Acoustic parameter combinations underlying mapping of pseudoword sounds to multiple domains of meaning: representational similarity analyses and machine-learning models

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], Neuroscience & Experimental Therapeutics[2], and Psychology[3],
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
Penn State Health Milton S. Hershey Medical Center
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, the resemblance between the sound structure of words and their meaning, is often studied using pseudowords. Previously, we showed that representational dissimilarity matrices (RDMs) of shape ratings of pseudowords correlated significantly with RDMs of acoustic parameters reflecting spectro-temporal variations; the ratings also correlated significantly with voice quality features. Here, we examined how perceptual ratings relate to spectro-temporal and vocal parameters of pseudowords across eight meaning domains. We largely replicated our previous findings for shape, while observing different patterns for other domains. Using a k-nearest-neighbor (KNN) algorithm, we compared 4095 combinations of twelve acoustic parameters (3 spectro-temporal and 9 characterizing vocal quality) to find the optimal combination associated with sound-symbolic ratings in each domain. We found sound-symbolic mappings linked to domain-specific combinations of acoustic parameters. One spectro-temporal parameter, the fast Fourier transform, contributed to all domains, indicating the significance of time-varying spectral properties for sound-symbolic judgments. We applied the KNN approach to generate shape ratings for 160 real words. These generated ratings strongly correlated with perceptual ratings of real words, indicating the value of the KNN approach to assess sound-symbolic mapping in natural languages, independent of semantic bias. Our findings support the relevance of sound symbolism to language.

I. 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[1,2]. 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[3,4], size[5], weight[6], and brightness[7]; as well as to non-sensory, abstract, domains, such as valence[8], arousal[9], and social dominance[10]. Moreover, non-arbitrary correspondences between sound and meaning have been found at different developmental stages[11,12], and in various linguistic, geographical and cultural contexts[13]. While converging lines of evidence establish the prevalence and robustness of sound symbolism in natural languages, it remains unclear how various acoustic features combine and contribute to sound-meaning associations in different domains of meaning.

A major focus of previous research has been on the contributions of phonetic and articulatory features to sound symbolism[5,14–17]. For instance, pseudowords containing the phonemes /i/ and /e/ connote smallness[5], brightness[18], and sharpness[17], compared to pseudowords containing the vowels /o/ and /u/ that have been associated with roundedness[19], heaviness[20] and social dominance[10]. Similarly, the occurrences of certain consonants versus vowels [21], voiced versus unvoiced consonants[19,22], obstruents versus sonorants[19], or vowel formants[23] underlie inherent biases in assigning meaning to certain sounds (e.g., rounded versus pointed and dark versus bright). In a related study, we show that perceptual ratings exhibit domain-specific associations with specific phonetic features, across eight sound-symbolic meaning domains[20]. These associations are important since the sound structure of any language is defined by such features.

Whether and how acoustic properties, of either real words or pseudowords, contribute to sound-symbolic mappings has received less attention[23]. Acoustic properties may be important because they relate to the listener's end of the 'speech chain'[24], with the spectro-temporal characteristics of speech mapping to phonetic features. For example, intracranial recordings show that, during listening to natural speech, the superior temporal gyrus (STG) shows phonetic selectivity arising from neuronal populations tuned to the specific spectro-temporal profiles of each phonetic feature[25–27]. Accordingly, cortical responses to phonetic features of speech could be decoded from low-level acoustic parameters[28]. Additionally, 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[23,29]. 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[30]. Thus, defining sound symbolism in terms of acoustic parameters complements analyses of phonetic features alone.

Few studies have examined the role of different acoustic features in sound symbolism, either focusing on sound-to-meaning mapping in one or two domains, or employing only a small set of

pseudowords[23,29,31]. Consequently, the relative importance of different acoustic parameters in mediating sound-symbolic mapping across diverse meaning domains remains largely unknown. In a related study[9], phonetic properties across multiple meaning domains were analyzed for a large set of 537 pseudowords; however, the acoustic features underlying sound symbolism and the parameters most relevant to this mapping are still uncertain.

While identifying the key acoustic drivers of sound symbolism is important, an equally crucial open question is the extent to which these acoustic parameters function within the lexicons of natural languages[32,33]. Since most studies have employed pseudowords to investigate the importance of different phonetic, articulatory and acoustic 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[34], 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[34], while others did not[35]. Moreover, cross-linguistic and cultural differences, such as the reversal of vowel–meaning associations reported by Diffloth, further complicate the application of this approach to real words[36]. Another approach is the best-worst judgement format developed by Hollis[37], 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. Bridging this gap requires not only identifying the key acoustic parameters underlying sound-symbolic associations, but also determining whether such patterns hold in the more complex and variable context of natural language.

Therefore, the purpose of this study is threefold: (1) To provide evidence for the role of acoustic properties in sound-symbolic mapping by replicating our previous work in the shape (rounded/pointed) domain and more critically, extending it to seven additional domains: size (small/big), brightness (bright/dark), hardness (hard/soft), roughness (rough/smooth), weight (light/heavy), arousal (calming/exciting), and valence (good/bad); (2) Of primary theoretical interest, to identify the combination of acoustic parameters that best supports sound-symbolic associations across the aforementioned eight meaning domains; and (3) As an application and proof of concept, to use these insights to generate sound-symbolic ratings for a set of real words based solely on their acoustic properties, independent of semantic bias. To these ends, we compared perceptual ratings for 537 pseudowords to their corresponding spectro-temporal and vocal parameters, using representational similarity analyses (RSA) for the spectro-temporal parameters and conventional correlation analysis for the vocal parameters, as in a previous study

by Lacey et al.[29]. Next, 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 eight meaning domains. 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. RSA were conducted independently of the combinations identified through our machine-learning approach. Taken together, this combined approach allowed us to examine the relationship of each parameter to sound-symbolic ratings in each meaning domain, and also identify which parameter combinations best account for sound-symbolic associations across different meaning domains. Finally, we extended this approach to generate predicted ratings of real words, assessing sound symbolism without the confounds of semantic processing.

II. METHODS

A. Participants

As detailed in the report of a related study[9], participants were recruited and compensated through the Prolific online participant pool[38] (https://prolific.com). Eligibility criteria included being between 18 and 35 years old and a native speaker of American English, without a history of language or hearing disorders. The final sample consisted of 389 participants (171 male, 208 female, 6 non-binary, 2 agender, 1 gender-fluid, and 1 who chose not to disclose their gender), with a mean age of 27 years and 1 month (SD = 8 months). The rating tasks were administered via the Gorilla platform[39] (https://gorilla.sc), where participants also provided informed consent. Participant numbers are reported below for each meaning domain. All procedures received approval from the Emory University Institutional Review Board. Additionally, data from 24 monolingual speakers of American English (20 females, mean age 30 years, SD 7 years) recruited for a different experiment[40] were used in this study. These participants were emailed a link to the experiment on Pavlovia.org after they gave consent over the telephone to participate; this study was approved by the Penn State University Institutional Review Board.

B. Stimuli

We used two sets of audio recordings comprising (1) pseudowords, and (2) real words. The pseudowords were a set of 537 two-syllable CVCV (i.e., consonant/vowel/consonant/vowel) pseudowords created by McCormick et al.[19], comprising only phonemes and combinations of phonemes that occur in American English[19]. The pseudowords were recorded by a female native speaker of American English, in random order and with neutral intonation, and digitized at a 44.1 kHz sampling rate. Each pseudoword was then down-sampled at 22.05 kHz, which is standard for speech, and amplitude-normalized using PRAAT speech analysis software[41]. The mean duration

of the pseudowords was 457 ms (SD ± 62 ms). We also chose 160 real words from Sidhu et al. (2021)[33], 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)[40].

C. Perceptual rating tasks

Participants recruited via Prolific were randomly assigned to rate 537 pseudowords on one of two 7-point Likert type scales representing categorical opposites across eight meaning domains: shape (rounded/pointed, N = 30/30 respectively), size (small/big, N = 32/31), brightness (bright/dark, N = 24/26), hardness (hard/soft, N = 27/27), roughness (rough/smooth, N = 29/29), weight (light/heavy, N = 29/26), arousal (calming/exciting, N = 30/28), and valence (good/bad, N = 26/23)[9].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, roughness, 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 yielded a unified scale ranging from 1 to 7, consistently anchored from one categorical extreme to the other, e.g., for shape, the rescaled rating 1 denoted 'most rounded' and 7 denoted 'most pointed'. These unified scales were used to rank the pseudowords along a continuum from one categorical extreme to the other within each meaning domain (see Section II.E.1) and then used for identifying the optimal acoustic parameter combinations in our machine-learning algorithm (see Section II.F).

For rating the 160 real words, 24 participants (on Pavlovia) 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. These two sets of ratings were then converted into a unified scale, with 1 representing the most rounded and 7 representing the most pointed, following the procedure described above.

In both studies, each participant rated all stimuli once, with the order of presentation randomized. Participants were compensated for their time, with the pseudoword ratings study lasting approximately 2 hours and the real word rating study about 1 hour. To ensure data quality, both tasks incorporated attention checks consisting of responding to tasks unrelated to the primary rating task, e.g. what is the color of carrot?

D. Acoustic parameters

Following Lacey et al.[29], we selected three spectro-temporal acoustic parameters: the fast Fourier transform (FFT), spectral tilt, and speech envelope, measuring the distributions of frequency and amplitude over time[42,43]; and seven parameters capturing the degree and regularity of voicing. These included 3 parameters indexing the proportion and characteristics of voiced segments[41]: the

fraction of unvoiced frames (FUF), mean autocorrelation (MAC), and pulse number; and 3 parameters capturing the regularity of voicing or voice quality during the production of each pseudoword: jitter, shimmer, and harmonics-to-noise ratio (HNR)[44]; as well as the standard deviation of the fundamental frequency of the speech sound ($F0_{SD}$)[45]. We also included 2 additional parameters: the mean F0, and item duration. Each parameter is described in detail below:

1. **Spectro-temporal parameters**
   a. *Fast Fourier transform (FFT)*
   The FFT extracts the frequency components of the speech signal and captures how their energy changes over time, thereby representing the power spectrum throughout the duration of the spoken pseudowords. The FFT is visualized through the spectrogram, which displays distribution of power across frequencies over time. For instance, obstruents are typically reflected in sudden changes in power across different frequencies, whereas sonorants show more gradual variations in power with across frequencies.

   b. *Spectral tilt*
   Spectral tilt captures relative differences in power across frequencies, providing an estimate of the overall slope of the power spectrum based on measurements taken across the entire utterance. When higher frequencies contain less power than lower ones, the power spectrum shows a steep downward slope from low to high frequencies. This slope becomes flatter when more power is concentrated in the higher frequencies.

   c. *Speech envelope*
   The speech envelope captures fluctuations in amplitude over time, closely aligning with phonemic features and syllabic transitions[43]. When these transitions are abrupt, such as with stops, affricates, or fricatives, the envelope becomes irregular and discontinuous. In contrast, smoother and more gradual transitions, as seen with sonorants, result in a more continuous and uniform envelope.

2. **Vocal parameters**
   a. *Mean autocorrelation (MAC)*
   This is a measure of the periodicity of a signal[29,41]. Periodicity is expected to be high for a long vowel such as "ooo" or a consonant like "mmm" where each successive period should be very similar to the previous one, that is, they should be highly correlated. Higher autocorrelation values reflect smoother, more regular voicing and a greater presence of voiced segments, while lower values indicate a more irregular pattern and fewer voiced or periodic segments.

   b. *Pulse number*
   This refers to the number of glottal pulses, meaning the opening and closing of the vocal folds, that occur during the production of vowels or voiced consonants, measured across the entire utterance[29,41]. A lower pulse number suggests a more uneven voice pattern or

fewer voiced segments, whereas a higher pulse number suggests a smoother voice pattern or more voiced segments.

c. *Fraction of unvoiced frames (FUF)*

The fraction of unvoiced frames (FUF) indicates the proportion of time windows in which the vocal folds are not active, expressed as a percentage[29,41] . FUF is influenced by the phonemic composition of a pseudoword — higher values are observed in pseudowords containing unvoiced features, such as obstruents, while lower values are typical for those with voiced, periodic elements like vowels.

d. *Shimmer*

Shimmer measures the variability in peak-to-peak amplitude of the glottal waveform[46] . It serves as an indicator of vocal stability, with low shimmer corresponding to smooth, stable speech and high shimmer indicating irregular amplitude fluctuations, often perceived as hoarseness in the voice.

e. *Jitter*

Jitter refers to the variation in frequency between successive vocal cycles and serves as an indicator of voice quality by capturing irregularities in vocal fold vibration[47]. High jitter values are associated with a rough or "breaking" voice quality. It is typically assessed using sustained vowel sounds, where minimal frequency variation is expected. In pseudoword production, elevated jitter indicates greater vocal instability and contributes to the perception of vocal roughness.

f. *F0 & $F0_{SD}$*

F0 is the average fundamental frequency of the speech signal, while $F0_{SD}$ captures the variability in F0 across the utterance[29,41]. $F0_{SD}$ reflects degree of change in intonation: low $F0_{SD}$ corresponds to a flat, monotone voice, whereas high $F0_{SD}$ indicates a more dynamic or expressive vocal tone[45].

g. *Duration*

The duration of the speech signal, measured in milliseconds (ms), is influenced by several factors, including the distinction between long and short vowels[48,49], the surrounding phonetic context[48], and the speaker's rate of speech[50].

E. Analysis of acoustic parameters

1. *Spectro-temporal parameter analysis*

Representational similarity analysis (RSA), originally developed for analyzing functional magnetic resonance imaging (fMRI) data[51], was applied in a previous study[29] to compare perceptual ratings of pseudowords on the rounded-pointed continuum to their spectro-temporal acoustic features. In the present study, we employed a similar RSA approach across all eight domains.

For each domain, we constructed a 537 × 537 representational dissimilarity matrix (RDM) based on participants' perceptual ratings of the pseudowords. Using the unified scale (see section II.C), pseudowords were ranked left to right by their mean rating along the domain-specific dimension **(Figure 1**, Step 1): from most rounded to most pointed (shape), smallest to biggest (size), hardest to softest (hardness), smoothest to roughest, lightest to heaviest (weight), brightest to darkest (brightness), most calming to most exciting (arousal), and best to worst (valence). Based on these ordered values, we then created reference RDMs for each domain (**Figure 1**, Step 2). This involved calculating the first-order Pearson correlation between the vectorized perceptual ratings of each pair of pseudowords across participants using the original (unrecoded) data, since RDMs capture dissimilarity irrespective of the scale used by any individual. The dissimilarity between each pair was defined as 1 minus the correlation coefficient (1 – r), and this value was entered into each cell of the RDM. Note that, while correlations range from +1 to -1, the corresponding dissimilarity values range from 0 to 2.

We also created corresponding acoustic RDMs (**Figure 1**, Step 3) by vectorizing the spectro-temporal measurements for each pseudoword, computing pairwise correlations between items, and converting these to dissimilarity values using the 1 – r metric. These acoustic RDMs were computed separately for the speech envelope, spectral tilt, and FFT. Spectro-temporal parameters and their associated RDMs were computed in MATLAB (version 2021a; The MathWorks, Natick, MA), following the procedures outlined in Lacey et al. (2020). To ensure consistency across items, we normalized the duration of all pseudowords to the mean duration of 457 ms using MATLAB's resampling function. At a down-sampled rate of 22.05 kHz (see section II.B), each pseudoword was represented by a common vector length of 10,077 data points (22,050 × 0.457). This normalization introduced only minimal noise, proportional to the relative variability in duration (13.5%, calculated as SD/mean = 62/457 ms). Standardizing the length allowed for equal dimensionality across pseudowords in the RDM computations. It is worth noting that the spectro-temporal parameters are not all independent, e.g., both spectral tilt and FFT capture aspects of the distribution of frequency content, and that each parameter comprises multiple values per pseudoword. The exact measurements extracted from these vectors varied across parameters, for example differing window lengths and overlaps (details provided in the Supplementary Material). As a result, the number of data points underlying the dissimilarity calculations differed by parameter. For instance, the speech envelope and FFT had 10077 and 5489 data points respectively, and pairwise dissimilarities were computed using Pearson correlations. In contrast, spectral tilt was represented by only 8 data points, so dissimilarities were computed using Spearman (non-parametric) correlations.

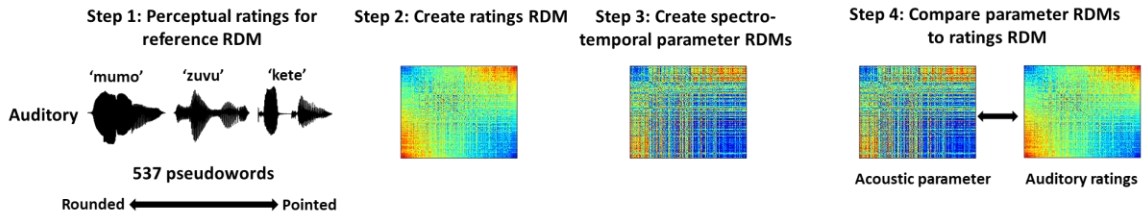

Figure 1: Schematic illustration of the analysis pipeline. Step 1: auditory pseudowords were rated along different dimensions, e.g., shape (rounded/pointed), brightness (bright/dark). Step 2: these perceptual ratings were used to create reference representational dissimilarity matrices (RDMs). Step 3: RDMs for each of the spectro-temporal parameters were created for each domain. Step 4: in each domain, the spectro-temporal parameter RDMs were compared to the ratings RDM by way of a second-order correlation.

Finally, each acoustic parameter RDM was compared to the corresponding perceptual ratings RDM using a second order Spearman correlation (**Figure 1**, Step 4). Significant correlations at this level indicate which spectro-temporal features are associated with perceptual judgments of the pseudowords.

2. *Voice parameter analysis*

As reported previously[29], the voice parameters were measured in PRAAT[41], using the original pseudoword sound files, without applying the resampling procedure described earlier. Similar to the spectro-temporal parameters, the voice parameters are not necessarily independent of one another. For example, FUF and pulse number both measure the proportion of an utterance that is voiced. However, unlike the spectro-temporal parameters, the voice parameters are represented by a single value for each pseudoword. For calculating jitter and shimmer, we used the local measures, which represent the mean absolute difference in frequency and amplitude, respectively, between consecutive periods of the speech waveform. These differences were then divided by the overall mean difference across all periods and expressed as a percentage[41]. Each voice parameter, represented by a single value per pseudoword, was directly compared to the recoded perceptual ratings using standard correlational analyses.

F. *k*-nearest neighbor models

A *k*-nearest neighbor (KNN) machine-learning algorithm was employed to find the set of acoustic parameters that correlated best with the unified ratings (see section II.C) for the 537 pseudowords in each of the seven meaning domains studied. The algorithm (**Figure 2**) comprised five steps as described below.

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. Pairwise distances between pseudowords were computed for these parameters using the methods described in section II.E.1. The dissimilarity values derived from the spectro-temporal parameters were then linearly rescaled

to range between 0 and 1.. For the nine voice parameters, each pseudoword was represented by a single value. Pairwise distances between pseudowords for these parameters 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. The rescaling was performed to ensure that the pairwise distances between pseudowords based on the spectro-temporal and voice parameters were on a uniform scale.

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 correlated with the ratings of the pseudowords, we varied the number of parameters from 1-12 and iteratively performed the subsequent steps with all 4095 possible parameter combinations.

3. Step 3: k-nearest neighbors

The KNN model rating for every pseudoword was iteratively generated in each of the multidimensional parameter spaces defined by a particular combination of acoustic parameters. In each of these spaces, the computed 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: $k = \sqrt{537} \approx 23$ [52]. In each parameter space, the rating for a given pseudoword was computed 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 computed on a participant-wise basis in all 4095 parameter spaces (thus, for shape, N = 60; for size, 63; brightness, 50; hardness, 54; roughness, 58; weight, 55; arousal, 58; valence, 49).

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.

4. Steps 4 and 5: Model evaluation and selection

In each meaning domain, participant-wise computed 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 computing model-based pseudoword ratings in that domain.

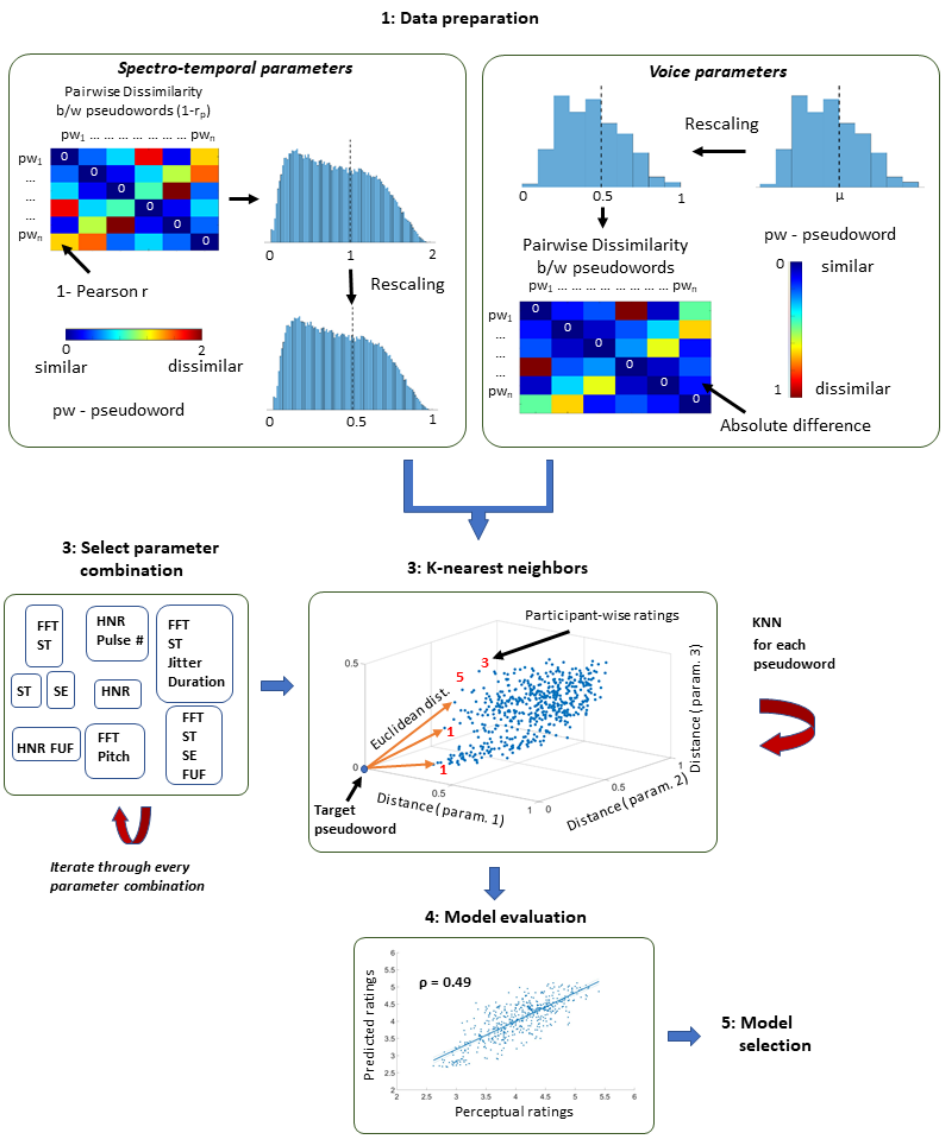

Figure 2: Analysis flowchart for generating pseudoword ratings based on acoustic parameters. Step 1: The pairwise distance between pseudowords was computed based on their differences in spectro-temporal and voice parameters, as detailed in the text. **Step 2.** The number of parameters entering the algorithm was varied from 1-12 and all 4095 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 model-based ratings on a participant-by-participant basis were used to quantify

G. Shape ratings of real words

As a proof of concept, we sought to apply our approach to predict the shape ratings of real words independent of semantic knowledge. Shape ratings for real words were based on the perceptual ratings of pseudowords in the acoustic neighborhood of each real word. Using the KNN procedure outlined in Section II.F, we determined the most relevant acoustic parameters for classifying

pseudowords' shape ratings. These parameters defined the multidimensional acoustic space within which the ratings for real words were estimated. Each real word was iteratively projected into this space, and its 23 nearest pseudoword neighbors ($k = \sqrt{538} \approx 23$) were determined using Euclidean distances from all 537 pseudowords. The modal shape rating of these neighbors was then assigned to the real word, as it reflects the sound-symbolic judgment derived from the perceptual ratings of pseudowords occupying the same acoustic neighborhood. As outlined in Section II.F.3, the set of 23 nearest pseudoword neighbors for each real word remains invariant because the Euclidean distances to the pseudowords are determined solely by their acoustic features (i.e., the optimal parameter combination for shape). However, the ratings of these neighbors vary across participants. Consequently, each real word's rating was predicted separately for each participant. These participant-specific predicted ratings were then averaged across all participants, and the mean predicted ratings were correlated with the mean perceptual ratings for each item using Pearson's correlation. We acknowledge that participants' perceptual ratings for real words may not be entirely free from semantic influences, even though the participants were instructed to respond based solely on the sound of the real words. However, the perceptual ratings of the real words are used only as a means to validate our findings.

In addition to the aforementioned analyses, we performed three supplementary procedures to examine perceptual variability across participants and assess the robustness of our findings. We first quantified participant-level variability in perceptual ratings by calculating confidence intervals for each pseudoword. Next, we explored the effect of varying *k* in the KNN algorithm across semantic domains to assess the sensitivity of model performance to parameter choice. Finally, we conducted subsampling analyses to evaluate the stability of feature selection across participant subsets. Detailed descriptions of the methods and results for these analyses are provided in the Supplementary Materials.

III. RESULTS

A. Acoustic parameters mediating sound-symbolic associations for each meaning domain

For each domain, the following sections present: RSA of individual spectro-temporal parameters, conventional correlations between each voice parameter and the perceptual ratings, and the optimal parameter combination identified by the KNN model.

1. Shape

a. RSA for spectro-temporal parameters

Consistent with our previous study[29], RSA showed that the RDMs for all three spectro-temporal parameters were significantly positively correlated with the RDM for ratings of the pseudowords from rounded to pointed (**Figure 3 B-D**). The strongest correlation ($r = .41$, $p < .001$) was for spectral tilt, which was steeper for pseudowords rated as more rounded where power was concentrated in low-frequency bands, but flatter for pseudowords rated as pointed in which power migrated to the higher-frequency bands (see representative highly rated rounded and pointed pseudowords in **Supplementary Figure 1a**). The next strongest relationship ($r = .28$, $p < .001$)

was for the FFT: the spectrograms showed power variations across frequencies and time that were gradual for a representative rounded pseudoword but abrupt for a representative pointed pseudoword in **Supplementary Figure 1b**. The weakest correlation was for the speech envelope ($r = .18$, $p < .001$), which was smoother and more continuous for pseudowords rated as strongly rounded, compared to strongly pointed, where the speech envelope was discontinuous and uneven (illustrated for representative pseudowords in **Supplementary Figure 1c**).

b. Conventional correlations for voice parameters

For the voice parameters (**Figure 3 E-M**), the mean HNR ($r = -.58$, $p<0.001$), pulse number ($r = -.48$, $p<0.001$), and MAC ($r = -.42$, $p<0.001$) were significantly negatively correlated with ratings of the pseudowords, reflecting the presence of more sonorants in items rated as rounded compared to those rated as pointed. The FUF ($r = .50$, $p<0.001$) was significantly positively correlated with ratings, reflecting the greater presence of unvoiced segments in items rated as more pointed. Shimmer ($r = .39$, $p<0.001$) and jitter ($r = .30$, $p<0.001$) were significantly positively correlated with ratings, reflecting increasing variability in voice quality, as ratings of the pseudowords transitioned from rounded to pointed. Thus far, the voice parameter results replicated our earlier study[29] but, in the present study, and in contrast to our earlier study, $F0_{SD}$ ($r = .17$, $p<0.001$) was correlated significantly and positively with ratings, increasing as ratings moved from rounded to pointed. This indicates greater variability in the fundamental frequency for pointed, compared to rounded, pseudowords. Thus, overall, ratings moved from rounded to pointed as noisiness and variability in the speech pattern increased. For the two new voice parameters that were not included in our earlier study, there was a significant tendency for pseudowords to be rated as more pointed as pseudoword duration ($r = -.13$, $p=0.002$) shortened. F0 ($r = .03$, $p = 0.5$) did not show a significant correlation with the ratings.

c. Optimal combination

The KNN algorithm identified an optimal model comprising five parameters: two spectro-temporal parameters including the FFT and spectral tilt, and three vocal parameters including the FUF, shimmer, and duration. This combination yielded ratings that showed the highest correlation (r_s = 0.59), among all the combinations tested, with participants' (N = 60) perceptual ratings of the 537 pseudowords in the shape domain (**Table 1**).

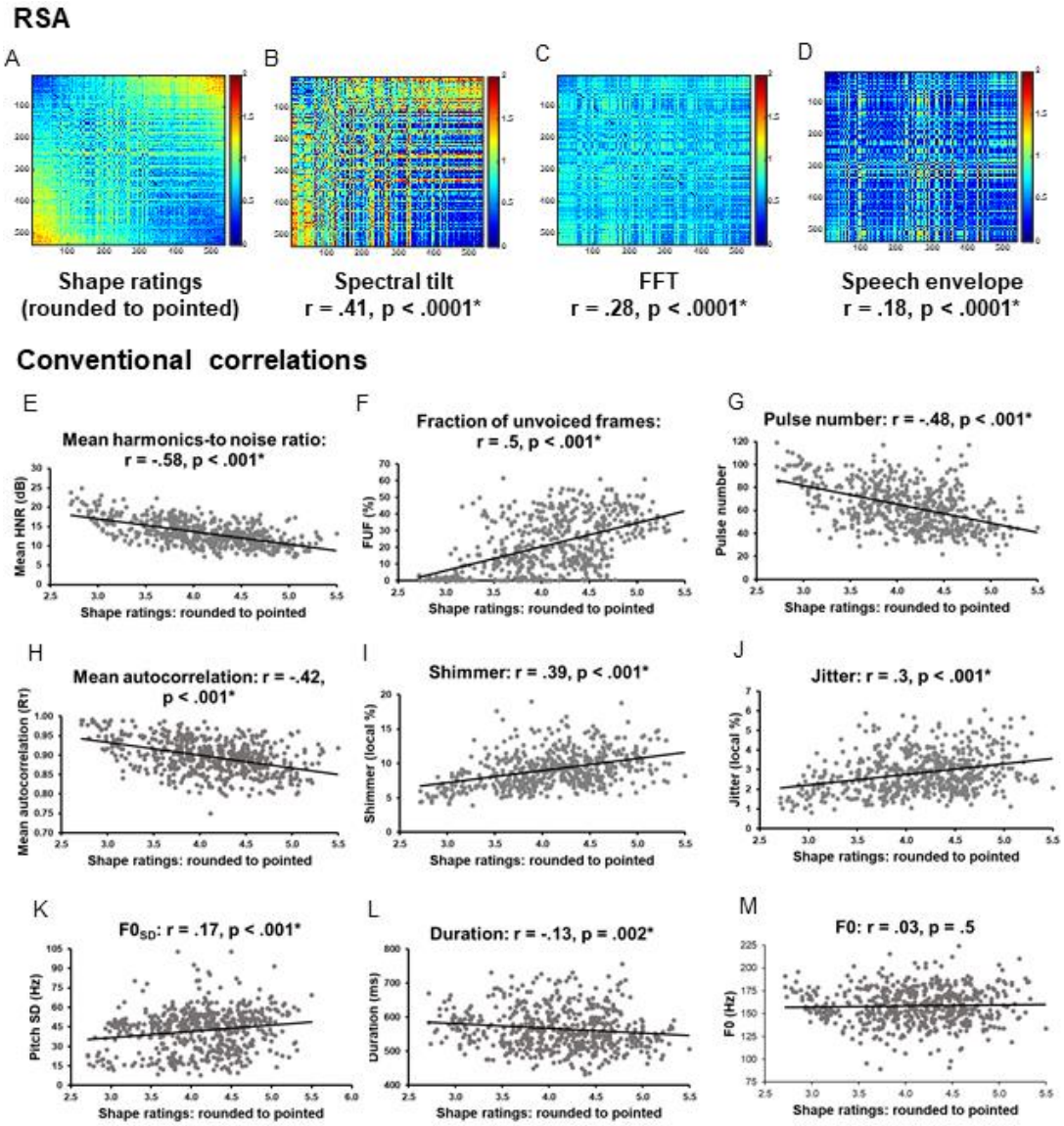

Figure 3: Representational similarity analyses and conventional correlations relating auditory shape ratings and acoustic parameters. **A**: Ratings RDM. **B–D**: RDMs for the three spectro-temporal parameters. Pseudowords are ordered, left to right, from most rounded to most pointed. Spectro-temporal parameter correlations are ordered, left to right, from the strongest to the weakest relationship. Color bar shows pairwise dissimilarity, where 0 = zero dissimilarity (items are identical) and 2 = maximum dissimilarity (items are completely different). r = Spearman correlation coefficient for second-order correlation between the parameter RDM and ratings RDM; *correlation passes the Bonferroni-corrected $\alpha = .0167$ for three tests; df = 143,914 for all three spectro-temporal parameters. **E-M**: Conventional correlations between auditory perceptual ratings for shape and nine voice-quality parameters, ordered, top to bottom, left to right, from the strongest to the weakest relationship, whether positive or negative. r = Pearson correlation coefficient; *correlation passes the Bonferroni-corrected α of 0.0056 for nine tests; df = 535 for all nine voice parameters.

2. Size

a. RSA for spectro-temporal parameters

RSA revealed a weak correlation between the RDMs for spectral tilt and size ratings ($r = .13$, $p < .001$), with even weaker correlations observed between the size ratings RDM and the RDMs for the speech envelope ($r = .11$, $p < .001$) and FFT ($r = .08$, $p < .001$) (**Figure 4 B-D**), although these correlations were all significant. We observed that spectral tilt was broadly similar for both pseudowords rated as small as well as big, the slope of the spectral density being initially steep before flattening out (**Supplementary Figure 2a**), although for pseudowords rated as big, broadband energy remained relatively higher across the frequency range. Both small- and big-rated pseudowords showed discontinuous speech envelopes, the only difference being that big-rated pseudowords showed somewhat greater amplitude for the second syllable, compared to the small-rated pseudowords (see representative pseudowords in **Supplementary Figure 2b**). In the spectrograms, there were relatively distinct changes in energy for pseudowords regardless of their rating, although there was somewhat more energy in the higher frequencies for the small-rated, compared to the big-rated, pseudowords (**Supplementary Figure 2c**).

b. Conventional correlations for voice parameters

There were no significant correlations between voice parameters and size ratings that survived correction (**Figure 4 E-M**); in particular, we did not replicate the finding that short and long pseudoword duration reflect small and large size[23], respectively.

c. Optimal combination

The KNN algorithm showed that perceptual ratings of size of the 537 pseudowords (N = 63) were best correlated ($r_s = 0.45$) with the ratings predicted by a combination of five parameters. This combination included all three spectro-temporal parameters: the FFT, spectral tilt, and speech envelope, and two voice parameters: the FUF and duration (**Table 1**).

3. Brightness

a. RSA for spectro-temporal parameters

RSA revealed weak but statistically significant correlations between the ratings RDM and all three spectro-temporal parameter RDMs (**Figure 5 B-D**). The strongest among them was for spectral tilt ($r = .14$, $p < .001$), followed by weaker correlations for the FFT ($r = .08$, $p < .001$) and the speech envelope ($r = .04$, $p < .001$).

Spectral tilt was flatter for brighter-rated pseudowords, reflecting more high-frequency energy, whereas for darker-rated pseudowords, power was concentrated in the lower frequency bands (**Supplementary Figure 3a**). The spectrogram showed greater power in the higher frequencies for pseudowords rated as bright, compared to those rated as dark (**Supplementary Figure 3b**). However, there were relatively small differences in the abruptness of power variations across frequencies. The speech envelope was discontinuous for both bright- and dark-rated pseudowords but amplitude was perhaps greater for the latter (**Supplementary Figure 3c**).

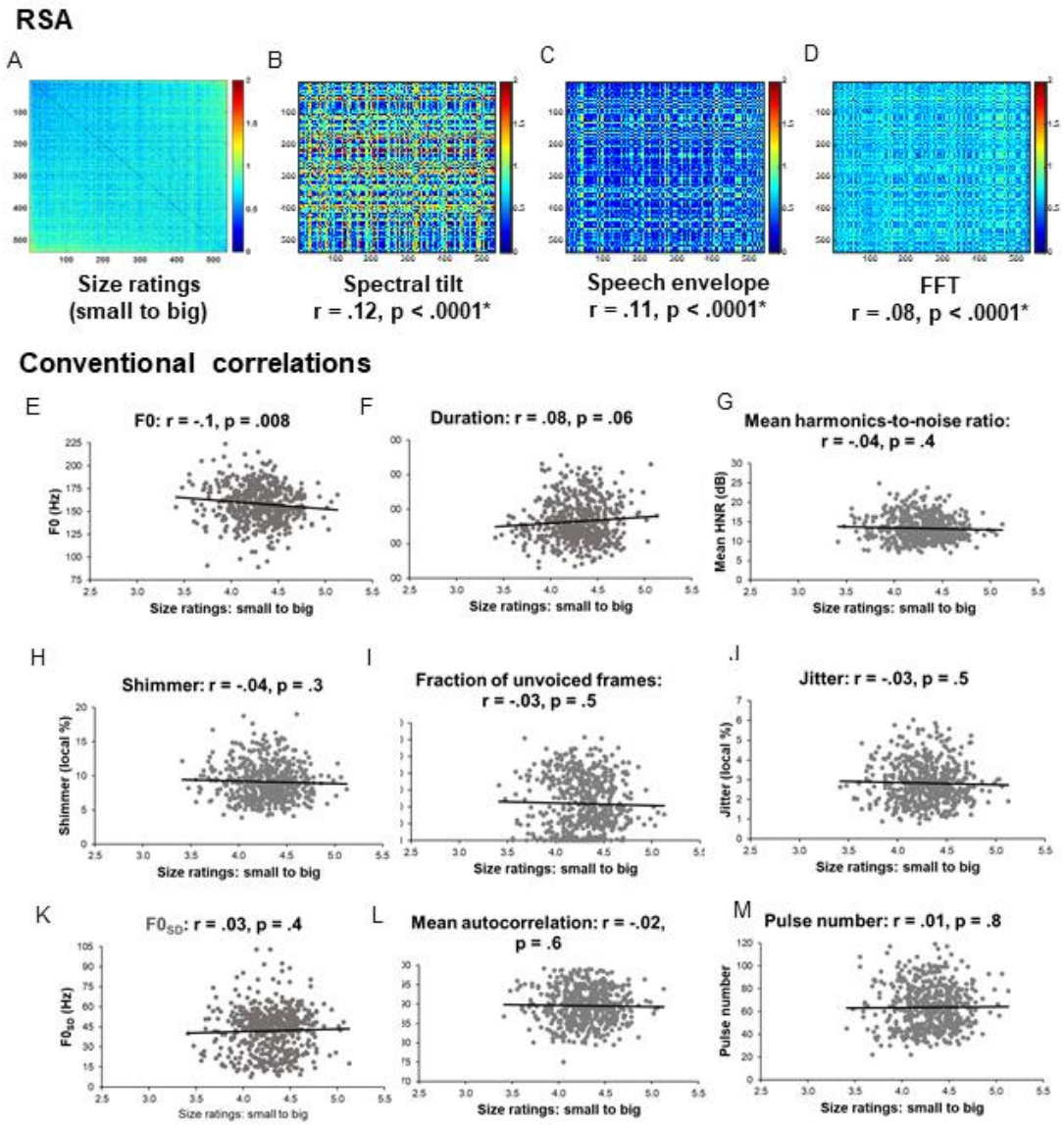

Figure 4: Representational similarity analyses and conventional correlations relating auditory size ratings and acoustic parameters. **A**: Ratings RDM. **B–D**: RDMs for the three spectro-temporal parameters. Pseudowords are ordered, left to right, from smallest to largest. Spectro-temporal parameter correlations are ordered, left to right, from the strongest to the weakest relationship. *correlation passes the Bonferroni-corrected α = .0167 for three tests; df = 143,914 for all three spectro-temporal parameters. **E-M**: Conventional correlations between auditory perceptual ratings for size and nine voice-quality parameters, ordered, top to bottom, left to right, from the strongest to the weakest relationship, whether positive or negative. *correlation passes the Bonferroni-corrected α of 0.0056 for nine tests; df = 535 for all nine voice parameters.

b. Conventional correlations for voice parameters

We observed that the FUF (r = -.20, p < .001) decreased significantly as ratings increased towards the dark end of the scale, while mean HNR (r = .20, p < .001) and pulse number (r = .13, p < .001) were significantly positively correlated with ratings such that voice quality became more periodic

and smoother for pseudowords that were rated as darker. There was a small effect of duration (r = .1, p = 02) in which longer pseudowords were rated as darker, but this did not survive correction. There were also weak relationships for shimmer (r = -.11, p = .01) and MAC (r = .1, p = .02) that did not survive correction, and non-significant relationships were found for $F0_{SD}$ (r = -.05, p = .2) and jitter (r = -.04, p = .4) (**Figure 5 E-M**).

c. Optimal combination

A combination of ten parameters including all three spectro-temporal parameters: FFT, spectral tilt, and speech envelope; and seven voice parameters: FUF, pulse number, jitter, MAC, F0, $F0_{SD}$ and duration, constituted the optimal model that showed the best correlation (r_s = 0.52) between perceptual brightness ratings (N = 50) and those generated by the KNN algorithm (**Table 1**).

4. Hardness

a. RSA for spectro-temporal parameters

The RDMs for all three spectro-temporal parameters were significantly positively correlated with the RDM for ratings of the pseudowords from hard to soft (**Figure 6 B-D**). The FFT RDM showed the strongest correlation (r = .36, p < .001). The RDM of the speech envelope was similarly strongly correlated (r = .33, p < .001) while the relationship between ratings and spectral tilt was weaker (r = .11, p < .001), with more high-frequency energy for the hard-, compared to the soft-rated, pseudowords. Hard-rated pseudowords exhibited more abrupt changes in power across frequency and time than pseudowords rated as soft, where such changes were more gradual. More energy was concentrated at the higher-frequency bands for hard-rated pseudowords compared to the soft-rated ones (see representative spectrograms in **Supplementary Figure 4a**). A higher-rated hard pseudoword exhibited a more discontinuous envelope than a higher-rated soft pseudoword, reflecting the sustained higher amplitude for the latter (**Supplementary Figure 4b**). More power was concentrated in higher frequencies for hard-rated pseudowords compared to soft-rated pseudowords (**Supplementary Figure 4c**).

b. Conventional correlations for voice parameters

For the voice parameters, soft-rated pseudowords were associated with greater periodicity and less vocal variability, while the reverse was true for the hard-rated pseudowords (**Figure 6 E-M**). The mean HNR (r = .42, p < .001), pulse number (r = .38, p < .001), and MAC (r = .30, p < .001) were significantly positively correlated with ratings, while the FUF (r = -.28, p < .001) was significantly negatively correlated. Shimmer (r = -.26, p < .001) and jitter (r = -.23, p < .001) were significantly negatively correlated with hardness ratings, meaning that as ratings moved from hard to soft, there was decreasing variability in amplitude and frequency, respectively. Hardness ratings were uncorrelated with F0 (r = -0.01, p = 0.7); there was a weak effect in which $F0_{SD}$ (r = -0.12, p =0.006) decreased for softer pseudowords but this did not survive correction.

c. Optimal combination

The optimal combination that yielded the highest correlation between the KNN-generated ratings and participants' perceptual ratings (N = 54) of hardness comprised five acoustic parameters (r_s =

0.54). This set included two spectro-temporal parameters: FFT and spectral tilt, and three voice parameters: FUF, shimmer, and duration (**Table 1**).

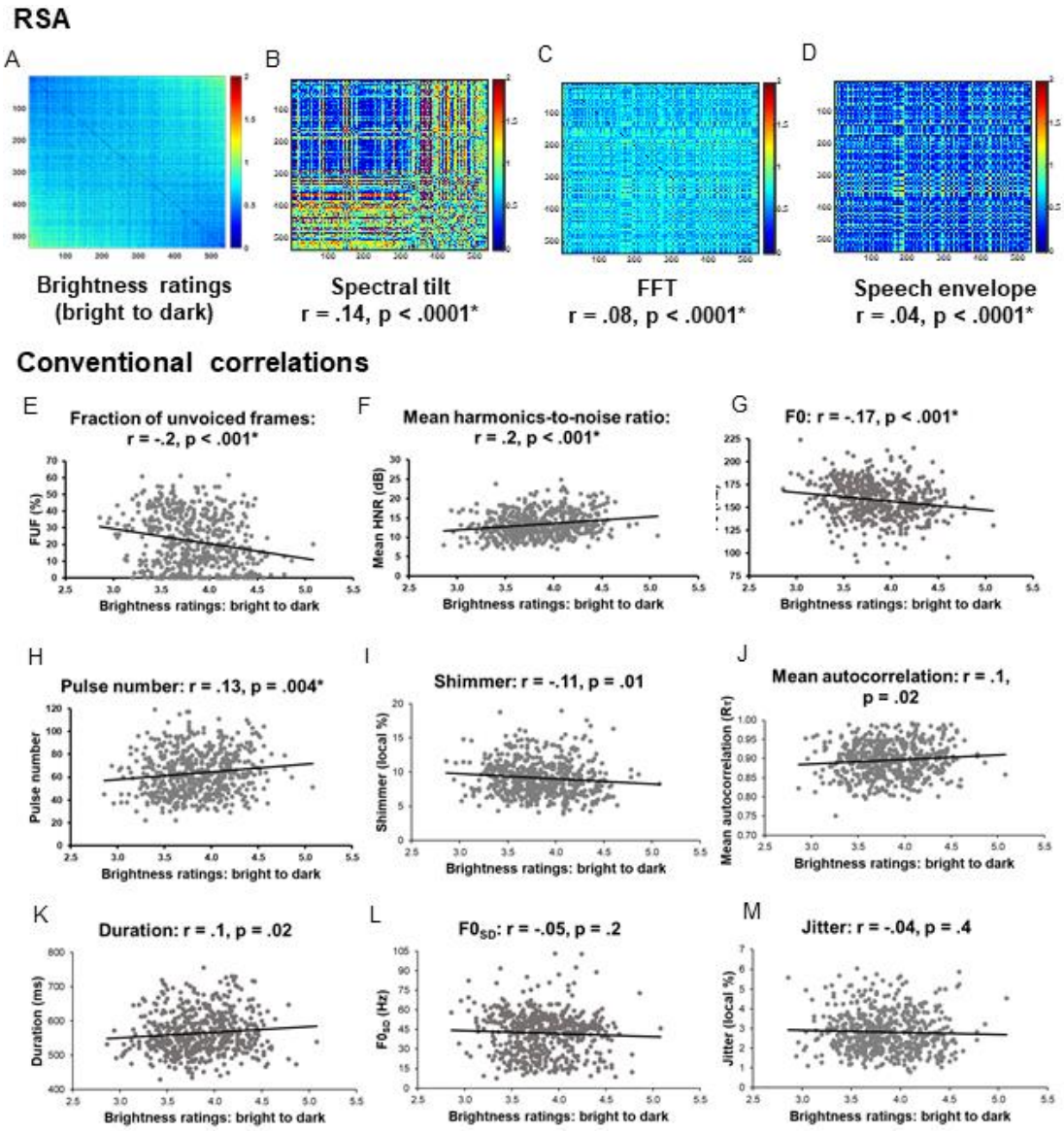

Figure 5: Representational similarity analyses and conventional correlations relating auditory brightness ratings and acoustic parameters. **A**: Ratings RDM. **B–D**: RDMs for the three spectro-temporal parameters. Pseudowords are ordered, left to right, from brightest to darkest. Spectro-temporal parameter correlations are ordered, left to right, from the strongest to the weakest relationship. *correlation passes the Bonferroni-corrected α = .0167 for three tests; df = 143,914 for all three spectro-temporal parameters. **E-M**: Conventional correlations between auditory perceptual ratings for brightness and nine voice-quality parameters, ordered, top to bottom, left to right, from the strongest to the weakest relationship, whether positive or negative. *correlation passes the Bonferroni-corrected α of 0.0056 for nine tests; df = 535 for all nine voice parameters.

