Between Bouba and Kiki lies Grougi: Modelling the Rough Middle Ground of Sound–Shape Correspondences

Asterios Zacharakis[1] [0000−0003−2024−6924],
Konstantinos Velenis[1] [0009−0006−7299−4763], and
Emilios Cambouropoulos[1] [0000−0002−6887−4287]

Aristotle University, Thessaloniki, Greece
aszachar@mus.auth.gr

Abstract. Cross-modal correspondences between audition and vision have been extensively studied in relation to pitch and loudness, but timbre remains more challenging to capture empirically and model acoustically due to its multidimensional nature. Nevertheless, recent research has shown that timbral qualities such as brightness, roughness, and clarity can be reflected in the visual properties of images. Building on the well-known Bouba/Kiki effect, the present study extends this framework to musical timbres and investigates visual forms that span a continuum ranging from full roundness to high angularity. We hypothesised not only that visually induced roughness would occupy the perceptual middle ground between these two extremes, but also that this categorisation would be meaningfully reflected in auditory perception. To test this, 27 participants were asked to associate 30 short sound stimuli with visual shapes using a morphing interface that enabled continuous transformation from a circle to a jagged, starlike form. The results revealed strong, non-random associations between sounds and the three shape categories (smooth, rough, and sharp). Acoustic modelling further supported these distinctions, showing that established models of auditory roughness and sharpness effectively predicted participants' assignments of the sound stimuli within these categories. Together, these findings offer empirical backing for perceptually grounded mappings between sound and image, with potential applications in artistic creation, education, and interactive media.

Keywords: timbre · roughness · sharpness · cross-modal correspondences · shapes.

All rights remain with the authors under the Creative Commons Attribution 4.0 International License (CC BY 4.0).
Proc. of the 17th Int. Symposium on Computer Music Multidisciplinary Research, London, United Kingdom, 2025

1 Introduction

Musical sounds —even when experienced in isolation from their original context— can convey extra-musical meaning [15,8], often by evoking qualities that extend beyond audition in the strict sense [15]. Growing interest in timbre semantics over the past fifteen years has advanced our understanding of how such qualities are cognitively organised and what acoustic features underlie them (for an overview of the domain until 2019 see [33,25]). