, p<0.001) (Figure 3A and **Supplementary** T**able 1B**). The coefficient of the speech envelope was not significant (β = 0.0001, p < 0.98) in the regression equation indicating that it does not contribute materially to predicting the ratings of the pseudowords in this domain.

3.1.3. Brightness

A combination of five parameters including all three spectro-temporal parameters: FFT, spectral tilt, and speech envelope; and two voice parameters: FUF and mean pitch; constituted the optimal model that best predicted (r_s = 0.53) the perceptual ratings (N = 50) of brightness of the pseudowords (**Figure 2**).

The multiple regression equation explained 18.9% of the variance ($F_{(5,143910)}$ = 6686.14, p<0.001) (**Supplementary** T**able 1C**), where each of the parameters significantly predicted brightness ratings in the order: FFT, spectral tilt, FUF, speech envelope and mean pitch (**Figure 3A**).

3.1.4. Texture

The optimal combination that best predicted perceptual ratings of the participants (N = 54) in the texture (hard/soft) domain consisted of five acoustic parameters (r_s = 0.54) which included two spectro-temporal parameters: FFT and spectral tilt; and three voice parameters: FUF, shimmer, and duration (**Figure 2**).

Multiple regression analysis showed that the regression equation explained 38.9% of the variance ($F_{(5,143910)}$ = 18321.51, p<0.001) (**Supplementary** T**able 1D**), and all five parameters significantly predicted theratings in the order: FFT, FUF, spectral tilt, shimmer, and duration (**Figure 3A**).

3.1.5. Weight

A combination of six parameters including the three spectro-temporal parameters: FFT, spectral tilt, and speech envelope; and three voice parameters: FUF, HNR, and duration, comprised the optimal model that best predicted participants' (N = 55) perceptual ratings (r_s = 0.45) of pseudowords in the weight domain (**Figure 2**).

Multiple regression analysis in this meaning domain showed that the regression equation explained 44.4% of the variance ($F_{(6,143909)}$ =18995.85, p<0.001) (**Supplementary** T**able 1E**), and all six parameters significantly predicted the ratings in the order: FFT, HNR, spectral tilt, FUF, duration and speech envelope (**Figure 3A**).

3.1.6. Arousal

The optimal combination that best predicted participants' (N = 58) perceptual ratings of arousal for the pseudowords comprised seven acoustic parameters (r_s = 0.51), including two spectro-temporal parameters: FFT and spectral tilt; and five voice parameters: FUF, shimmer, HNR, mean pitch, and duration (**Figure 2**).

The multiple regression equation explained 32.6% of the variance ($F_{(7,143908)}$ = 9930.56, p<0.001) (**Supplementary** T**able 1F**), and all seven parameters significantly predicted the ratings in the order: FFT, HNR, spectral tilt, FUF, duration, shimmer, and mean pitch (**Figure 3A**).

3.1.7. Valence

The KNN algorithm identified an optimal combination of six parameters comprising two spectro-temporal parameters: FFT and speech envelope; and four vocal parameters: FUF, HNR, PSD, and duration, that best predicted (r_s =.66) the perceptual ratings (N = 49) of pseudowords in the valence domain (**Figure 2**).

Multiple regression analysis used to evaluate the relative contributions of these parameters showed a significant regression equation that explained 23.1% of the variance ($F_{(6,143909)}$ = 7191.64, p<0.001) (**Supplementary** T**able 1G**). We found that all six parameters significantly predicted the ratings in the following order: FFT, HNR, FUF, duration, PSD, and speech envelope (**Figure 3A**).

3.2. Cross-domain relationships

Figure 3A illustrates that the relative contributions of the various acoustic parameters clearly differed across the domains studied. It is apparent that the FFT was a major predictor of ratings for each domain, and the strongest predictor for all except the shape domain, for which spectral tilt was the strongest predictor. The FFT and FUF were among the relevant acoustic parameters for all seven meaning domains, while spectral tilt and pseudoword duration were present in the combinations for all but two domains (valence and brightness, respectively). Conversely, PSD was relevant only for the valence domain, while jitter, mean autocorrelation and pulse number did not contribute to the ratings on any domain.

Figure 3B illustrates cosine dissimilarities between domains, computed based on the relative weights of acoustic parameters in each domain. Some interesting patterns of differences emerged. For example, the sensory domains of shape and size were quite dissimilar from the abstract domains of valence and arousal (see **Figure 3B and Supplementary Tables 2 & 3**), with shape and valence exhibiting the highest dissimilarity. Within the sensory domains, shape and size, both of which can be assessed visually as well as haptically, were similar to each other, but dissimilar to texture (hard/soft) and weight, both of which are assessed purely haptically and which were themselves similar to each other. Brightness, a purely visual property, was most similar to size, followed by weight and texture (hard/soft). Within the sensory domains, shape and texture were most dissimilar. The pairwise dissimilarity between weight and the abstract domains was very low, as was that between the abstract domains of valence and arousal.

[INSERT FIGURE 3 ABOUT HERE]

3.3. Predicting sound-symbolic shape ratings of real words

We used the multidimensional space defined by the six acoustic parameters that best predicted the shape ratings of the pseudowords (see section 3.1.1) to predict the shape ratings of real words. Each real word's predicted rating was derived from the modal perceptual rating of 23 pseudowords within its acoustic neighborhood. With 60 participants rating the pseudowords, each real word's rating was predicted 60 times, and the mean of these ratings was used as the predicted rating of the real words. The mean predicted ratings were strongly correlated ($r_p = 0.52, p < 0.001$) with mean perceptual ratings of the real words.

4. DISCUSSION

Behavioral, developmental, and neuroimaging studies have widely employed pseudowords to demonstrate the phenomenon of sound symbolism and to study the cognitive processes that underlie sound-symbolic associations; however, the mechanisms that underpin sound-symbolism remain unknown. Here, employing a supervised machine-learning approach, we show that sound-symbolic associations in seven different meaning domains are mediated by different combinations of acoustic parameters, the relative contributions of these parameters varying by meaning domain. Furthermore, as a proof of concept, we used the same approach to predict the sound-symbolic shape ratings of real words based entirely on their acoustic properties, unconfounded by semantic bias.

Acoustic parameter combinations relevant to sound-symbolic ratings of pseudowords

Prior studies investigating acoustic properties that mediate sound symbolism suggest that the influence of these properties varies by semantic domain (Knoeferle et al., 2017, Lacey et al. 2020, 2024b). To our knowledge, there are no published reports attempting to evaluate the combinations of acoustic parameters that are optimally predictive of sound-symbolic associations, and the relative contributions of individual parameters within such combinations. We reasoned that if a set of sound-symbolic pseudowords were projected onto a multidimensional acoustic space, defined by the optimal combination of acoustic parameters for a particular domain of meaning, then pseudowords in that space would tend to form coherent clusters, i.e., pseudowords with similar degrees of sound-symbolic associations would be close to each other. Based on this premise, we employed a KNN machine-learning approach and found that varying combinations of acoustic parameters predicted the perceptual ratings of pseudowords in seven different meaning domains. Further, we used multiple regression analyses to quantify the relative contributions of the relevant parameters, and demonstrated that these also varied by domain.

Shape

We found that shape ratings were best predicted by a combination of six parameters including all three spectro-temporal parameters: the FFT, spectral tilt, and the speech envelope; and three voice parameters: the FUF, shimmer and duration (**Figure 2**). In the combination, the spectro-temporal parameters carried an overall higher influence in predicting the ratings compared to the voice parameters (**Figure 3A and supplementary table 1A**). Interestingly, the relative contributions of the spectro-temporal parameters in predicting perceptual ratings, as indexed by their beta weights in the regression, were concordant with the relative effect sizes of the correlations between the representational dissimilarity matrices (RDMs) of the spectro-temporal parameters and the ratings demonstrated by Lacey et al. (2020, 2024b). However, while Lacey et al. (2024b) found significant correlations of the ratings with eight vocal parameters, the optimal combination only featured three, measuring the duration, voicing (FUF) and amplitude variation (shimmer) of the utterances.

Among the spectro-temporal parameters, we observed that spectral tilt carried the highest influence, followed by the FFT, whose beta value was two-thirds of that for the spectral tilt, while the speech envelope had the least influence, with a beta value only a quarter of that for the

spectral tilt. Existing studies show 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), indicating the significance of formant frequencies in mediating sound-shape relationships. Each formant is associated with a specific frequency range and is characterized by a concentration of spectral power around those frequencies (Fisher et al., 2018). These formant properties fit with the higher influence of spectral tilt compared to the FFT, because spectral tilt is primarily sensitive to the distribution of power across different frequencies, e.g., rounded/pointed words have power concentrated at lower/higher frequencies resulting in a steeper/flatter spectral tilt (Lacey et al. 2020, 2024b), whereas the FFT reflects the power spectrum across the duration of the spoken pseudowords (Lacey et al., 2024b), capturing the instantaneous spectral profile as the pseudoword unfolds in time.

The distinction between rounded and pointed pseudowords is primarily based on broad phonetic categories, such as sonorants versus obstruents. However, within these categories, consonants with specific voicing, place and manner of articulation appear to be more influential in predicting roundedness or pointedness (Lacey 2024a). Sonorants and obstruents exhibit distinct spectro-temporal patterns (Lacey et al. 2024b). Broad distinctions are effectively represented in spectral tilt, while finer distinctions related to voicing, place and manner of articulation are thoroughly captured by the FFT. This may be why spectral tilt carries the highest weight, followed by the FFT.

Among the vocal parameters, we found that duration carried the highest weight, even higher than that of the speech envelope, with a beta weight whose absolute value for duration was half that for the spectral tilt. This might stem from the fact that the duration of the speech signal carries information related to the component vowels (Hillenbrand et al., 1995; Kluender et al., 1988) that influence sound-shape judgments, e.g., rounded (e.g., o) and unrounded vowels (e.g., I) that influence roundedness and pointedness, respectively, have distinctive temporal profiles. Additionally, the type of vowels that precede or follow a consonant can influence the duration of coarticulated phoneme or syllable, thereby affecting the overall duration of the pseudoword. Robust sound-shape associations have been shown between sonorants (e.g., /l/, /m/, /n/) and round shapes, and voiceless stops (/p/, /t/, /k/) and pointed shapes (McCormick et al., 2015), supporting the importance of the FUF that characterizes features related to voicing in speech. However, the relative weighting of the speech envelope was a bit higher than that of the FUF, perhaps because these sonorants and obstruents have distinctive temporal profiles (Lacey et al., 2020, 2024b). The presence of shimmer in the model is consistent with the importance of the amplitude profile represented by the speech envelope, as shimmer indexes the temporal variation in amplitude. Nevertheless, shimmer exerts about two-thirds of the psychoacoustic influence of the speech envelope, implying that momentary amplitude variations are less important than the overall amplitude profile in predicting the ratings. Parameters representing frequency variability, e.g., jitter and PSD, and periodicity, e.g., HNR, pulse number and mean autocorrelation, did not contribute materially to predicting ratings. This is discussed below.

Size

Our KNN algorithm showed that a combination of five parameters including three spectro-temporal parameters: the FFT, spectral tilt and the speech envelope, and two voice parameters: the FUF and duration, best predicted sound-symbolic judgments of size (**Figure 2**). However, our regression analysis using the parameters in the optimal combination showed that the speech envelope was not a significant contributor in predicting the ratings (**Figure 3A and supplementary table 1B**). Spectral tilt and the FFT carried comparable influence, while the influence of FUF was nearly two-thirds of that for the spectro-temporal parameters. Duration carried the least influence which was roughly one-third of that for FFT. In general, the frequency of the vocalizations produced by animals is related to their size, i.e., animals with larger bodies produce vocalizations with lower frequencies, and vice versa (Fitch, 1997; Gingras et al., 2013; Bowling et al., 2017). It has been argued that experience of such co-occurring attributes (e.g. body size and frequency profile of the vocalization) in the environment leads to internalization of these probabilities and might underlie sound-symbolic associations (Sidhu & Pexman, 2018). Furthermore, small/big ratings were found to be associated with unvoiced/voiced consonants (Lacey et al., 2024a), which are characterized by distinctive spectral profiles (Lacey et al. 2024b). Thus, a pseudoword's spectral characteristics outweigh other features in relation to sound-symbolic judgments of size, accounting for the higher influence of the FFT and spectral tilt which both encode the spectral characteristics of the utterances. Furthermore, broad categorization of sonorants and obstruents predicted small/big ratings. However, unlike for the shape domain, voicing, place and manner of articulation of consonants were equally significant in predicting the smallness/bigness ratings of pseudowords. Broad phonetic distinctions (sonorants versus obstruent) are well represented in spectral tilt, while finer distinctions related to voicing, place and manner of articulation are thoroughly captured by FFT (Lacey et al. 2024b). Therefore, both FFT and spectral tilt carry comparable influence in predicting the size ratings.

Anikin et al. (2021) found that vocal roughness exaggerates the apparent body size of the vocalizer. Since the proportion of unvoiced segments is indicative of vocal roughness, the FUF is an important parameter. Nevertheless, the influence of FUF was less than that of the spectro-temporal parameters. This might be related to how well these parameters characterize the relevant phonetics. While FUF is sensitive to the broad phonetic features (e.g. obstruents, sonorants) of an utterance, the spectro-temporal parameters carry more nuanced information capable of distinguishing phonetic characteristics within a category (e.g. stops and affricates). Previously, Knoeferle et al. (2017) showed that the duration of vowels in pseudowords influence sound-size judgements, however, with the current set of pseudowords Lacey et al. (2024a) found that bigness, but not smallness, ratings were associated with the vowels. Thus, pseudoword duration, which is partly contingent on the component vowels, is relevant for only one scale of the ratings. This may explain why it carried the least influence, among the significant parameters, in predicting the sound-size ratings of the pseudowords. Again, parameters representing frequency variability, e.g., jitter and PSD; amplitude variability, e.g., shimmer; periodicity, e.g., HNR, pulse number and mean autocorrelation; and fundamental frequency, e.g., mean pitch, did not feature in the combination.

Brightness

We found that the combination of parameters that best predicted the perceptual ratings of brightness comprised the three spectro-temporal parameters: the FFT, spectral tilt and the speech envelope, and two voice parameters: the FUF and mean pitch (**Figure 2**). Among the spectro-temporal parameters, the FFT and spectral tilt carried comparable influence, with slightly more weight for the FFT, whereas the speech envelope was less influential, its beta weight being about a third of that for the FFT. The FUF carried nearly half the influence of that for spectral tilt while the mean pitch carried the least influence, less than a third of that for the speech envelope (**Figure 3A and Supplementary Table 1C**). Notably, the relative importance of the FFT, spectral tilt and the FUF was similar to that in the size domain. Lacey et al. (2024a) found that, similar to the size domain, consonant voicing and place of articulation were key determinants of the brightness ratings of pseudowords. These detailed characteristics are captured by the FFT, explaining its high influence. Additionally, consonant voicing affects spectral tilt, e.g., voiced consonants have more energy in the lower frequencies, while the unvoiced consonants have more energy in the higher frequencies (Lacey et al. 2024b). Therefore, it has an influence comparable to the FFT. The speech envelope exerted an influence on brightness ratings comparable to that of the FUF. This might be related to the fact that the temporal evolution of pseudowords reliably captures information related to the rounding of the vowels, which is important for brightness ratings (Lacey et al. 2024a). Furthermore, information related to voicing, represented by FUF, is crucial because the key difference between pseudowords rated as bright and dark is the presence of unvoiced and voiced consonants (Lacey et al., 2024a). The contribution of the mean pitch is consistent with the finding of Hubbard (1996) that participants match high-pitched sounds to lighter, and low-pitched sounds to darker, stimuli. Nevertheless, the mean pitch carried a very small influence in predicting brightness ratings, perhaps because the pseudowords were spoken with a neutral intonation by a single speaker, therefore attenuating sound-symbolic attributes of the mean pitch. Pseudoword duration does not seem to influence brightness ratings, even though bright/dark pseudowords were characterized by the presence of unrounded/rounded vowels, like in shape. This might be because the durational cues provided by the combination of vowels and consonants that influence bright/dark ratings are not as significant as the cues related to voicing. For example, while the combination of unvoiced obstruents and front unrounded vowels (bright) could result in a different duration compared to the combination of voiced sonorants and back rounded vowels (dark), the primary effect would be due to consonant voicing. This is because unvoiced/voiced consonants could influence the voicing of the surrounding vowels. Here, the unvoiced obstruents diminish the voicing of the front unrounded vowels and sonorants accentuate the voicing of the back rounded vowels, thus, making the voicing relevant than duration Once again, parameters related to frequency variability, e.g., jitter and PSD, amplitude variability, e.g., shimmer and periodicity, e.g., HNR, pulse number and mean autocorrelation, did not contribute to predicting the sound-brightness ratings.

Texture

The combination of features represented in five parameters, including two spectro-temporal parameters: the FFT and spectral tilt, and three voice parameters: the FUF, shimmer, and duration, best predicted the perceptual ratings in the texture (hardness) domain (**Figure 2**). Interestingly, we found that the FFT had almost 2.5 times as much influence as the next most

important parameter, the FUF, while the relative influence of each succeeding parameter (in order of beta weights) was one-half that of its predecessor (**Figure 3A and Supplementary Table 1D**). The place of articulation of the sonorants significantly predicts softness ratings, while the voicing, manner and place of articulation are key predictors of hardness ratings (Sakamoto & Watanabe, 2018; Sidhu et al., 2022; Lacey et al., 2024a). These features have very distinct spectro-temporal profiles, effectively captured by the FFT, giving it the highest influence. While these finer details are crucial for sound-texture associations, hard- and soft-rated pseudowords can also be distinguished by the broad phonetic categories of obstruents and sonorants (Lacey et al. 2024a), respectively, which are represented by spectral tilt. Consequently, spectral tilt is also influential. Furthermore, the utterance of an obstruent is characterized by discontinuous and turbulent airflow with major constrictions to the vocal tract, and the FUF reflects such changes, although this parameter was measured as a single value over the utterance and thus lacks the finer-grained temporal information of the FFT, accounting for the substantially greater effect of the latter. The contributions of shimmer and duration are consistent with prior studies showing that noisy and punctuated sounds with abrupt and highly variable onsets, involving voiceless plosives (e.g. /t/ and /k/) and bilabial plosives (e.g. /p/ and /b/) are associated with the tactile perception of softness (Sakamoto & Watanabe, 2018). Such phonemes are characterized by distinctive temporal variations in amplitude, measured by shimmer, and by durational cues. Duration carries very little influence perhaps because the vowels, which constrain the duration of pseudowords, have very low influence in this domain. As for other domains, parameters that indicate frequency variability, e.g. jitter and PSD; the amplitude profile, i.e. the speech envelope; and periodicity, e.g. HNR, pulse number and mean autocorrelation; and fundamental frequency, e.g., mean pitch, did not contribute to predicting sound-texture ratings.

Weight

We found that the perceived heaviness of pseudowords was mediated by a combination of six parameters including the three spectro-temporal parameters: the FFT, spectral tilt and the speech envelope, and three voice parameters: the HNR, the FUF and duration. Overall, the FFT had the highest influence, followed by the HNR and spectral tilt, each exerting an effect that was half of that for the FFT. The FUF had about a third the influence of that of the FFT, followed by the speech envelope and duration, whose influences were an order of magnitude lower (**Figure 3A and Supplementary Table 1E**). Interestingly, we observed that, except for HNR, all other parameters in the model were the same as for the size domain, although the relative influence of these parameters differed in the two domains. Notably, Lacey et al. (2024b) found that ‘heavy’ ratings for the pseudowords were correlated with their ‘big’ ratings, and ‘light’ with ‘small’. The shared acoustic parameters might stem from the fact weight and size typically co-vary, i.e. big objects are generally heavier than small objects. Furthermore, Lacey et al. (2024a) showed that the combinations of vowel rounding and the consonant features that influence light and heavy ratings are unique. For example, light ratings are influenced by the presence of unrounded vowels along with unvoiced bilabial and labiodental stops, while heavy ratings were reliant on back rounded vowels along with voiced post alveolar/velar stops and a/fricatives. Labiodental sounds (e.g., /f/) involve turbulent airflow, resulting in lower HNR. In contrast, alveolar (e.g.,

/d/) and velar (e.g., /g/) sounds, produced by the tongue contacting the alveolar ridge and soft palate respectively, have a more periodic nature and less turbulent airflow. Since HNR indicates the proportion of periodic and aperiodic components in speech, it significantly influences weight ratings of pseudowords. However, the importance of the HNR is less than that of the FFT, because the latter provides detailed information reflecting syllable changes within a phonetic category (e.g., voiced and unvoiced sonorants). Sonorants and obstruents that determine the lightness and heaviness of the pseudowords, respectively, have been shown to exhibit distinct spectral tilts (Lacey et al. 2024b), making the influence of spectral tilet comparable to that of the HNR. A characteristic distinction between these two classes of phonemes (i.e., sonorants and obstruents) is also their voicing. While FUF is sensitive to changes in voicing, the speech envelope also carries information related to the voiced and unvoiced segments in the utterance. However, for the weight domain, we found that the FUF carries considerably higher influence than the speech envelope. This indicates that information related to the proportion of voiced segments in the utterance is more crucial than the temporal evolution of the pseudoword in sound-weight judgements. Duration carried relatively a small influence which might be reflective of the component vowels. But the acoustic significance of those vowels appears to be less than that of sonorants and obstruents. As for other domains, parameters that feature frequency variability, e.g. jitter and PSD; amplitude variability, e.g. shimmer; and periodicity, e.g. HNR, pulse number and mean autocorrelation; and fundamental frequency, e.g., mean pitch, did not contribute to predicting sound-weight ratings.

Arousal

A combination of seven parameters comprising two spectro-temporal parameters: the FFT and spectral tilt, and five voice parameters including the HNR, the FUF, shimmer, mean pitch and duration, best predicted the perceptual ratings in the arousal domain (**Figure 2**). Among these, the FFT exerted the highest influence, followed by the HNR and spectral tilt, each carrying nearly one-half of the influence of that of the FFT. The FUF and duration carried comparable influence, about a third of that of the FFT. Shimmer and mean pitch carried the least influence that was an order of magnitude less than that of the other parameters. (**Figure 3A and Supplementary Table 1F**). Pseudowords rated as calming comprise sonorants and back rounded vowels while those rated as exciting primarily comprise unvoiced obstruents and front unrounded vowels (Sidhu et al. 2022; Lacey et al. 2024a). Spectro-temporal features that differentiate sonorants from obstruents, as well as those related to voicing, are effectively captured by the FFT. Consequently, the FFT exerts the greatest influence. While spectral tilt captures the broad distinction in the spectral profile of sonorants and obstruents that affect calming and exciting ratings, information related to voicing and place of articulation are more pronounced in the FFT, therefore spectral tilt has a lesser influence compared to FFT. Affective states induced by stress have been shown to influence the HNR in speech produced during these states (Kappen et al., 2022). Additionally, the presence of obstruents increases turbulent airflow at the glottis during phonation, leading to aperiodic vocal fold vibration, which impacts the HNR. Consequently, its influence is comparable to spectral tilt. Voicing of obstruents underlie exciting ratings, explaining the significance of the FUF in the model. However, the FUF may not reliably encode differences between kinds of obstruents, e.g., stops,

af/fricatives, in contrast to the FFT. This might be why the FUF is less important than the FFT. Vowel rounding carries significant durational cues with rounded and unrounded vowels having distinct duration, explaining the relevance of duration in the model. Calming/exciting pseudowords are marked by sonorants/obstruents, with corresponding smooth/jagged temporal profiles (Lacey et al. 2024b). Shimmer effectively captures these distinctions making it a crucial parameter. Although the speech envelope also contains this information, it also includes additional information related to temporal evolution of the utterance which could possibly diminish or obscure the cues that distinguish calming and exciting pseudowords. Finally, the relevance of pitch is supported by evidence that the vocal fundamental frequency, represented by the mean pitch, correlates with arousal states (Rodero, 2022; Rodero et al., 2017). The reason it is least influential here may be because the pseudowords were uttered by a single speaker with neutral intonation or variation in pitch. Yet again, parameters that feature frequency variability, e.g. jitter and PSD; amplitude profile, i.e. the speech envelope; and periodicity, e.g. pulse number and mean autocorrelation; did not contribute to predicting the sound-weight ratings.

Valence

A combination of features represented in two spectro-temporal parameters: the FFT and the speech envelope, and four vocal parameters: the HNR, the FUF, duration and PSD, best predicted the valence ratings of the pseudowords (**Figure 2**). Among these, the FFT carried the highest influence, followed by the HNR, whose beta weight was a bit over half of that for the FFT. The FUF and duration carried equal influence, one-fourth of that for the FFT, while PSD and the speech envelope carried the least influence, one-eighth of that for the FFT (**Figure 3A and Supplementary Table 1G**). Notably, except for PSD and the speech envelope, the acoustic parameters in the optimal combination for valence also featured in that for arousal. Moreover, the shared parameters in both domains also carried comparable relative weights. This might be related to the fact that the degree of valence (good or bad) associated to the external stimulus determines the intensity of arousal. Furthermore, it is striking that, similar to the weight and arousal domains, the primary characteristic distinguishing the pseudowords representing the categorical opposites in the valence domain is also the incidence of sonorants/obstruents (Lacey et al, 2024a). Interestingly, unlike the arousal domain that featured shimmer as a contributor to ratings, the speech envelope was relevant to the valence domain, suggesting that information encoded in the overall temporal profile of the sound facilitates valence ratings. Finally, the presence of PSD in the combination is supported by the finding of Rodero (2022) that pitch variations influence participants' self-report of valence. This domain is the only one in which spectral tilt does not feature. This might be because the combination of consonants and vowels that mediate good (sonorants + front unrounded vowels) and bad (obstruents + back rounded vowels) ratings of pseudowords likely ambiguates the otherwise distinct cues for sonorants and obstruents represented in spectral tilt. For example, sonorants and obstruents typically have power concentrated in the lower and higher frequencies, respectively, while unrounded and rounded vowels tend to have higher power in the higher and lower frequencies, respectively. This overlap obfuscates the distinction between good and bad pseudowords based on spectral tilt. And once more, parameters that feature frequency variability, e.g. jitter; amplitude variability, e.g. shimmer; and periodicity, e.g., pulse number and mean autocorrelation; did not contribute to predicting the arousal ratings.

Similarities and differences between domains

We found that the best-performing combinations of acoustic parameters for different meaning domains had notable similarities and differences (**Figure 2**). Interestingly, the FFT and spectral tilt were amongst the top three parameters accounting for the most variance across the seven domains, suggesting that changes in a pseudoword's spectral profile over time and the distribution of its power across different frequencies play vital roles underpinning sound-symbolic associations. These features are significant from a neural standpoint because the initial stages of speech perception involve processing the spectro-temporal properties represented by these two parameters (Nourski, 2017; Khalighinejad et al., 2021). This indicates that spectro-temporal signatures encode rich information that is fundamental for non-arbitrary sound-to-meaning associations spanning multiple meaning domains.

Measures of vocal fold vibration and durational cues represent information fundamental for distinguishing phonological voicing (Hualde et al., 2015). We observed that the FUF, which measures the percentage of time windows that do not engage the vocal folds, predicted sound-symbolic ratings across all domains (**Figure 2**). The FUF provides cues about the phonemic content of an utterance (Boersma & Weenink, 2012), being lower if the utterance includes exclusively voiced elements like sonorants and vowels, and higher if it includes elements like unvoiced obstruents. Such phonemic cues are critical in forming sound-symbolic associations (McCormick et al., 2015) and therefore, sound-symbolic judgments across multiple domains are sensitive to features reflected in this parameter.

We found that stimulus duration featured in six domains. Our findings extend and generalize those of Knoeferle et al. (2017) demonstrating the significance of durational cues in sound-size symbolism. Most existing evidence shows how the phonological features that comprise a word determine the nature of its sound-symbolic association. Interestingly, the vowels and consonants that represent these phonological features are characterized by distinctive time profiles, with vowel duration being modulated by the consonant environment (Crystal & House, 1988; Chen, 2009) and duration influencing the perception of vowel quality (Hillenbrand et al., 2000). Taken together, the duration of a pseudoword is related to the phonological characteristics comprising the spoken word, and thus provides relevant cues for making sound-symbolic judgements.

None of the optimal combinations for the seven domains featured mean autocorrelation, pulse number and jitter. This might be because these parameters are less diagnostic of acoustic-phonetic properties that differentiate meaning domains. For instance, both FUF and pulse number represent how much phonation or voicing occurs in the pseudoword. However, FUF would provide a comparative measure of unvoiced to voices periods while the pulse number simply reports the number of glottal pulses found in the signal. The relationship of a feature such as voicing to particular meaning domains may depend on the relative degree of voicing. For our current set of pseudowords, we speculate that FUF might encode more crucial information related to voicing for making sound-symbolic judgments compared to pulse number. Similarly, for the current set of pseudowords, it is possible that critical information related to periodicity and variability, including the relative stability of vocal fold vibrations is carried by shimmer, PSD, HNR as compared to mean autocorrelation and jitter. Both mean autocorrelation and jitter

indicates the regularity in the frequency domain of voiced segments and may be less diagnostic than other measures of voicing. Whether the relevance of these analogous parameters is a function of the pseudoword set or the speaker requires further investigation.

Cross-domain dissimilarity

Even though the best combinations of acoustic parameters predicting ratings for different meaning domains comprised common parameters in some cases, each of the combinations was unique, and each domain was characterized by a unique regression equation in terms of the acoustic parameters contributing to ratings (**Figure 3A**). This offers support to the idea that sound-symbolic associations in specific meaning domains stem from features encoded not in individual parameters, but by the unique combinations and relative strengths of acoustic parameters. We quantified the degree of this uniqueness by evaluating how dissimilar the domains were in terms of the unique acoustic parameters and their relative weights. Interestingly, the patterns of cross-domain dissimilarity appeared to be broadly organized based on the sensory modality related to processing the attributes represented by the meaning domains (**Figure 3B**). For instance, the pairwise dissimilarities between the domains of shape and size were relatively low compared to their pairwise dissimilarity with texture (hard/soft). While an object's shape and size can be gauged haptically, these properties are typically first assessed visually (Pasupathy et al., 2018; Julian et al., 2017; Meese & Baker, 2023). Conversely, to gain in-depth information about hardness or softness (object compliance), we primarily rely on the sense of touch (Kim et al., 2021). This pattern of dissimilarity suggests that sound-symbolic associations may rely on modal simulations, i.e., mental recreations of our sensorimotor experiences, and perhaps sound features required to generate simulations (of items, e.g. rounded shape, big object) for less dissimilar modalities (e.g., arousal and valence) are represented in similar acoustic parameters.

We observed that brightness was more similar to weight than texture. This may be related to the fact that the prominent perceptual dimensions for brightness and weight represent variations of intensity (Purves et al., 2004). Brightness was also found to be more similar to weight compared to shape. We speculate that internalization of the brightness-weight illusion in which darker/brighter objects are perceived as heavier/lighter may underlie this discrepancy (Walker et al., 2010). Interestingly, we found that the pairwise dissimilarity between weight and both the abstract domains: valence and arousal, was low compared to its dissimilarity with other domains. This may be grounded in cross-domain perceptual/physiological relationships as shown by a study where participants associated valence-related emotions like happiness/anger to bodily lightness/heaviness (Hartmann et al., 2023). Analogous cross-domain associations were also observed by Lacey et al. (2024a) between valence and weight in which the sound-symbolic ratings for good/bad were related to light/heavy, and between arousal and weight in which lightness was correlated positively to calming and negatively to exciting.

Sound-symbolic ratings of real words

Extending our KNN approach, we demonstrated that the predicted shape ratings for real words, derived from pseudowords in their acoustic neighborhoods, were strongly correlated with their perceptual ratings. Although corpus-linguistic studies demonstrate instances of *iconicity* in

natural languages, where a word's form somehow bears resemblance to its meaning, the vast majority of these studies only examine whether iconicity exists in the first place (Winter et al., 2017, 2023). The ratings obtained this way are not divided into specified meaning domains and more importantly, these ratings are prone to influence by subjective semantic knowledge. Conversely, our approach allows us to use pseudowords to first narrow down the most relevant acoustic parameters that aid in sound-symbolic associations in a given meaning domain, and then, based on those acoustic parameters, to find the nearest pseudoword neighbors for each real word. The real word rating is determined by the mode of the perceptual ratings of these nearest pseudoword neighbors. Following this procedure allowed us to apply the principles emerging from pseudowords to generate ratings for real words, free of any semantic bias.

Conclusions

Our KNN-based approach represents a principled way to empirically identify the acoustic parameters underlying sound-symbolic judgements. The approach depends on the premise that, in certain parameter spaces, data points belonging to similar categories are closer to each other than data points that belong to different categories, analogous to our natural ability to form categories and classify items based on specific sets of features (Eimas & Quinn, 1994; Deng & Sloutsky, 2015). In addition, the operational principle of the KNN algorithm makes our approach transparent to the underlying relationships learned by the acoustic parameter models, as opposed to the general opacity of widely employed machine-learning approaches like support vector machines and deep learning methods.

Above all, the most significant contribution of our approach is that it offers a way to employ the principles of sound-symbolic associations, deduced from pseudowords, to measure the degree of sound symbolism in natural languages. Notably, since the measures are based solely on the acoustic features that drive sound-symbolic associations in pseudowords, they are not influenced by the conventional meaning of the words in that language. This has potential implications in the study of language development and applications in the field of consumer psychology.

Research suggests that during the initial stages of acquiring a vocabulary, sound symbolism may provide an effective, even essential, scaffold that facilitates early acquisition of words in that vocabulary (Imai & Kita, 2014; Kantartzis et al., 2011). Interestingly, in Japanese, sound-symbolic words, especially those which refer to action, are used more in speech by and toward children, and children are sensitive to the meanings of sound-symbolic verbs (Imai et al., 2008). In light of such evidence, our findings offer avenues for studying the impact of different acoustic features, fundamental for sound-symbolic associations in a particular domain, that affect the learning of vocabulary of that domain. In addition, given the significance of sound symbolism in early language development, a potential remedy for children facing difficulty in language learning could involve adjusting the inputs given to them, to include more sound-symbolic words. To this end, our approach could potentially aid in finding a sound symbolically rich vocabulary that could guide parents and caregivers of children facing difficulty in language learning. Similar considerations apply to rehabilitation of acquired language difficulties, e.g. in post-stroke aphasia (Meteyard et al., 2015).

Research in consumer psychology suggests that sound-symbolic brand names that can convey information about a product, e.g. its size, strength, weight, etc., can enhance awareness and create a favorable image of the product, even in the absence of a marketing communication (Klink, 2000, 2001). However, these studies have examined sound-symbolic effects of brand names by varying limited, mostly broadly contrasting categories, e.g. rounded vs unrounded vowels for naming big- and small-sized products respectively (Yorkston & Menon, 2004). Our findings provide a roadmap for making more nuanced variations to brand names by modulating the acoustic parameters most relevant to mediating sound-symbolic effects for given meaning domains, e.g. based on our results in the shape and size domains, a rounder or bigger product can be most effectively expressed by a longer name that has higher power in the lower frequencies.

In sum, our findings indicate that different acoustic parameters underlie sound-symbolic associations in different domains. In addition, we present an approach that can be used to assess sound symbolism in natural languages, unconfounded by semantic bias. Further research can extend our findings and methods to understand the roles of these acoustic parameters in natural language learning and processing. In addition, future neuroimaging and electrophysiological studies can provide insight into how the patterns represented in these acoustic parameters are represented neurally, which could pave the way to understanding the neural basis of sound symbolism.

ACKNOWLEDGEMENTS
This work was supported by grants to KS and LCN from the National Eye Institute at the NIH (R01EY025978) and the Emory University Research Council. Earlier versions of these data were presented at the 2022 meeting of the Cognitive Neuroscience Society (Hoffmann et al., 2022; Nygaard et al., 2022).

Author contributions: SL, KS, LCN and GVK designed research; JD collected data; GVK analyzed data; GVK, SL, KS and LCN wrote the paper.

DATA AVAILABILITY
Stimuli are available at https://osf.io/ekpgh/ and rating data and scripts for the spectro-temporal analyses are available at https://osf.io/y9zjc/.

References

Aiken, S. J., & Picton, T. W. (2008). Human cortical responses to the speech envelope. *Ear and Hearing*, *29*(2), 139–157. https://doi.org/10.1097/AUD.0B013E31816453DC

Anikin, A., Pisanski, K., Massenet, M., & Reby, D. (2021). Harsh is large: Nonlinear vocal phenomena lower voice pitch and exaggerate body size. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1954). https://doi.org/10.1098/rspb.2021.0872

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., & Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. *Cortex*, *63*, 196–205. https://doi.org/10.1016/j.cortex.2014.08.025

Auracher, J. (2017). Sound iconicity of abstract concepts: Place of articulation is implicitly associated with abstract concepts of size and social dominance. *PLoS ONE*, *12*(11), 1–25. https://doi.org/10.1371/journal.pone.0187196

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*(39), 10818–10823. https://doi.org/10.1073/pnas.1605782113

Bowling, D. L., Garcia, M., Dunn, J. C., Ruprecht, R., Stewart, A., Frommolt, K. H., & Fitch, W. T. (2017). Body size and vocalization in primates and carnivores. *Scientific Reports 2017 7:1*, *7*(1), 1–11. https://doi.org/10.1038/srep41070

Crystal, T. H., & House, A. S. (1988). The duration of American-English stop consonants: an overview. *Journal of Phonetics*, *16*(3), 285–294. https://doi.org/10.1016/s0095-4470(19)30503-0

Cuskley, C., Simner, J., & Kirby, S. (2017). Phonological and orthographic influences in the bouba–kiki effect. *Psychological Research*, *81*(1), 119–130. https://doi.org/10.1007/s00426-015-0709-2

Deng, W., & Sloutsky, V. M. (2015). *Inference Training on Category Representation*. *51*(3), 392–405. https://doi.org/10.1037/a0038749.

Dorsi, J., Lacey, S., Nygaard, L., & Sathian, K. (2023, November). *Acoustic stimulus parameters underlying sound symbolism in words and pseudowords*. Presented at the Annual Meeting of the Society for Neuroscience, Washington D.C., USA

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification by Richard O. Duda, David G. Stork, Peter E.Hart .pdf* (p. 738). http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471056693.html

Eimas, P. D., & Quinn, P. C. (1994). Studies on the Formation of Perceptually Based Basic-Level Categories in Young Infants. *Child Development*, *65*(3), 903–917. https://doi.org/10.1111/J.1467-8624.1994.TB00792.X

Fisher, J. M., Dick, F. K., Levy, D. F., & Wilson, S. M. (2018). Neural representation of vowel formants in tonotopic auditory cortex. *NeuroImage*, *178*(February), 574–582. https://doi.org/10.1016/j.neuroimage.2018.05.072

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America*, *102*(2 Pt 1), 1213–1222. https://doi.org/10.1121/1.421048

Fort, M., Martin, A., & Peperkamp, S. (2015). Consonants are More Important than Vowels in the Bouba-kiki Effect. *Language and Speech*, *58*(2), 247–266. https://doi.org/10.1177/0023830914534951

Gingras, B., Boeckle, M., Herbst, C. T., & Fitch, W. T. (2013). Call acoustics reflect body size across four clades of anurans. *Journal of Zoology*, *289*(2), 143–150. https://doi.org/10.1111/J.1469-7998.2012.00973.X