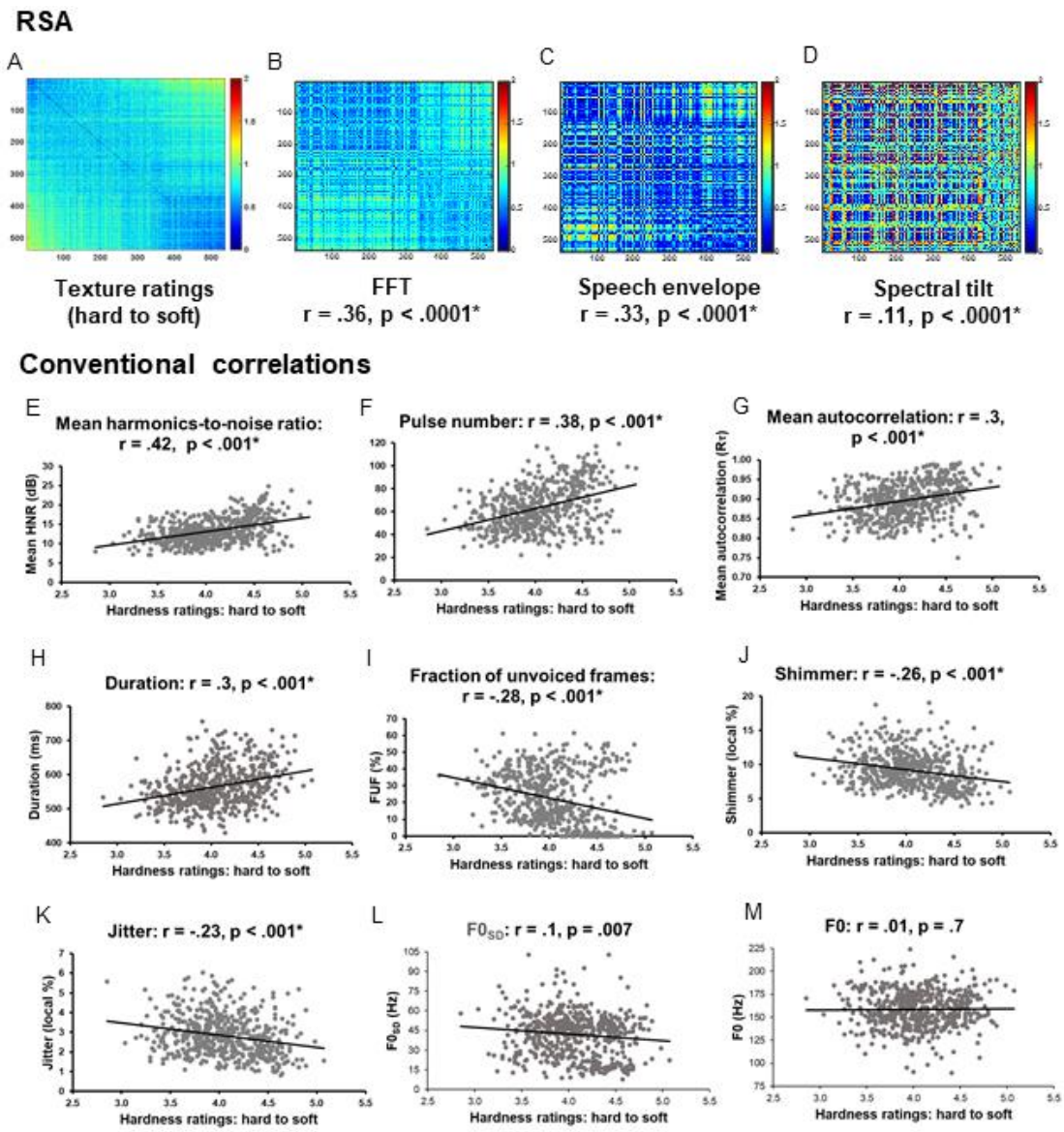

Figure 6. Representational similarity analyses and conventional correlations relating auditory hardness ratings and acoustic parameters. **A**: Ratings RDM. **B–D**: RDMs for the three spectro-temporal parameters. Pseudowords are ordered, left to right, from hardest to softest. Spectro-temporal parameter correlations are ordered, left to right, from the strongest to the weakest relationship. *correlation passes the Bonferroni-corrected α = .0167 for three tests; df = 143,914 for all three spectro-temporal parameters. **E-M**: Conventional correlations between auditory perceptual ratings for hardness and nine voice-quality parameters, ordered, top to bottom, left to right, from the strongest to the weakest relationship, whether positive or negative. *correlation passes the Bonferroni-corrected α of 0.0056 for nine tests; df = 535 for all nine voice parameters.

5. Roughness

a. RSA for spectro-temporal parameters

RSA showed that the RDMs for all three spectro-temporal parameters were significantly positively correlated with the RDM for ratings of the pseudowords from smooth to rough (**Figure 7 B-D**).

The strongest correlation was for the FFT RDM (r = .32, p < .001), followed by the speech envelope RDM (r = .28, p <.001). Spectral tilt had the weakest relationship to the ratings in terms of between-RDM correlations (r = .1, p < .001). Pseudowords rated as smooth were characterized by more smooth transitions in energy across frequencies and time (**Supplementary Figure 5a**) compared to those rated rough which had more abrupt transitions. The speech envelope was even and continuous for smooth-rated pseudowords, but uneven and discontinuous for the rough-rated pseudowords (**Supplementary Figure 5b**). More power was concentrated in higher frequencies for smooth-rated, compared to rough-rated pseudowords (**Supplementary Figure 5c**).

b. Conventional correlations for voice parameters

For voice parameters (**Figure 7 E-M**), the HNR (r = -.51, p < .001), pulse number (r = -.4, p < .001), MAC (r = -.36, p < .001) and duration (r = -., p < .001) were significantly negatively correlated with ratings of the pseudowords, indicating that periodicity and duration decreased as ratings transitioned from smooth to rough. The FUF (r = .37, p < .001) increased as ratings transitioned from smooth to rough, also indicating decreasing periodicity as the proportion of unvoiced, aperiodic, segments increased. Similarly, shimmer (r = .33, p < .001) and jitter (r = .23, p < .001) were significantly positively correlated to the size ratings. The $F0_{SD}$ relationship (r = .1, p = .01) did not survive multiple comparison correction while the mean F0 was not significantly correlated with roughness ratings (r = .02, p = .51).

c. Optimal combination

The KNN algorithm identified an optimal combination comprising the two spectro-temporal parameters: FFT and speech envelope, that produced ratings for the 537 pseudowords which correlated (r_s =.56) best with the perceptual ratings (N = 58) of pseudowords in the roughness domain (**Table 1**).

6. Weight

a. RSA for spectro-temporal parameters

RSA showed that the RDMs for all three spectro-temporal parameters were significantly positively correlated with the RDM for ratings of the pseudowords from light to heavy (**Figure 8 B-D**). The strongest correlation was for the FFT RDM (r = .38, p < .001), followed by the speech envelope RDM (r = .28, p <.001). Spectral tilt had the weakest relationship to the ratings in terms of between-RDM correlations (r = .19, p < .001). Pseudowords rated as heavy were characterized by more abrupt transitions in energy across frequencies and time, as shown in the spectrograms (**Supplementary Figure 6a**). The speech envelope was even and continuous for light-rated pseudowords, but uneven and discontinuous for the heavy-rated pseudowords (**Supplementary Figure 6b**). Each of the representative pseudowords in **Supplementary Figure 6c** shows energy at both low and high frequencies, though those rated as heavy display slightly steeper spectral slopes compared to those rated as light.

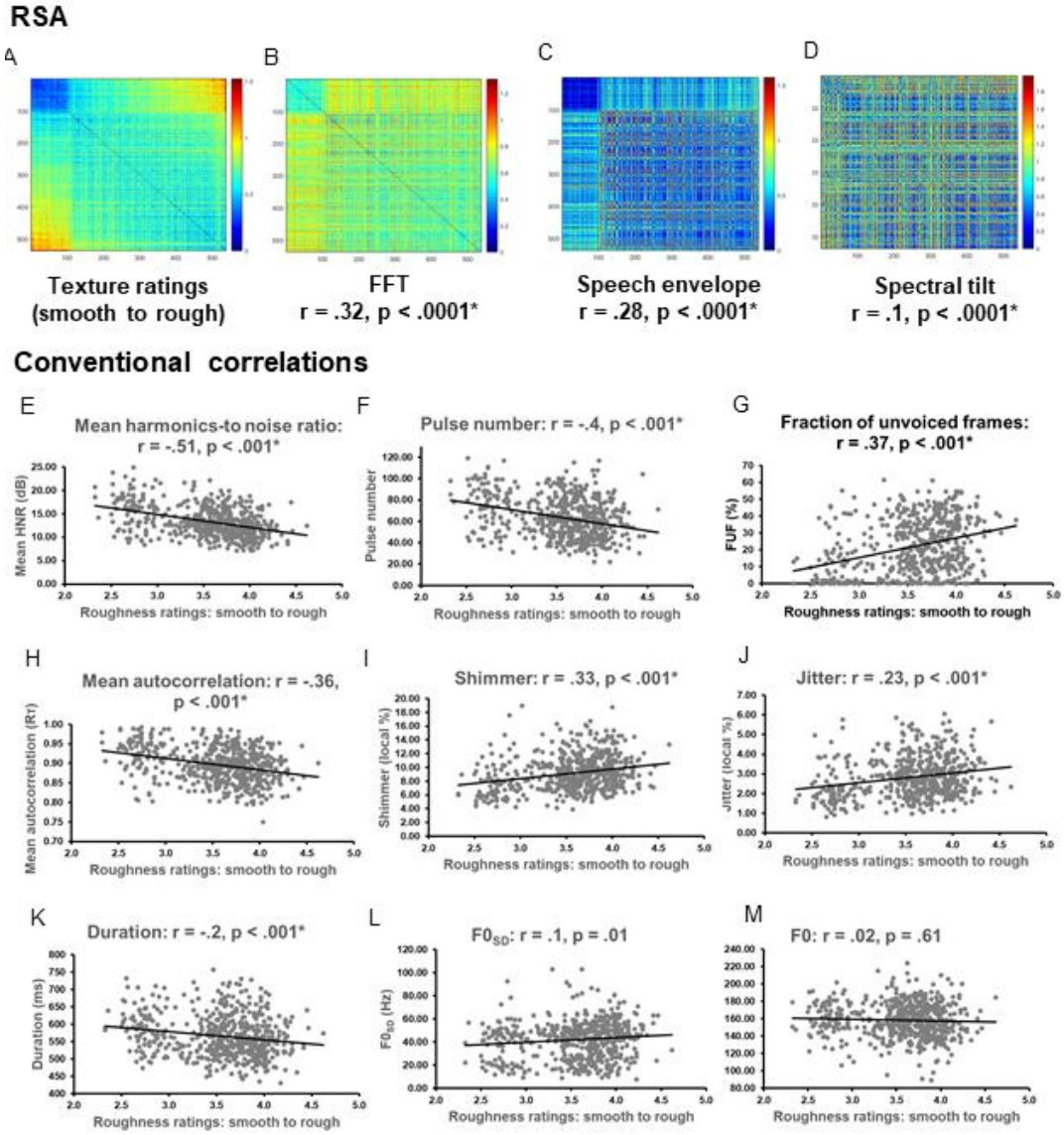

Figure 7: Representational similarity analyses and conventional correlations relating auditory roughness ratings and acoustic parameters. **A**: Ratings RDM. **B–D**: RDMs for the three spectro-temporal parameters. Pseudowords are ordered, left to right, from smoothest to roughest. Spectro-temporal parameter correlations are ordered, left to right, from the strongest to the weakest relationship. *correlation passes the Bonferroni-corrected α = .0167 for three tests; df = 143,914 for all three spectro-temporal parameters. **E-M**: Conventional correlations between auditory perceptual ratings for roughness and nine voice-quality parameters, ordered, top to bottom, left to right, from the strongest to the weakest relationship, whether positive or negative. *correlation passes the Bonferroni-corrected α of 0.0056 for nine tests; df = 535 for all nine voice parameters. All other details as for Fig. 3

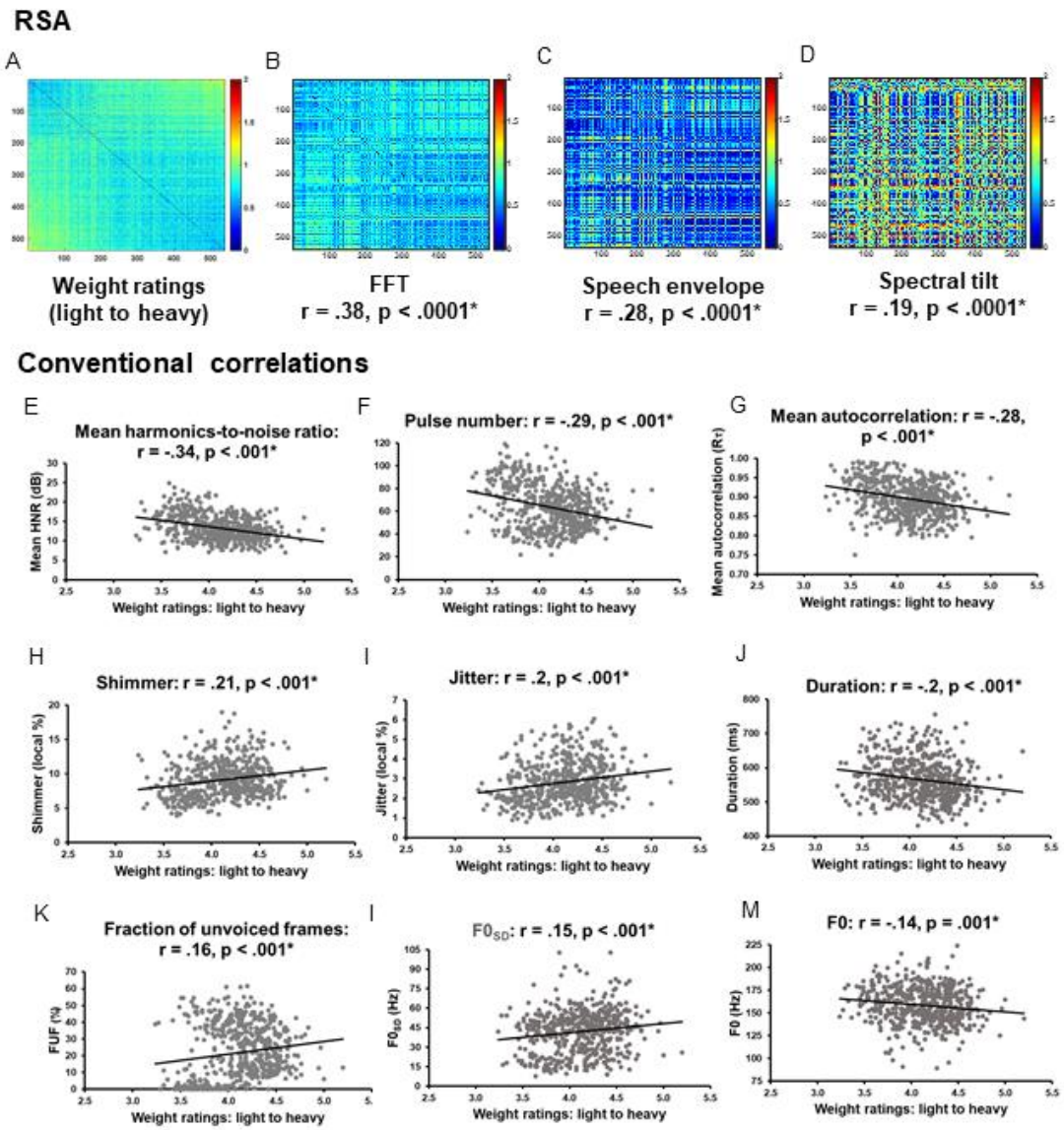

Figure 8: Representational similarity analyses and conventional correlations relating auditory ratings of weight and acoustic parameters. **A**: Ratings RDM. **B–D**: RDMs for the three spectro-temporal parameters. Pseudowords are ordered, left to right, from lightest to heaviest. Spectro-temporal parameter correlations are ordered, left to right, from the strongest to the weakest relationship. *correlation passes the Bonferroni-corrected α = .0167 for three tests; df = 143,914 for all three spectro-temporal parameters. **E-M**: Conventional correlations between auditory perceptual ratings of weight and nine voice-quality parameters, ordered, top to bottom, left to right, from the strongest to the weakest relationship, whether positive or negative. *correlation passes the Bonferroni-corrected α of 0.0056 for nine tests; df = 535 for all nine voice parameters. All other details as for Fig. 3

b. Conventional correlations for voice parameters

For voice parameters, the HNR (r = -.34, p < .001), pulse number (r = -.29, p < .001), and MAC (r = -.28, p < .001) were significantly negatively correlated with ratings of the pseudowords (**Figure**

8 E-M), indicating that periodicity decreased as ratings transitioned from light to heavy. The FUF (r = .16, p < .001) increased as ratings transitioned from light to heavy, also indicating decreasing periodicity as the proportion of unvoiced, aperiodic, segments increased. Similarly, shimmer (r = .21, p < .001) and jitter (r = .20, p < .001) were significantly positively correlated to the size ratings. F0 (r = -.14, p < .001) was significantly lower for heavy-, compared to light-rated, pseudowords. Finally, duration (r = -.2, p < .001) was negatively correlated with ratings, such that shorter items were rated as heavier than longer items.

c. Optimal combination

The KNN algorithm identified a combination of six parameters including three spectro-temporal parameters: FFT, spectral tilt, and speech envelope, and three voice-related parameters: FUF, HNR, and duration, that generated predicted ratings for the 537 pseudowords with the strongest correlation (r_s = 0.45) to the perceptual ratings provided by participants (N = 55) in the weight domain (**Table 1**).

7. Arousal

a. RSA for spectro-temporal parameters

Significant positive correlations with the ratings RDM were found for all three spectro-temporal parameter RDMs (**Figure 9 B-D**) although they were weaker than those for the shape, weight, and hardness domains. The highest correlation with the arousal ratings RDM was for the FFT RDM (r = .21, p < .001), followed by the speech envelope RDM (r = .18, p < .001) and then the spectral tilt RDM (r = .12, p < .001). Changes in the distribution of power across frequencies over time were more gradual for pseudowords rated as calming, compared to those rated as exciting, where the power changes were more abrupt (**Supplementary Figure 7a**). For calming-rated pseudowords, the speech envelope was relatively continuous, and spectral tilt was steep, reflecting the lower frequencies. For exciting-rated pseudowords, the speech envelope was more discontinuous, and spectral tilt was flatter as energy was dispersed across all frequency bands (representative pseudowords in **Supplementary Figure 7b & c**).

b. Conventional correlations for voice parameters

For voice parameters, HNR (r = -.42, p < .001), pulse number (r = -.30, p < .001), and MAC (r = -.26, p < .001) were negatively correlated with ratings, indicating that periodicity decreased as ratings transitioned from calming to exciting (**Figure 9 E-M**). The FUF was positively correlated (r = .40, p < .001) with ratings, meaning that the number of unvoiced segments increased from calming to exciting. Increasing variability in amplitude and frequency from cycle to cycle of the speech waveform, as measured by shimmer (r = .29, p < .001) and jitter (r = .14, p < .001) respectively, reflected the change in ratings from calming to exciting. However, F0 (r = .07, p = .1), $F0_{SD}$ (r = .07, p = .08), and duration (r = .02, p = .5) were not significantly correlated with arousal ratings.

c. Optimal combination

The optimal combination that produced ratings correlating best (r_s = 0.51, N = 58) with participants' perceptual ratings of arousal comprised five acoustic parameters, including the three

spectro-temporal parameters: FFT, spectral tilt and speech envelope; and two voice parameters: FUF, and HNR (**Table 1**).

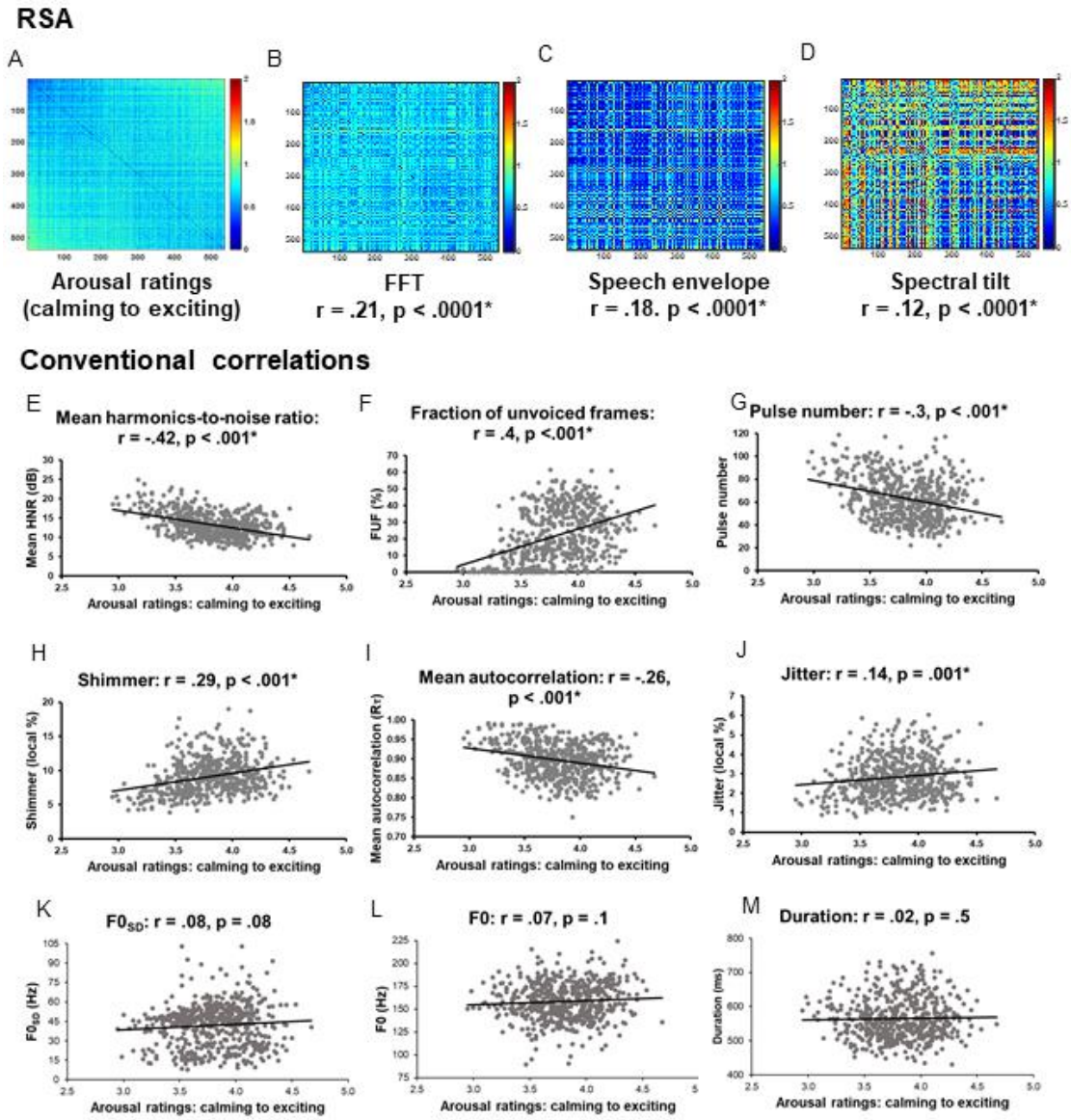

Figure 9: Representational similarity analyses and conventional correlations relating auditory ratings of arousal and acoustic parameters. **A**: Ratings RDM. **B–D**: RDMs for the three spectro-temporal parameters. Pseudowords are ordered, left to right, from calming to exciting. Spectro-temporal parameter correlations are ordered, left to right, from the strongest to the weakest relationship. *correlation passes the Bonferroni-corrected α = .0167 for three tests; df = 143,914 for all three spectro-temporal parameters. **E-M**: Conventional correlations between auditory perceptual ratings of arousal and nine voice-quality parameters, ordered, top to bottom, left to right, from the strongest to the weakest relationship, whether positive or negative. *correlation passes the Bonferroni-corrected α of 0.0056 for nine tests; df = 535 for all nine voice parameters. All other details as for Fig. 3

8. Valence

a. RSA for spectro-temporal parameters

Weak, but significant, positive correlations were found for all three spectro-temporal parameters RDMs with the ratings RDM (**Figure 10 B-D**). The FFT RDM had the highest correlation of the three (r = .17, p < .001), followed by the speech envelope RDM (r = .16, p < .001) and then the spectral tilt RDM (r = .07, p < .001). The spectrograms showed changes in the distribution of power across frequency over time that were more diffuse for pseudowords rated as good, compared to those rated as bad (**Supplementary Figure 8a)**. Similarly, the speech envelope was somewhat more continuous for good-rated, compared to bad-rated, pseudowords (**Supplementary Figure 8b**). A relatively flat spectral tilt was observed whether pseudowords were rated as good or bad (**Supplementary Figure 8c)**.

b. Conventional correlations for voice parameters

For voice parameters, the mean HNR (r = -.40, p < .001), pulse number (r = -.37, p < .001), and MAC (r = -.35, p < .001) were significantly negatively correlated with ratings, and FUF (r = .36, p < .001) was significantly positively correlated with ratings. Shimmer (r = .36, p < .001), jitter (r = .26, p < .001), and $F0_{SD}$ (r = .23, p < .001) were significantly positively correlated with ratings, reflecting increasing variability in voice quality for pseudowords perceived as bad, compared to those perceived as good. A significant negative correlation with ratings indicated that F0 (r = -.20, p < .001) was lower for bad-rated, compared to good-rated, pseudowords. Duration (r = .16, p < .001) was not significantly correlated with the valence ratings (**Figure 10 E-M**).

c. Optimal combination

The KNN algorithm identified an optimal combination comprising the two spectro-temporal parameters: FFT and speech envelope, that produced ratings for the 537 pseudowords which correlated (r_s =.66) best with the perceptual ratings (N = 49) of pseudowords in the valence domain (**Table 1**).

B. Predicting sound-symbolic shape ratings of real words

We used the multidimensional space defined by the five acoustic parameters that best predicted the shape ratings of the pseudowords (see section III.A.1) to predict the shape ratings of real words. The predicted rating of each real word was based on the modal perceptual rating of the 23 pseudowords in its acoustic neighborhood. Since the pseudowords were rated by 60 participants, the mean across-participant pseudoword ratings were used to predict ratings of the real words. The predicted ratings were strongly correlated ($r_p = 0.52, p < 0.001$) with mean perceptual ratings of the real words.

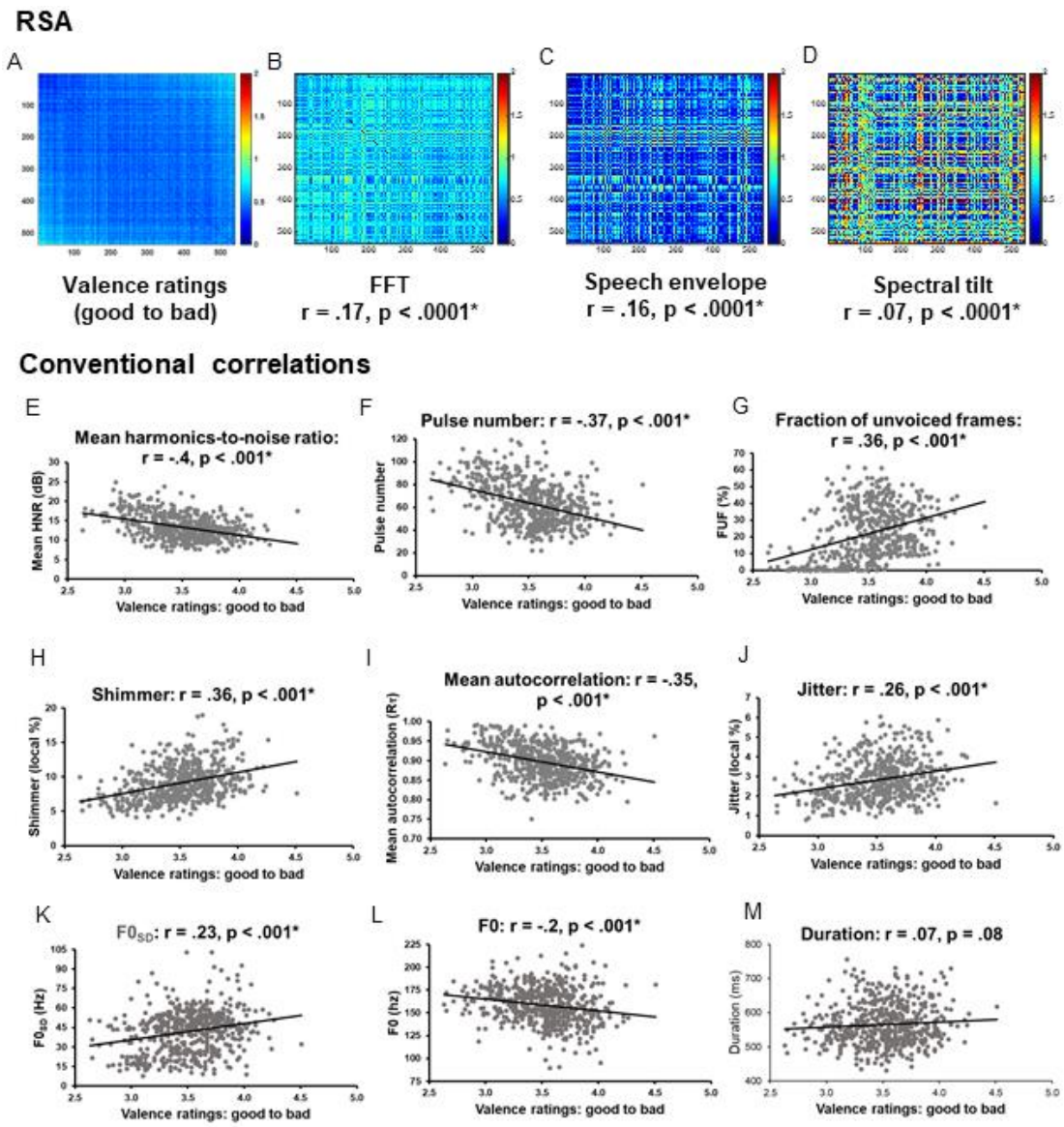

Figure 10: Representational similarity analyses and conventional correlations relating auditory ratings of valence and acoustic parameters. **A**: Ratings RDM. **B–D**: RDMs for the three spectro-temporal parameters. Pseudowords are ordered, left to right, from good to bad. Spectro-temporal parameter correlations are ordered, left to right, from the strongest to the weakest relationship. *correlation passes the Bonferroni-corrected α = .0167 for three tests; df = 143,914 for all three spectro-temporal parameters. **E-M**: Conventional correlations between auditory perceptual ratings of valence and nine voice-quality parameters, ordered, top to bottom, left to right, from the strongest to the weakest relationship, whether positive or negative. *correlation passes the Bonferroni-corrected α of 0.0056 for nine tests; df = 535 for all nine voice parameters. All other details as for Fig. 3

Table 1: Summary of KNN results

FFT: Fast Fourier transform, ST: Spectral tilt, SE: Speech envelope, PN: Pulse number, FUF: Fraction of unvoiced frames, HNR: Harmonics-to-noise ratio, MAC: Mean autocorrelation

Domain	N	Max. r_s	Model
Shape	60	0.59	FFT, ST, FUF, shimmer and duration
Size	63	0.45	FFT, ST, SE, FUF and duration
Brightness	50	0.52	FFT, ST, SE, PN, FUF, jitter, MAC, F0, $F0_{SD}$ and duration
Hardness	54	0.54	FFT, ST, FUF, shimmer and duration
Roughness	58	0.57	FFT, SE
Weight	55	0.45	FFT, ST, SE, FUF, HNR and duration
Arousal	58	0.51	FFT, ST, SE, FUF and HNR
Valence	49	0.66	FFT and SE

IV. 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. The acoustic parameters that underpin sound-symbolism have been less well studied. Prior studies investigating acoustic properties that mediate sound symbolism suggest that the influence of these properties varies by semantic domain[23,29]. However, the associations of acoustic properties with sound-symbolic properties have not previously been investigated systematically across a range of semantic domains. Here, we assessed the relationship between each of twelve acoustic parameters and sound-symbolic ratings in eight different meaning domains, using RSA for the three spectro-temporal parameters studied and conventional correlations for nine parameters associated with voice quality.