Hartmann, M., Lenggenhager, B., & Stocker, K. (2023). Happiness feels light and sadness feels heavy: introducing valence-related bodily sensation maps of emotions. *Psychological Research*, *87*(1), 59–83. https://doi.org/10.1007/s00426-022-01661-3

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*(6), 3013–3022. https://doi.org/10.1121/1.1323463

Hirata, S., Ukita, J., & Kita, S. (2011). Implicit phonetic symbolism in voicing of consonants and visual lightness using Garner's speeded classification task. *Perceptual and Motor Skills*, *113*(3), 929–940. https://doi.org/10.2466/15.21.28.PMS.113.6.929-940

Hollis, G. (2018). Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior Research Methods*, *50*(2), 711–729. https://doi.org/10.3758/s13428-017-0898-2

Hualde, J. I., Eager, C. D., & Nadeu, M. (2015). Catalan voiced prepalatals: Effects of nonphonetic factors on phonetic variation? *Journal of the International Phonetic Association*, *45*(3), 243–267. https://doi.org/10.1017/S0025100315000031

Hubbard, T. (1996). *Synesthesia-like mappings of lightness, pitch, and melodic interval*. Am J Psychol. https://pubmed.ncbi.nlm.nih.gov/8644886/

Humphries, C., Liebenthal, E., & Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage*, *50*(3), 1202–1211. https://doi.org/10.1016/j.neuroimage.2010.01.046.Tonotopic

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*(1651). https://doi.org/10.1098/rstb.2013.0298

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, *109*(1), 54–65. https://doi.org/10.1016/j.cognition.2008.07.015

Johansson, N., Anikin, A., & Aseyev, N. (2020). Color sound symbolism in natural languages. *Language and Cognition*, *12*(1), 56–83. https://doi.org/10.1017/langcog.2019.35

Julian, J. B., Ryan, J., & Epstein, R. A. (2017). Coding of Object Size and Object Category in Human Visual Cortex. *Cerebral Cortex*, *27*(6), 3095–3109. https://doi.org/10.1093/cercor/bhw150

Kantartzis, K., Imai, M., & Kita, S. (2011). Japanese sound-symbolism facilitates word learning in English-speaking children. *Cognitive Science*, *35*(3), 575–586. https://doi.org/10.1111/j.1551-6709.2010.01169.x

Kappen, M., van der Donckt, J., Vanhollebeke, G., Allaert, J., Degraeve, V., Madhu, N., Van Hoecke, S., & Vanderhasselt, M. A. (2022). Acoustic speech features in social comparison: how stress impacts the way you sound. *Scientific Reports*, *12*(1), 1–10. https://doi.org/10.1038/s41598-022-26375-9

Katz, A. N. (1986). Meaning conveyed by vowels: Some reanalyses of word norm data. *Bulletin of the Psychonomic Society*, *24*(1), 15–17. https://doi.org/10.3758/BF03330490

Khalighinejad, B., Patel, P., Herrero, J. L., Bickel, S., Mehta, A. D., & Mesgarani, N. (2021). Functional characterization of human Heschl's gyrus in response to natural speech. *NeuroImage*, *235*. https://doi.org/10.1016/j.neuroimage.2021.118003

Kim, J. H., Kim, J., Yeon, J., Park, J. Y., Chung, D., & Kim, S. P. (2021). Neural correlates of tactile hardness intensity perception during active grasping. *PeerJ*, *9*, 1–19. https://doi.org/10.7717/peerj.11760

Klink, R. R. (2000). Creating Brand Names with Meaning: The Use of Sound Symbolism. *Marketing Letters*, *11*(1), 5–20. https://doi.org/10.1023/A:1008184423824

Klink, R. R. (2001). Creating meaningful new brand names: a study of semantics and sound symbolism. *Journal of Marketing Theory and Practice*, *9*, 27–34. https://doi.org/10.2307/40470030 (2001)

Kliper, R., Portuguese, S., & Weinshall, D. (2016). Prosodic analysis of speech and the underlying mental state. *Communications in Computer and Information Science*, *604*, 52–62. https://doi.org/10.1007/978-3-319-32270-4_6/COVER

Kluender, K. R., Diehl, R. L., & Wright, B. A. (1988). Vowel-length differences before voiced and voiceless consonants: an auditory explanation. *Journal of Phonetics*, *16*(2), 153–169. https://doi.org/10.1016/s0095-4470(19)30480-2

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 2017 7:1*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-05965-y

Kohler, W. (1929). *Gestalt Psychology*. New York: Liveright. https://www.biblio.com/gestalt-psychology-by-kohler-dr-wolfgang/work/2062021

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The Tracking of Speech Envelope in the Human Cortex. *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0053398

Lacey, S., Jamal, Y., List, S. M., McCormick, K., Sathian, K., & Nygaard, L. C. (2020). Stimulus parameters underlying sound-symbolic mapping of auditory pseudowords to

visual shapes. *Cognitive Science*, *44*(9), e12883. https://doi.org/10.1111/COGS.12883

Lacey, S., Matthews, K.L., Sathian, K. & Nygaard, L.C. (2024a). Phonetic underpinnings of sound symbolism across multiple domains of meaning. bioRxiv 2024.09.03.610970; doi: https://doi.org/10.1101/2024.09.03.610970

Lacey, S., Matthews, K.L., Hoffmann, A.M., Sathian, K. & Nygaard, L.C. (2024b). Differential contributions of the spectro-temporal and vocal characteristics of auditory pseudowords to multiple sound-symbolic mappings. bioRxiv 2024.09.03.610973; doi: https://doi.org/10.1101/2024.09.03.610973

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental Science*, *9*(3), 316–322. https://doi.org/10.1111/j.1467-7687.2006.00495.x

McCormick, K., Kim, J. Y., List, S., & Nygaard, L. C. (2015). Sound to Meaning Mappings in the Bouba-Kiki Effect. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, *November 2017*, 1565–1570.

Meese, T. S., & Baker, D. H. (2023). Object Image Size Is a Fundamental Coding Dimension in Human Vision: New Insights and Model. *Neuroscience*, *514*, 79–91. https://doi.org/10.1016/j.neuroscience.2023.01.025

Menard, S. (2002). Applied Logistic Regression Analysis. *Applied Logistic Regression Analysis*. https://doi.org/10.4135/9781412983433

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

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. https://doi.org/10.1016/j.neuropsychologia.2015.01.042

Myers, R. H. (1986). *Classical and modern regression with applications*. 359. https://books.google.com/books/about/Classical_and_Modern_Regression_with_App.html?id=y2gZAQAAIAAJ

Nielsen, A. K. S., & Rendall, D. (2013). Parsing the role of consonants versus vowels in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology*, *67*(2), 153–163. https://doi.org/10.1037/a0030553

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*(2), 115–124. https://doi.org/10.1037/a0022268

Nourski, K. V. (2017). Auditory processing in the human cortex: An intracranial electrophysiology perspective. *Laryngoscope Investigative Otolaryngology*, *2*(4), 147–156. https://doi.org/10.1002/lio2.73

Nuckolls, J. B. (1999). The Case for Sound Symbolism. *Annu. Rev. Anthropol.*, *28*, 225–252.

Parise, C. V., & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research*, *214*(3), 373–380. https://doi.org/10.1007/S00221-011-2836-3/METRICS

Pasupathy, A., El-Shamayleh, Y., & Popovkina, D. V. (2018). Visual Shape and Object Perception. *Oxford Research Encyclopedia of Neuroscience*, *July*, 1–28. https://doi.org/10.1093/acrefore/9780190264086.013.75

Purves, D., Williams, S. M., Nundy, S., & Lotto, R. B. (2004). Perceiving the Intensity of Light. *Psychological Review*, *111*(1), 142–158. https://doi.org/10.1037/0033-295X.111.1.142

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia - A window into perception, thought and language. *Journal of Consciousness Studies*, *8*(12), 3–34.