To our knowledge, there are no published reports attempting to evaluate the combinations of acoustic parameters that are optimally predictive of sound-symbolic associations. 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 supervised KNN machine-learning approach and found that domain-specific combinations of acoustic parameters yielded predicted ratings that correlated with the perceptual ratings of pseudowords across the eight different meaning domains. Finally, as a proof of concept, we used the KNN approach to predict the sound-symbolic shape ratings of real words based entirely on their acoustic properties, unconfounded by semantic bias.

We chose the KNN algorithm because it is well suited to our idea that pseudowords with similar sound-symbolic ratings cluster together in a multidimensional acoustic parameter space. Unlike approaches relying on methods such as support vector machines (SVM), regression, or deep learning, which require fitting a global decision boundary, a parametric mapping, or a complex transformation of the input features, KNN operates directly on the local geometry of the parameter space. This makes it easily interpretable in the context of our goal: if two pseudowords are acoustically similar, they should be close in the parameter space and therefore have similar ratings, and our KNN based algorithm captures this relationship. Moreover, KNN's distance-based metric allows us to quantify acoustic proximity, which in turn reflects perceptual similarity (of the real words). While SVM and regression models can perform well for prediction, they tend to be more or less opaque regarding their underlying neighborhood structure; and deep learning methods, though powerful, would require substantially larger datasets, risk overfitting, and make interpretability of specific parameter contributions more challenging. In contrast, KNN allows us to interpret the multidimensional clusters of pseudowords, making it a theoretically and methodologically appropriate choice for our study.

Shape

Consistent with our previous findings[29], RSA demonstrated strong relationships between specific spectro-temporal acoustic parameters and shape-based sound symbolic ratings. Spectral tilt showed the strongest alignment with roundedness judgments, with greater energy concentrated in lower frequencies for pseudowords rated as more rounded and higher-frequency energy dominating for pointed pseudowords rated as more pointed (**Figure 3 B-D**). This pattern likely reflects underlying phonetic contrasts in sonorant versus obstruent content, as well as the influence of place of articulation, with bilabial and labiodental consonants contributing to lower-frequency spectra and alveolar or velar consonants contributing to higher-frequency components[9]. Similarly, FFT patterns changed more gradually over the duration of the pseudoword for rounded-rated pseudowords and more abruptly for pointed-rated ones, further distinguishing the smoothness of sonorants from the sharp onsets of obstruents. The speech envelope also showed a modest but consistent effect, being smoother and more continuous for rounded pseudowords rated as more rounded, consistent with sustained vocal energy and reduced temporal variability.

For the voice parameters, we replicated earlier findings that reduced periodicity (lower HNR, fewer pulses, lower MAC) and increased variability of the acoustic waveform (higher shimmer and jitter) were associated with pointedness, while more regular, periodic signals were associated

with roundedness (**Figure 3 E-M**). The positive correlation between the fraction of unvoiced frames and pointed ratings further supports this link. Unlike our earlier study, $F0_{SD}$ also emerged as a significant factor, with greater variability marking more pointed items. Of the two newly examined parameters, duration was inversely related to pointedness, aligning with underlying phonetic contrasts in sonorant versus obstruent content for rounded and pointed pseudowords, while F0 was not significantly correlated with ratings. Together, these findings provide a robust replication of our prior results and offer some new insights into the acoustic dimensions that underlie sound-symbolic associations in the rounded–pointed domain. Importantly, they highlight how increasing acoustic variability, temporal irregularity, and spectral shift toward higher frequencies all contribute to the perceptual transition from rounded to pointed.

We found that shape ratings were best predicted by a combination of five parameters including two spectro-temporal parameters: the FFT and spectral tilt, and three voice parameters: the FUF, shimmer and duration (**Table 1**). Existing studies show that formant frequencies are higher for more pointed words[23] and for vocalizations produced in response to more pointed shapes[53], 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[54]. These formant properties may be related to the influence of spectral tilt and 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, whereas the FFT reflects the power spectrum across the duration of the spoken pseudowords (see representative pseudowords **Supplementary Figure 1a**), capturing the instantaneous spectral profile as the pseudoword unfolds in time.

The duration of the speech signal carries information related to the component vowels[48,55] 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 the coarticulated phoneme and in turn, the 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[919], supporting the importance of the FUF that characterizes features related to the degree of voicing in speech. The presence of shimmer in the model is consistent with the importance of the amplitude profile, as shimmer indexes the peak-to-peak variation in the amplitude of the periodic speech signal, implying that momentary amplitude variations are important in mediating sound-symbolic judgements of shape. Importantly, these duration effects are probably not limited to English; duration is a cross-linguistic prosodic cue[56]. For example, final lengthening, where syllables or segments at the ends of words and phrases are universally produced with extra duration, providing listeners with a robust cue for speech segmentation and perception[56,57]. However, the pseudowords

in our set were produced with neutral intonation, so the observed duration differences are more likely tied to intrinsic vowel–consonant dynamics.

Size

RSA demonstrated that spectral tilt, speech envelope, and spectrogram RDMs were all weakly but significantly correlated with the size ratings RDM, with spectral tilt showing the strongest association (**Figure 4**). Visual inspection of the RDMs suggested these effects were diffuse rather than localized. Both small and big rated pseudowords showed broadly similar spectral profiles and discontinuous speech envelopes, with only minor differences in energy distribution and envelope amplitude. No voice parameters, including duration, significantly predicted size ratings after correction. These findings suggest that while there may be some spectro-temporal cues contributing to size-related sound symbolism, the associations were neither strong nor straightforward in the present pseudoword set. The weak correlation with spectral tilt indicates that both small-rated and big-rated pseudowords contain a mix of low- and high-frequency energy due to the combination of voiced/unvoiced consonants and their places of articulation[9]. In small-rated pseudowords, the use of unvoiced consonants (typically high-frequency) at bilabial/labiodental places may balance out to produce mid-range spectral characteristics. Similarly, big-rated pseudowords incorporate both high-frequency fricatives and lower-frequency features from post-alveolar/velar articulations and back rounded vowels, leading to a comparable spectral profile overall. The discontinuous speech envelopes observed across both size types likely reflect the presence of stops, which are common to both. However, the slightly greater amplitude for big pseudowords, particularly in the second syllable, may stem from voicing differences, which influence overall intensity and periodicity. These subtle acoustic variations may contribute marginally to size perception, even if they are not robust predictors on their own.

The lack of significant associations between voice parameters and size ratings, including duration, might be because conventional correlations may not adequately capture the underlying sound-size relationships for this specific pseudoword set. It remains to be determined whether this pattern reflects characteristics of the particular pseudowords used or idiosyncrasies of the speaker, which should be addressed in future research. While prior work has linked duration and phonation type to size symbolism, these effects may depend on intentional exaggeration by speakers or specific experimental contexts. In our study, the speaker used natural intonation and voicing, which may have muted such associations. Taken together, these results highlight that while certain acoustic properties may weakly reflect sound-symbolic size, they do not provide strong or consistent cues. The overlap in spectral and temporal features across small and big pseudowords complicates any simple mapping between form and meaning, underscoring the need for further research into more fine-grained or context-sensitive acoustic features that might better capture these sound-symbolic effects.

The 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 (**Table 1**). Evidence suggest

that 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[58–60]. 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[17]. Furthermore, small/big ratings were found to be associated with unvoiced/voiced consonants[9], which are characterized by distinctive spectral profiles. Thus, a pseudoword's spectral characteristics indexed by FFT and spectral tilt feature in the combination. Anikin and colleagues [61] found that vocal roughness exaggerates the apparent body size of the vocalizer[62]. Since the proportion of unvoiced segments is indicative of vocal roughness, the FUF is an important parameter. Previously, Knoeferle et al.[23] showed that the duration of vowels in pseudowords influence sound-size judgements, however, with the current set of pseudowords, we found that bigness, but not smallness, ratings were associated with the back rounded vowels[9]. Thus, pseudoword duration, which is partly contingent on the component vowels, is relevant for size ratings. Parameters representing frequency variability, e.g., jitter and $F0_{SD}$; amplitude variability, e.g., shimmer; periodicity, e.g., HNR, pulse number and mean autocorrelation; and fundamental frequency, e.g., F0, did not feature in the combination.

Although no single vocal parameter showed a significant conventional correlation with size judgments on its own, these parameters may have contributed to KNN model for size because the model exploits joint patterns across features. Even if each parameter contributes only minimal information individually, the combined interactions can capture more complex relationships that support sound-symbolic judgments.

Brightness

The RSA results indicated that spectro-temporal features captured aspects of the bright–dark distinction in pseudowords (**Figure 5 B_D**). Spectral tilt was flatter for brighter pseudowords, consistent with greater energy at higher frequencies, while darker pseudowords showed energy concentrated in lower-frequency bands (**Supplementary Figure 3a**). These differences likely reflect the association of unvoiced consonants and front unrounded vowels with bright pseudowords, and voiced consonants and back rounded vowels with dark. For the FFT, although the spectrogram did show greater power in the higher frequencies for bright, compared to dark, pseudowords (consistent with the difference in spectral tilt), there were relatively small differences in the abruptness of power variations across frequencies (**Supplementary Figure 3b**), reflecting that the main difference between bright and dark pseudowords was voicing rather than manner of articulation. The speech envelope was discontinuous for both bright and dark pseudowords, but amplitude was perhaps greater for the latter (**Supplementary Figure 3c**), reflecting that bright/dark pseudowords tend to involve unvoiced/voiced consonants, respectively.

Our voice parameter results align with prior work on crossmodal correspondences between auditory and visual dimensions. Brighter pseudowords tended to have higher F0, whereas darker pseudowords were characterized by lower F0, consistent with the association between high F0 and brightness and low F0 and darkness. In line with observations linking voicing and darkness[7,9,63],

pseudowords rated as darker also showed greater periodicity, a higher proportion of voiced segments, and elevated HNR and pulse counts. Duration tended to be longer for darker items, similar to Tzeng et al.[64], although this effect was less robust, potentially reflecting task or stimuli differences. By contrast, measures of F0 and amplitude variability (jitter, $F0_{SD}$, shimmer, MAC) did not contribute in a meaningful way, consistent with the view that brightness and darkness mappings are primarily driven by features that reflect overall spectral and temporal properties of the signal rather than their variability.

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 seven voice parameters: pulse number, FUF, jitter MAC, F0, $F0_{SD}$ and duration (**Table 1**). Lacey et al. found that, similar to the size domain, consonant voicing and place of articulation were key determinants of the brightness ratings of pseudowords[9]. These characteristics change the frequency profiles of the speech signal such as shifting energy toward lower or higher bands and are therefore captured by the FFT, explaining its presence in the combination. Additionally, consonant voicing and vowel category affect spectral tilt, e.g., voiced consonants and back rounded vowels have more energy in the lower frequencies, while unvoiced consonants and front unrounded vowels have more energy in the higher frequencies[20] explaining the importance of spectral tilt in the combination. The temporal evolution of pseudowords also reliably captures information related to vowel rounding, which influences brightness ratings[9]. This is represented by speech envelope which explains its presence in the model. Furthermore, information related to voicing, represented by FUF, is crucial because the key difference between pseudowords rated as bright vs. dark is the presence of unvoiced vs. voiced consonants[9]. The contribution of F0 and $F0_{SD}$ is consistent with the finding of Hubbard that participants match high-pitched sounds to lighter, and low-pitched sounds to darker, stimuli[65]. As in for the shape domain, bright/dark pseudowords were characterized by the presence of unrounded/rounded vowels, which influences the duration of pseudowords, thus explaining the relevance of duration in brightness judgements. The inclusion of MAC, pulse number and jitter reflect that measures of periodicity and stability of vocal fold vibration mediate sound-brightness judgements. Pulse number is influenced by phonetic properties such as consonant voicing and vowel category: voiced consonants and vowels involve regular vocal fold vibration, resulting in more glottal pulses, whereas voiceless consonants introduce aperiodic segments with fewer pulses. Jitter and mean autocorrelation capture regularity in the waveform; greater regularity (high mean autocorrelation and low jitter) is associated with voiced and sonorant segments, while irregular or absent periodicity reflects unvoiced segments. Titze [66] and Boersma[67] have shown that measures like jitter and MAC robustly quantify voicing stability, and are sensitive to the presence/absence of vocal fold vibration across segments. Thus, their inclusion provides cues that distinguish voiced vs. unvoiced contrasts that appear to be central to brightness judgments. Parameters related to amplitude variability, e.g., shimmer and periodicity, e.g., HNR did not feature in the combination for brightness.

Hardness

RSA revealed that ratings along the hard-to-soft dimension were significantly associated with spectro-temporal features (**Figure 6 B-D**). Among these, the FFT and speech envelope RDMs showed the strongest correlations with the perceptual ratings RDM, followed by the spectral tilt RDM. These results indicate that hard-rated pseudowords exhibited more energy at higher frequencies and greater discontinuities in their spectro-temporal profiles, while soft-rated pseudowords were marked by smoother, more continuous acoustic patterns and energy concentrated at lower frequencies.

In terms of vocal features, higher perceived softness was associated with greater periodicity and reduced variability in the vocal signal. Specifically, HNR, pulse number, and MAC were positively correlated with softness ratings, while FUF, shimmer, and jitter were negatively correlated (**Figure 6 E-M**). These findings suggest that softness is signaled by acoustic regularity and smoothness, whereas hardness is conveyed through aperiodicity and vocal roughness. Together, these results support the view that sound-symbolic hardness judgments are grounded in systematic acoustic cues. The sharper spectral transitions and high-frequency emphasis characteristic of hard-rated pseudowords align with the presence of obstruents, particularly fricatives, which are noisy and aperiodic[68]. In contrast, the acoustic smoothness of soft-rated pseudowords reflects the properties of sonorants, which are more periodic and less acoustically variable[9]. F0 and duration did not show consistent relationships with ratings indicating that these parameters do not carry cues that mediate sound-symbolic judgments of hardness.

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, produced ratings that best correlated with the perceptual ratings of hardness of the pseudowords (**Table 1**). The place of articulation of the sonorants significantly predicts softness ratings, while voicing and manner and place of articulation are key predictors of hardness ratings[9,69,70]. These features have very distinct spectro-temporal profiles, effectively captured by the FFT, explaining its importance in the combination. While these finer details are crucial for sound-hardness associations, hard- and soft-rated pseudowords can also be distinguished by the broad phonetic categories of obstruents and sonorants[9], respectively, which are reflected in 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. 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 [69]. Such phonemes are characterized by distinctive temporal variations in amplitude, measured by shimmer, and by durational cues. As for other domains, parameters that indicate frequency variability, e.g. jitter and $F0_{SD}$; the amplitude profile, i.e. the speech envelope; and periodicity, e.g. HNR, pulse number and MAC; and F0 did not contribute to predicting hardness ratings.

Roughness

The RSA revealed that all three spectro-temporal parameter RDMs: for the FFT, speech envelope, and spectral tilt, were significantly correlated with the perceptual ratings RDM in the roughness domain, with FFT showing the strongest correspondence (**Figure 7 B-D**). These associations are consistent with studies linking phonetic and acoustic structure to tactile perception of smoothness. Pseudowords rated as smooth therefore tend to feature sonorant-dominated segments along with voiced stops and fricatives[69,71]. These are produced with relatively open and stable vocal-tract configurations, yielding harmonic-rich, periodic signals and gradual formant transitions, which manifest as continuous energy patterns across time and frequency[72]. By contrast, rough-rated pseudowords more frequently include voiceless obstruents such as stops and fricatives with turbulent airflow and abrupt articulatory closures, leading to aperiodicity and irregular amplitude envelopes[73,74]. This articulatory difference accounts for the observed acoustic profiles: smoother rated items exhibit gradual, even energy transitions and more sustained amplitude envelopes (**Supplementary Figure 5a–b**), while items rated as rougher show more discontinuous, abrupt temporal–spectral patterns. The higher-frequency energy concentration in smooth-rated items (**Supplementary Figure 5c**) is also consistent with the formant characteristics of sonorants and voiced segments versus the broadband noise bursts typical of voiceless obstruents[75].

Vocal parameters provide further support for these phonetic interpretations (**Figure 7 E–M**). Measures reflecting periodicity, such as HNR and voiced pulse number, were higher for smooth-rated items and decreased toward the rough end of the scale, consistent with the greater proportion of unvoiced and aperiodic segments in rough pseudowords. Conversely, indicators of voicing instability, including jitter and shimmer, were elevated for rough-rated items, reflecting the articulatory turbulence and irregular vocal fold vibration characteristic of voiceless obstruents. Duration also declined toward the rough end, suggesting that shorter, more temporally compact words may be perceived rougher, in line with previous findings that abrupt articulatory gestures cue tactile-like roughness[76]. In contrast, the mean F0 and its variability ($F0_{SD}$ and jitter) appeared less central to smooth–rough judgments, suggesting that frequency per se is not a primary driver of sound -symbolic roughness.

A combination of features represented in two spectro-temporal parameters: the FFT and the speech envelope, generated ratings that showed the highest correlation with the perceptual ratings for the roughness of the pseudowords (**Table 1**). Prior research indicates that, as noted above, smooth items are dominated by sonorants and voiced segments, comprising gradual formant transitions and sustained amplitude the features well captured by FFT and speech envelope. In contrast, rough items rated as rough are characterized by voiceless obstruents that show broadband noise bursts and abrupt energy changes, resulting in jagged, discontinuous envelopes. Because roughness perception depends on the balance between gradual/abrupt spectral energy and temporal continuity, the combination of information represented in the FFT and speech-envelope carries cues that mediate sound-symbolic judgements of roughness.

Weight

The RSA revealed that all three spectro-temporal parameter RDMs: for the FFT, speech envelope, and spectral tilt, were significantly correlated with the perceptual ratings RDM in the weight domain, with FFT showing the strongest correspondence (**Figure 8 B-D**). Vocal parameters associated with periodicity, such as HNR, pulse number, MAC, and FUF, also tracked the transition from light to heavy ratings, as did F0 and duration (**Figure 8 E-M**). Phonetic analysis suggests that perceived lightness is linked with sonorants and the bilabial/labiodental place of articulation, accompanied by unrounded vowels, whereas perceived heaviness aligns with voiced obstruents and post-alveolar/velar articulation, along with rounded vowels[9]. This distribution of articulatory features likely accounts for the observed spectro-temporal profiles, where sonorants contribute to smoother, more continuous energy distributions over time and frequency, while obstruents yield more abrupt, discontinuous patterns. The lack of a strong association between spectral tilt and perceptual ratings is consistent with the presence of both low- and high-frequency energy across light-rated and heavy-rated pseudowords, stemming from different articulatory sources. Given the real-world association between size and weight, analogies with the size domain provide interpretive insight. Higher pitch, commonly linked to smaller size, is consistent with lighter-rated pseudowords, whereas lower pitch aligns with perceived heaviness. Measures of periodicity decrease from light to heavy ratings, reflecting a shift from voiced to unvoiced segments, in line with the broader transition from sonorants to obstruents. Vocal variability metrics such as jitter, shimmer, and $F0_{SD}$, demonstrated positive relationships with heaviness. This may point to increased articulatory effort or irregularity in voicing patterns among heavier-rated items. The negative association between duration and heaviness raises the possibility that listeners may interpret shorter words as denser and therefore heavier, an idea that invites further investigation with direct assessments of perceived density.

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. Interestingly, we observed that, except for HNR, all other parameters in the model were the same as for the size domain **(Table 1)**. Notably, Lacey et al. [9] 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 that weight and size mostly co-vary, i.e. big objects in most cases are heavier than small objects. However, it is important to note that this relationship is not absolute, since weight also depends on density. As a result, a large but low-density object can weigh less than a smaller, high-density one. Furthermore, Lacey et al. showed that the combinations of vowel rounding and the consonant features that influence light and heavy ratings are unique[9]. 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. Fricative sounds (e.g., /f/) involve turbulent airflow, resulting in lower HNR. In contrast, voiced consonant sounds (e.g., alveolar: /d/ and velar /g/) have a more periodic nature and less turbulent airflow. Since the HNR indicates the proportion of periodic and aperiodic components in speech, it significantly influences weight ratings of pseudowords.

Sonorants and obstruents that determine the lightness and heaviness of the pseudowords, respectively, show distinct spectral tilts, accounting for the influence of spectral tilt. A characteristic distinction between these two classes of phonemes (i.e., sonorants and obstruents) is also their voicing. The FUF captures sensitivity to voicing, while the speech envelope reflects differences between obstruents and sonorants through amplitude. Together, these measures highlight that the proportion of voiced segments and their amplitude within an utterance play a crucial role in weight judgments. The presence of duration might be reflective of the influence of the component vowels. As for other domains, parameters that feature frequency variability, e.g. jitter and $F0_{SD}$; amplitude variability, e.g. shimmer; and periodicity, e.g. the HNR, pulse number and mean autocorrelation; and F0, did not contribute to predicting weight ratings.

Arousal

The ratings RDM showed the strongest correspondence with the FFT RDM, followed by the speech envelope and spectral tilt RDMs (**Figure 9 B-D**). For voice parameters, the FUF increased with arousal, while the mean HNR, pulse number, and MAC decreased. Jitter and shimmer rose with arousal, whereas F0, $F0_{SD}$, and duration showed no clear relationship (**Figure 9 E-M**). These patterns align with established phonetic associations that pseudowords rated as calming are more likely to contain sonorants and back rounded vowels, while those rated as exciting more often feature unvoiced obstruents and front unrounded vowels[9,70]. Calming ratings are also strongly linked to bilabial/labiodental articulation, which concentrates energy in the lower frequencies, whereas exciting ratings are not tied to a single place of articulation. Consistent with these phonetic characteristics, calming-rated pseudowords displayed gradual changes in spectral energy over time, continuous speech envelopes, and steep spectral tilt, reflecting the sustained amplitude and low-frequency dominance of sonorants and bilabials. Exciting-rated pseudowords, by contrast, showed abrupt spectral changes, discontinuous envelopes characteristic of stops, and flatter spectral tilt with energy spread across frequencies (see representative pseudowords **Supplementary Figure 7**), including the high-frequency components generated by fricatives[72].

The voice parameter trends also fit with these phonetic distinctions. The increase in FUF and reduction in HNR, pulse number, and MAC from calming to exciting ratings reflect a shift from voiced to unvoiced consonants, with reduced periodicity[9]. The rise in jitter and shimmer is consistent with findings that general, non-emergency arousal tends to increase vocal variability[77]. In contrast, voice stress research in emergency contexts often reports the opposite pattern, likely due to deliberate adjustments for clarity and volume that reduce variability[78]. The lack of clear effects for F0, $F0_{SD}$, and duration suggests that spectro-temporal and periodicity features carried more information about arousal than prosodic measures in the present study. Finally, our arousal scale was positively valenced overall, with both “calm” and “exciting” at the high-positive end of the Warriner et al. database[79]. Using a restful-stressful scale might yield a sharper distinction between positive and negative arousal dimensions and reveal different acoustic correlates.

A combination of five parameters comprising all three spectro-temporal parameters: the FFT, spectral tilt and speech envelope and two voice parameters including the FUF and HNR generated

ratings that best correlated with the perceptual ratings in the arousal domain (**Table 1**). Pseudowords rated as calming comprise sonorants and back rounded vowels while those rated as exciting primarily comprise unvoiced obstruents and front unrounded vowels[9,70]. Spectro-temporal features that differentiate sonorants from obstruents, as well as those related to voicing, are effectively captured by the FFT as well as the speech envelope. Additionally, spectral tilt captures the broad distinction in the spectral profile of sonorants and obstruents. This explains the presence of the spectro-temporal parameters in the optimal combination. Affective states induced by stress have been shown to influence the HNR in speech produced during these states[80]. Additionally, the presence of obstruents increases turbulent airflow at the glottis during phonation, leading to aperiodic vocal fold vibration, which impacts the HNR. Also, voicing of obstruents underlie the tendency for pseudowords to be rated as exciting, explaining the significance of the FUF in the model.

Valence

The RSA showed weak but significant positive correlations between the ratings RDM and all three spectro-temporal parameter RDMs, with FFT showing the strongest association, followed by speech envelope, and then spectral tilt (**Figure 10 B-D**). The stronger relationship for FFT may reflect the differing phonetic composition of good- and bad-rated pseudowords[9]: good-rated pseudowords, containing more sonorants and unrounded front vowels, tended to exhibit more gradual, diffuse spectral changes, whereas bad-rated pseudowords, containing more obstruents and rounded back vowels, displayed more abrupt spectral transitions (see representative pseudowords **Supplementary Figure 8**). The speech envelope findings were consistent with this pattern, being smoother and more continuous for pseudowords rated as good and more discontinuous for those rated as bad, reflecting the amplitude and temporal characteristics of their constituent phonemes. The very low correlation for spectral tilt aligns with the observation that both good- and bad-rated pseudowords contain energy at high and low frequencies, producing minimal differences in the overall spectral slope.

For voice parameters, lower HNR, pulse number, and MAC, together with higher FUF, shimmer, jitter, and PSD, were associated with pseudowords rated as bad (**Figure 10 E-M**). These measures indicate greater irregularity and instability in voice quality when pseudowords are perceived negatively. F0 was lower for bad-rated pseudowords, while duration showed no relationship with ratings. Overall, the results suggest that listeners' valence judgments of pseudowords are associated with systematic differences in both segmental composition and voice quality. The combination of phonetic structure, spectro-temporal patterns, and voice parameters appears to contribute to the perceptual mapping of pseudowords along a good-bad continuum.

A combination of features represented in two spectro-temporal parameters: the FFT and the speech envelope, generated ratings that showed the highest correlation with the perceptual ratings for the valence of the pseudowords (**Table 1**). 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 the incidence of sonorants/obstruents[9]. Spectro-temporal features that differentiate sonorants from obstruents, as well as those related to voicing, are effectively captured by the FFT as well as the speech envelope. Parameters that feature frequency variability, e.g. spectral tilt, jitter, F0, $F0_{SD}$; amplitude variability, e.g. shimmer; periodicity, e.g., FUF, pulse number and mean autocorrelation; and duration 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 (**Table 1**). Interestingly the spectro-temporal parameters featured in most of the meaning domains: FFT in all and spectral tilt and speech envelope in six, 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. Importantly, the phonetic features related to voicing, manner and place of articulation, and information related to vowel formants are comprehensively represented in the FFT but only partially in spectral tilt as FFT captures fine-grained energy distribution across the entire frequency range, allowing identification of specific formants and spectral patterns that carry information related to place, manner, and voicing. In contrast, spectral tilt summarizes the overall slope of the spectrum and reflects a more general characterization of the frequency distribution, so it lacks the detailed frequency-specific information over time[81]. The speech envelope captures the alternation between high-energy voiced segments such as sonorants and vowels, with sustained periodic energy, and the transient bursts and amplitude dips of obstruents including stops, fricatives and affricates[82]. In this way, the temporal envelope provides a coarse but informative representation of voicing and manner contrasts, complementing the spectral detail of the FFT and the broader distribution captured by spectral tilt. These features are also significant from a neural standpoint because the initial stages of speech perception involve processing the spectro-temporal properties represented by these two parameters[83,84]. 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[85]. We observed that the FUF, which measures the percentage of time windows that do not engage the vocal folds, featured in the combination of features for six meaning domains (**Table 1**). The FUF provides cues about the phonemic content of an utterance[41], 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[9,19] and therefore, sound-symbolic judgments across multiple domains are sensitive to features reflected in this parameter.

We found that stimulus duration featured in five domains. Our findings extend and generalize those of Knoeferle et al. demonstrating the significance of durational cues in sound-size symbolism[23]. 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[86] and duration influencing the perception of vowel quality[55]. 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.

The MAC, pulse number, jitter, F0 and $F0_{SD}$ featured in the optimal combination only for the brightness domain. This suggests that these parameters may carry specific information relevant for distinguishing brightness-related sound-symbolic judgments, rather than being broadly representative of acoustic-phonetic properties across meaning domains. For instance, both the FUF and pulse number capture aspects of phonation or voicing in the pseudoword. However, FUF reflects the relative proportion of unvoiced to voiced periods, whereas pulse number simply reports the number of glottal pulses in the signal. It is possible that for the current set of pseudowords, brightness-related distinctions rely more on absolute measures of voicing (pulse number) than on relative measures (FUF). Similarly, MAC and jitter, which index the regularity of vocal fold vibrations, may be particularly relevant for perceiving fine-grained periodicity or stability cues that listeners associate with brightness, even if these cues are not crucial in other meaning domains. In this way, brightness judgments might rely on a specific set of voicing stability and periodicity features, captured by the MAC, jitter, and pulse number, that do not play the same role for other domains. Whether the brightness-related relevance of these features reflects properties of the pseudoword set, the speaker, or both requires further investigation. The contribution of F0 and $F0_{SD}$ is consistent with the finding of Hubbard (1996) that participants match high-pitched sounds to lighter, and low-pitched sounds to darker, stimuli[65].

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[32,87]. 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[88,89]. 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 SVMs 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[90,91]. 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[92]. 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[92,93], a potential remedy for children facing difficulty in language learning could involve adjusting the inputs given to them, to include more sound-symbolic words. Although recent work suggests that the role of iconicity in early stages of language acquisition may be more limited than previously thought[94], we propose, nonetheless, that sound symbolic cues may still provide a valuable scaffold for preschool children experiencing language-learning difficulties, whether or not they play a primary role in typically developing infants. 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, where sound-symbolic associations seem to be preserved[95,96].

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 [97,98]. 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[99]. 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.

COMPETING INTERESTS

The authors declare that they have no competing interests.