Rodero, E. (2022). Effectiveness, Attractiveness, and Emotional Response to Voice Pitch and Hand Gestures in Public Speaking. *Frontiers in Communication*, *7*(June), 1–12. https://doi.org/10.3389/fcomm.2022.869084

Rodero, E., Potter, R. F., & Prieto, P. (2017). Pitch Range Variations Improve Cognitive Processing of Audio Messages. *Human Communication Research*, *43*(3), 397–413. https://doi.org/10.1111/hcre.12109

Sakamoto, M., & Watanabe, J. (2018). Bouba/Kiki in touch: Associations between tactile perceptual qualities and Japanese phonemes. *Frontiers in Psychology*, *9*(MAR), 1–12. https://doi.org/10.3389/fpsyg.2018.00295

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, *12*(3), 225–239. https://doi.org/10.1037/H0070931

Sidhu, D. M., & Pexman, P. M. (2018). Five mechanisms of sound symbolic association. *Psychonomic Bulletin and Review*, *25*(5), 1619–1643. https://doi.org/10.3758/s13423-017-1361-1

Sidhu, D. M., Vigliocco, G., & Pexman, P. M. (2022). Higher order factors of sound symbolism. *Journal of Memory and Language*, *125*(March 2021), 104323. https://doi.org/10.1016/j.jml.2022.104323

Sidhu, D. M., Westbury, C., Hollis, G., & Pexman, P. M. (2021). Sound symbolism shapes the English language: The maluma/takete effect in English nouns. *Psychonomic Bulletin and Review*, *28*(4), 1390–1398. https://doi.org/10.3758/S13423-021-01883-3/TABLES/5

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

Svantesson, J. O. (2017). Sound symbolism: the role of word sound in meaning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(5). https://doi.org/10.1002/wcs.1441

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. https://doi.org/10.1016/J.PROTCY.2014.10.138

Thorndike, E. L. (1945). On Orr'S Hypotheses Concerning the Front and Back Vowels. *British*

Journal of Psychology. General Section, *36*(1), 10–13. https://doi.org/10.1111/j.2044-8295.1945.tb01098.x

Tsur, R. (2006). Size-sound symbolism revisited. *Journal of Pragmatics*, *38*(6), 905–924. https://doi.org/10.1016/j.pragma.2005.12.002

Uno, R., Shinohara, K., Hosokawa, Y., Atsumi, N., Kumagai, G., & Kawahara, S. (2020). What's in a villain's name?: Sound symbolic values of voiced obstruents and bilabial consonants: *Review of Cognitive Linguistics*, *18*(2), 428–457. https://doi.org/10.1075/RCL.00066.UNO

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*(NOV), 1–25. https://doi.org/10.3389/fpsyg.2018.01994

Walker, P., Francis, B. J., & Walker, L. (2010). The brightness-weight illusion: Darker objects look heavier but feel lighter. *Experimental Psychology*, *57*(6), 462–469. https://doi.org/10.1027/1618-3169/a000057

Walker, P., & Parameswaran, C. R. (2019). Cross-sensory correspondences in language: Vowel sounds can symbolize the felt heaviness of objects. *Journal of Experimental Psychology: Learning Memory and Cognition*, *45*(2), 246–252. https://doi.org/10.1037/XLM0000583

Wallschläger, D. (1980). Correlation of song frequency and body weight in passerine birds. *Cellular and Molecular Life Sciences*, *36*(412). https://doi.org/:10.1007/BF01975119

Winter, B., Lupyan, G., Perry, L. K., Dingemanse, M., & Perlman, M. (2023). Iconicity ratings for 14,000+ English words. *Behavior Research Methods*, *March*. https://doi.org/10.3758/s13428-023-02112-6

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic? *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *18*(3), 443–464. https://doi.org/10.1075/is.18.3.07win

Yorkston, E., & Menon, G. (2004). A sound idea: Phonetic effects of brand names on consumer judgments. *Journal of Consumer Research*, *31*(1), 43–51. https://doi.org/10.1086/383422

Figure legends

Figure 1: Analysis flowchart for predicting pseudoword ratings based on acoustic parameters. Step 1: The pairwise distance between pseudowords was computed based on their differences in spectro-temporal and voice parameters, as detailed in the text. **Step 2.** The number of parameters entering the algorithm was varied from 1-12 and all 4094 combinations were iteratively tested. **Step 3.** For each combination, Euclidean distance was iteratively computed between each pseudoword and the remaining 536 pseudowords in a n-parameter space. The predicted ratings of each pseudoword were defined by the modal perceptual rating of its closest 23 neighbors ($k = \sqrt{537} \approx 23$). **Step 4.** Spearman correlations (r_s) between the perceptual and predicted ratings on a participant-by-participant basis was used to quantify the performance of the model. **Step 5.** The parameter combination that yielded the highest r_s was chosen as the best model for the corresponding meaning domain.

Figure 2: **Overlapping acoustic parameters across different meaning domains:** Each column corresponds to an acoustic parameter and each row to a meaning domain. Colored cells in each row signify the presence of an acoustic parameter. The histogram at the top shows the frequency of occurrence of each acoustic parameter in the optimal combinations across all meaning domains. The histogram to the right shows the number of acoustic parameters in the optimal combination for each meaning domain.

Figure 3: Cross-domain relationships: Heatmaps depicting (A) relative weights of different acoustic parameters in predicting the sound-symbolic ratings in each meaning domain. Beta weights of parameters that do not feature in a given meaning domain are represented by zeros and are depicted in off-white. (B) Cosine dissimilarity derived by subtracting cosine similarity from 1. Cosine similarity was computed based on the relative weights of the acoustic parameters that best predicted the ratings in each meaning domain.

Figure 1:

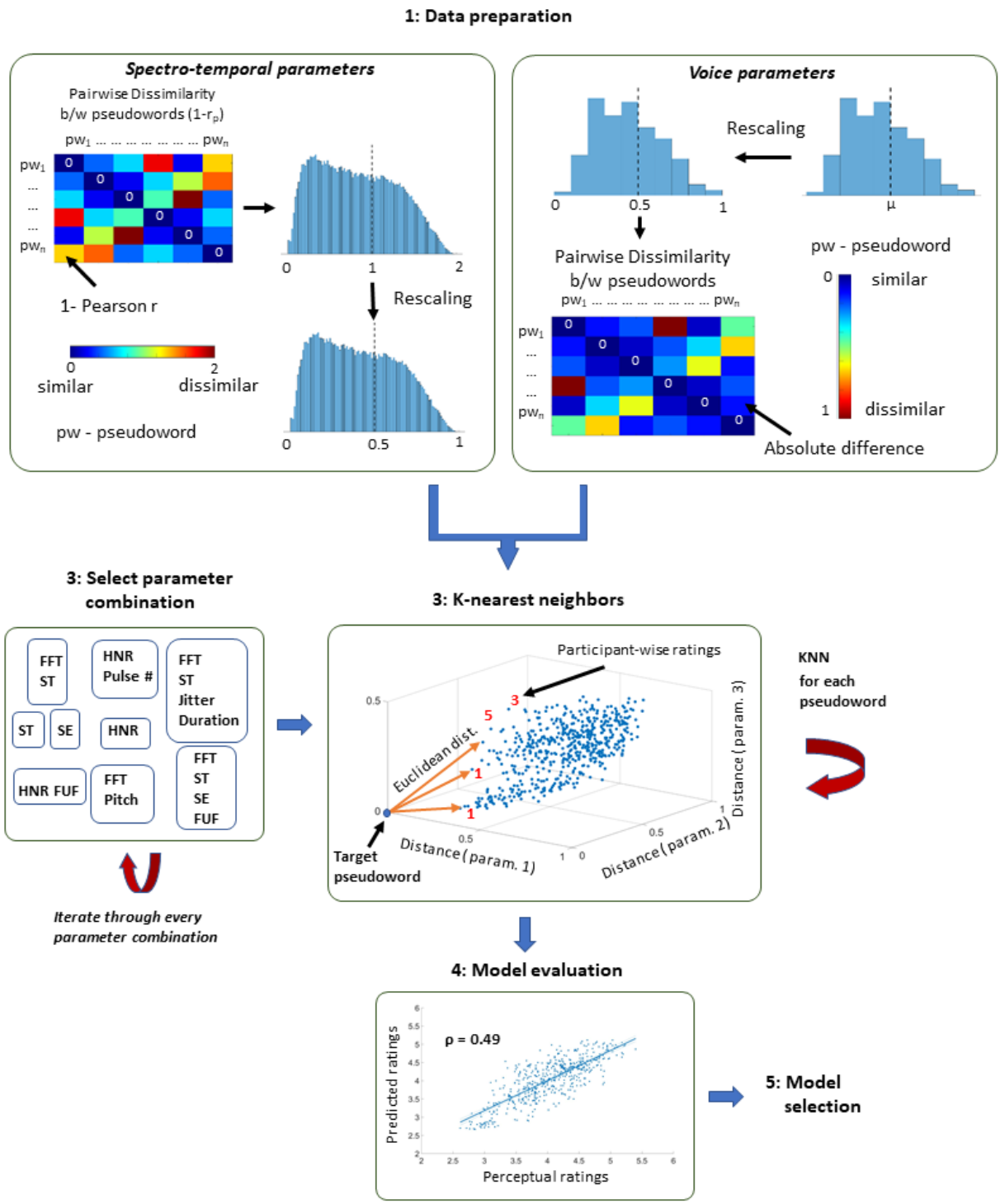

1: Data preparation
Spectro-temporal parameters
Pairwise Dissimilarity
b/w pseudowords (1-rp)
1- Pearson r
similar
dissimilar
pw - pseudoword
Rescaling
Voice parameters
Rescaling
Pairwise Dissimilarity
b/w pseudowords
pw - pseudoword
similar
dissimilar
Absolute difference
3: Select parameter combination
FFT ST
HNR Pulse #
FFT ST Jitter Duration
ST
SE
HNR
HNR FUF
FFT Pitch
FFT ST SE FUF
Iterate through every parameter combination
3: K-nearest neighbors
Participant-wise ratings
Euclidean dist.
Target pseudoword
Distance (param. 1)
Distance (param. 2)
Distance (param. 3)
KNN
for each pseudoword
4: Model evaluation
ρ = 0.49
Predicted ratings
Perceptual ratings
5: Model selection

Figure 2:

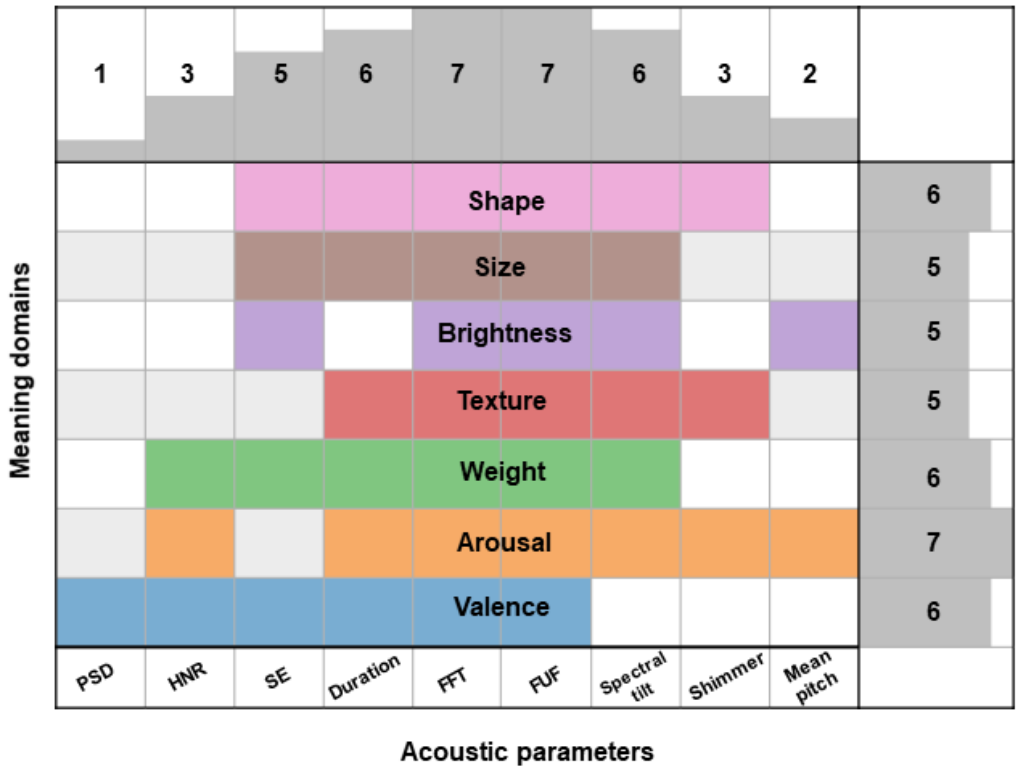
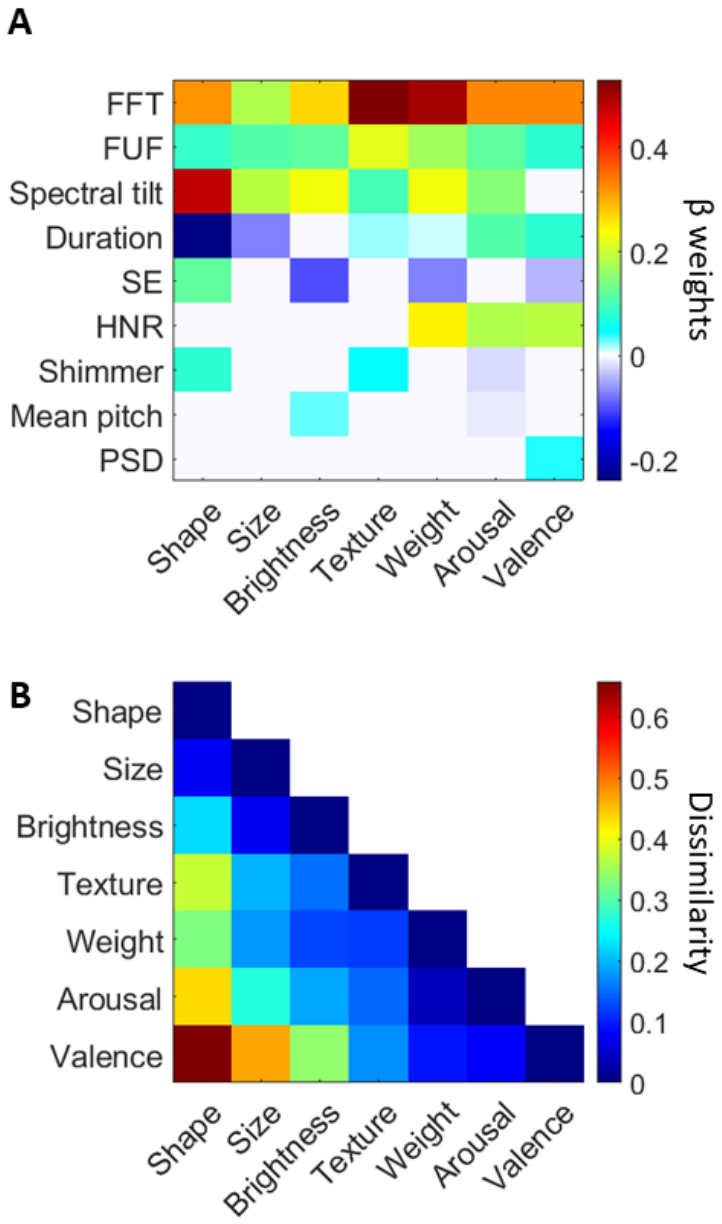

1
3
5
6
7
7
6
3
2
Shape
6
Size
5
Brightness
5
Texture
5
Weight
6
Arousal
7
Valence
6
PSD
HNR
SE
Duration
FFT
FUF
Spectral tilt
Shimmer
Mean pitch
Meaning domains
Acoustic parameters

Figure 3:


A
FFT
FUF
Spectral tilt
Duration
SE
HNR
Shimmer
Mean pitch
PSD
Shape
Size
Brightness
Texture
Weight
Arousal
Valence
0.4
0.2
0
-0.2
β weights
B
Shape
Size
Brightness
Texture
Weight
Arousal
Valence
0.6
0.5
0.4
0.3
0.2
0.1
0
Dissimilarity