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

1. F. de Saussure, *Course in General Linguistics* (McGraw-Hill, 1966).
2. J. O. Svantesson, "Sound symbolism: the role of word sound in meaning," *Wiley Interdiscip. Rev. Cogn. Sci.* 8 (2017).
3. W. Köhler, *Gestalt Psychology* (Liveright, New York, 1929), available at: https://www.biblio.com/gestalt-psychology-by-kohler-dr-wolfgang/work/2062021.
4. V. S. Ramachandran and E. M. Hubbard, "Synaesthesia – A window into perception, thought and language," *J. Conscious. Stud.* 8, 3–34 (2001).
5. E. Sapir, "A study in phonetic symbolism," *J. Exp. Psychol.* 12, 225–239 (1929).
6. P. Walker and C. R. Parameswaran, "Cross-sensory correspondences in language: Vowel sounds can symbolize the felt heaviness of objects," *J. Exp. Psychol. Learn. Mem. Cogn.* 45, 246–252 (2019).
7. S. Hirata, J. Ukita, and S. Kita, "Implicit phonetic symbolism in voicing of consonants and visual lightness using Garner's speeded classification task," *Percept. Mot. Skills* 113, 929–940 (2011).
8. R. Uno, et al., "What's in a villain's name?: Sound symbolic values of voiced obstruents and bilabial consonants," *Rev. Cogn. Linguist.* 18, 428–457 (2020).
9. S. Lacey, K. L. Matthews, K. Sathian, and L. C. Nygaard, "Phonetic underpinnings of sound symbolism across multiple domains of meaning," *bioRxiv*, doi:10.1101/2024.09.03.610970 (2024).
10. J. Auracher, "Sound iconicity of abstract concepts: Place of articulation is implicitly associated with abstract concepts of size and social dominance," *PLoS One* 12, 1–25 (2017).
11. D. Maurer, T. Pathman, and C. J. Mondloch, "The shape of boubas: Sound-shape correspondences in toddlers and adults," *Dev. Sci.* 9, 316–322 (2006).
12. M. Asano, et al., "Sound symbolism scaffolds language development in preverbal infants," *Cortex* 63, 196–205 (2015).
13. D. E. Blasi, S. Wichmann, H. Hammarström, P. F. Stadler, and M. H. Christiansen, "Sound-meaning association biases evidenced across thousands of languages," *Proc. Natl. Acad. Sci. U.S.A.* 113, 10818–10823 (2016).
14. J. B. Nuckolls, "The case for sound symbolism," *Annu. Rev. Anthropol.* 28, 225–252 (1999).
15. A. Nielsen and D. Rendall, "The sound of round: Evaluating the sound-symbolic role of consonants in the classic takete-maluma phenomenon," *Can. J. Exp. Psychol.* 65, 115–124 (2011).
16. A. K. S. Nielsen and D. Rendall, "Parsing the role of consonants versus vowels in the classic Takete-Maluma phenomenon," *Can. J. Exp. Psychol.* 67, 153–163 (2013).
17. D. M. Sidhu and P. M. Pexman, "Five mechanisms of sound symbolic association," *Psychon. Bull. Rev.* 25, 1619–1643 (2018).

18. N. Johansson, A. Anikin, and N. Aseyev, "Color sound symbolism in natural languages," *Lang. Cogn.* 12, 56–83 (2020).
19. K. McCormick, J. Y. Kim, S. List, and L. C. Nygaard, "Sound to meaning mappings in the Bouba-Kiki effect," *Proc. 37th Annu. Conf. Cogn. Sci. Soc.* 1565–1570 (2015).
20. S. Lacey, K. L. Matthews, A. M. Hoffmann, K. Sathian, and L. C. Nygaard, "Differential contributions of the spectro-temporal and vocal characteristics of auditory pseudowords to multiple sound-symbolic mappings," *bioRxiv* 1–44 (2024).
21. M. Fort, A. Martin, and S. Peperkamp, "Consonants are more important than vowels in the Bouba-Kiki effect," *Lang. Speech* 58, 247–266 (2015).
22. C. Cuskley, J. Simner, and S. Kirby, "Phonological and orthographic influences in the Bouba–Kiki effect," *Psychol. Res.* 81, 119–130 (2017).
23. K. Knoeferle, J. Li, E. Maggioni, and C. Spence, "What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings," *Sci. Rep.* 7, 1–11 (2017).
24. P. A. Denes and E. N. Pinson, *The Speech Chain: The Physics and Biology of Spoken Language*, 2nd ed., vol. 17 (1993).
25. Y. Oganian, I. Bhaya-Grossman, K. Johnson, and E. F. Chang, "Vowel and formant representation in the human auditory speech cortex," *Neuron* 111, 2105–2118.e4 (2023).
26. L. Hamilton, Y. Oganian, and E. F. Chang, "Topography of speech-related acoustic and phonological feature encoding throughout the human core and parabelt auditory cortex," *bioRxiv* 2020.06.08.121624 (2020).
27. N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science* 343, 1006–1010 (2014).
28. C. Daube, R. A. A. Ince, and J. Gross, "Simple acoustic features can explain phoneme-based predictions of cortical responses to speech," *Curr. Biol.* 29, 1924–1937.e9 (2019).
29. S. Lacey, et al., "Stimulus parameters underlying sound-symbolic mapping of auditory pseudowords to visual shapes," *Cogn. Sci.* 44, e12883 (2020).
30. R. Tsur, "Size-sound symbolism revisited," *J. Pragmat.* 38, 905–924 (2006).
31. K. Akita, "Phonation types matter in sound symbolism," *Cogn. Sci.* 45 (2021).
32. B. Winter, G. Lupyan, L. K. Perry, M. Dingemanse, and M. Perlman, "Iconicity ratings for 14,000+ English words," *Behav. Res. Methods*, doi:10.3758/s13428-023-02112-6 (2023).
33. D. M. Sidhu, C. Westbury, G. Hollis, and P. M. Pexman, "Sound symbolism shapes the English language: The Maluma/Takete effect in English nouns," *Psychon. Bull. Rev.* 28, 1390–1398 (2021).
34. E. L. Thorndike, "On Orr's hypotheses concerning the front and back vowels," *Br. J. Psychol. Gen. Sect.* 36, 10–13 (1945).
35. A. N. Katz, "Meaning conveyed by vowels: Some reanalyses of word norm data," *Bull. Psychon. Soc.* 24, 15–17 (1986).

36. G. Diffloth, “i: big, a: small,” in *Sound Symbolism*, edited by L. Hinton, J. Nichols, and J. J. Ohala (Cambridge University Press, 1994), pp. 107–114.
37. G. Hollis, “Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments,” *Behav. Res. Methods* 50, 711–729 (2018).
38. E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, “Beyond the Turk: Alternative platforms for crowdsourcing behavioral research,” *J. Exp. Soc. Psychol.* 70, 153–163 (2017).
39. A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, “Gorilla in our midst: An online behavioral experiment builder,” *Behav. Res. Methods* 52, 388–407 (2020).
40. J. Dorsi, S. Lacey, K. Sathian, and L. C. Nygaard, “Acoustic stimulus parameters underlying sound symbolism in words and pseudowords” (2023).
41. P. Boersma and D. Weenink, *PRAAT: doing phonetics by computer* (2012).
42. L. Singh, “Speech signal analysis using FFT and LPC,” *Int. J. Adv. Res. Comput. Eng. Technol.* 4, 1658–1660 (2015).
43. S. J. Aiken and T. W. Picton, “Human cortical responses to the speech envelope,” *Ear Hear.* 29, 139–157 (2008).
44. J. P. Teixeira and P. O. Fernandes, “Jitter, shimmer and HNR classification within gender, tones and vowels in healthy voices,” *Procedia Technol.* 16, 1228–1237 (2014).
45. R. Kliper, S. Portuguese, and D. Weinshall, “Prosodic analysis of speech and the underlying mental state,” *Commun. Comput. Inf. Sci.* 604, 52–62 (2016).
46. M. Brockmann, M. J. Drinnan, C. Storck, and P. N. Carding, “Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task,” *J. Voice* 25, 44–53 (2011).
47. J. P. Teixeira and P. O. Fernandes, “Jitter, shimmer and HNR classification within gender, tones and vowels in healthy voices,” *Procedia Technol.* 16, 1228–1237 (2014).
48. K. R. Kluender, R. L. Diehl, and B. A. Wright, “Vowel-length differences before voiced and voiceless consonants: An auditory explanation,” *J. Phon.* 16, 153–169 (1988).
49. J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, “Acoustic characteristics of American English vowels,” *J. Acoust. Soc. Am.* 97, 3099–3111 (1995).
50. J. L. Miller and L. E. Volaitis, “Effect of speaking rate on the perceptual structure of a phonetic category,” *Percept. Psychophys.* 46, 505–512 (1989).
51. K. N., et al., “Matching categorical object representations in inferior temporal cortex of man and monkey,” *Neuron* 60(6), 126–141 (2008).
52. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification* (Wiley, 2000).
53. C. V. Parise and F. Pavani, “Evidence of sound symbolism in simple vocalizations,” *Exp. Brain Res.* 214, 373–380 (2011).
54. J. M. Fisher, F. K. Dick, D. F. Levy, and S. M. Wilson, “Neural representation of vowel formants in tonotopic auditory cortex,” *Neuroimage* 178, 574–582 (2018).

55. J. M. Hillenbrand, M. J. Clark, and R. A. Houde, "Some effects of duration on vowel recognition," *J. Acoust. Soc. Am.* 108, 3013–3022 (2000).
56. L. White, S. Benavides-Varela, and K. Mády, "Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues?," *J. Phon.* 81 (2020).
57. J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?," *Behav. Brain Sci.* 33, 61–83 (2010).
58. W. T. Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," *J. Acoust. Soc. Am.* 102, 1213–1222 (1997).
59. B. Gingras, M. Boeckle, C. T. Herbst, and W. T. Fitch, "Call acoustics reflect body size across four clades of anurans," *J. Zool.* 289, 143–150 (2013).
60. D. L. Bowling, et al., "Body size and vocalization in primates and carnivores," *Sci. Rep.* 7, 1–11 (2017).
61. A. Anikin, K. Pisanski, M. Massenet, and D. Reby, "Harsh is large: Nonlinear vocal phenomena lower voice pitch and exaggerate body size," *Proc. R. Soc. B Biol. Sci.* 288 (2021).
62. B. Winter and M. Perlman, "Size sound symbolism in the English lexicon," *Glossa* 6 (2021).
63. S. Newman, "Further experiments in phonetic symbolism," *Am. J. Psychol.* 45, 53–75 (1933).
64. C. Y. Tzeng, J. Duan, L. L. Namy, and L. C. Nygaard, "Prosody in speech as a source of referential information," *Lang. Cogn. Neurosci.* 33, 512–526 (2018).
65. T. Hubbard, "Synesthesia-like mappings of lightness, pitch, and melodic interval," *Am. J. Psychol.* 109, 219–238 (1996).
66. I. R. Titze, *Principles of Voice Production* (Prentice Hall, 1994).
67. P. Boersma and D. Weenink, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *IFA Proc.* 17, 97–110 (1993).
68. H. Reetz and A. Jongman, *Phonetics: Transcription, Production, Acoustics, and Perception* (Wiley, 2020).
69. M. Sakamoto and J. Watanabe, "Bouba/Kiki in touch: Associations between tactile perceptual qualities and Japanese phonemes," *Front. Psychol.* 9, 1–12 (2018).
70. D. M. Sidhu, G. Vigliocco, and P. M. Pexman, "Higher order factors of sound symbolism," *J. Mem. Lang.* 125, 104323 (2022).
71. L. Y. Lo, H. M. Luk, and N. Thompson, "An investigation of sound-symbolism in the context of tactile feeling," *J. Gen. Psychol.* 144, 157–167 (2017).
72. P. Ladefoged and K. Johnson, *A Course in Phonetics* (Wadsworth, 2011).
73. A. Ćwiek, et al., "The alveolar trill is perceived as jagged/rough by speakers of different languages," *J. Acoust. Soc. Am.* 156, 3468–3479 (2024).
74. B. Winter, M. Sóskuthy, M. Perlman, and M. Dingemanse, "Trilled /r/ is associated with roughness, linking sound and touch across spoken languages," *Sci. Rep.* 12, 1–11 (2022).

75. J. J. Ohala, “The frequency code underlies the sound-symbolic use of voice pitch,” in *Sound Symbolism*, edited by L. Hinton, J. Nichols, and J. J. Ohala (Cambridge University Press, 1994), pp. 325–347.
76. A. K. Nielsen and D. Rendall, “The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete–Maluma phenomenon,” *Can. J. Exp. Psychol.* 65, 115–124 (2011).
77. M. Van Puyvelde, X. Neyt, F. McGlone, and N. Pattyn, “Voice stress analysis: A new framework for voice and effort in human performance,” *Front. Psychol.* 9, 1–25 (2018).
78. M. Brockmann-Bauser, “Acoustic perturbation measures improve with increasing vocal intensity in individuals with and without voice disorders,” *J. Voice* 32, 162 (2018).
79. A. B. Warriner, V. Kuperman, and M. Brysbaert, “Norms of valence, arousal, and dominance for 13,915 English lemmas
80. M. Kappen, et al., “Acoustic speech features in social comparison: how stress impacts the way you sound,” *Sci. Rep.* 12, 1–10 (2022).
81. C. T. Stilp and B. McMurray, “Spectral tilt versus formant frequency in static and dynamic vowels,” *J. Acoust. Soc. Am.* 135, 2666–2680 (2014).
82. S. Rosen, “Temporal information in speech: Acoustic, auditory and linguistic aspects,” *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 336, 367–379 (1992).
83. K. V. Nourski, “Auditory processing in the human cortex: An intracranial electrophysiology perspective,” *Laryngoscope Investig. Otolaryngol.* 2, 147–156 (2017).
84. B. Khalighinejad, et al., “Functional characterization of human Heschl’s gyrus in response to natural speech,” *Neuroimage* 235 (2021).
85. J. I. Hualde, C. D. Eager, and M. Nadeu, “Catalan voiced prepalatals: Effects of nonphonetic factors on phonetic variation?,” *J. Int. Phon. Assoc.* 45, 243–267 (2015).
86. T. H. Crystal and A. S. House, “The duration of American-English vowels: an overview,” *J. Phon.* 16, 263–286 (1988).
87. B. Winter, M. Perlman, L. K. Perry, and G. Lupyan, “Which words are most iconic?,” *Interact. Stud.* 18, 443–464 (2017).
88. P. D. Eimas and P. C. Quinn, “Studies on the formation of perceptually based basic-level categories in young infants,” *Child Dev.* 65, 903–917 (1994).
89. W. Deng and V. M. Sloutsky, “Inference training on category representation,” *Cogn. Psychol.* 51, 392–405 (2015).
90. K. Kantartzis, M. Imai, and S. Kita, “Japanese sound-symbolism facilitates word learning in English-speaking children,” *Cogn. Sci.* 35, 575–586 (2011).
91. M. Imai and S. Kita, “The sound symbolism bootstrapping hypothesis for language acquisition and language evolution,” *Philos. Trans. R. Soc. B Biol. Sci.* 369 (2014).
92. M. Imai, S. Kita, M. Nagumo, and H. Okada, “Sound symbolism facilitates early verb learning,” *Cognition* 109, 54–65 (2008).
93. H. Yoshida, “A cross-linguistic study of sound symbolism in children’s verb learning,” *J. Cogn. Dev.* 13, 232–265 (2012).

94. S. Aussems, L. Devey Smith, and S. Kita, “Do 14–17-month-old infants use iconic speech and gesture cues to interpret word meanings?,” *J. Acoust. Soc. Am.* 156, 638–654 (2024).

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

96. J. Dorsi, C. Sandberg, S. Lacey, L. C. Nygaard, and K. Sathian, “Sound symbolism for concrete, but not abstract, semantic domains is preserved in people with aphasia” (2024).

97. R. R. Klink, “Creating meaningful new brand names: A study of semantics and sound symbolism,” *J. Mark. Theory Pract.* 9, 27–34 (2001).

98. R. R. Klink, “Creating brand names with meaning: The use of sound symbolism,” *Mark. Lett.* 11, 5–20 (2000).

99. E. Yorkston and G. Menon, “A sound idea: Phonetic effects of brand names on consumer judgments,” *J. Consum. Res.* 31, 43–51 (2004).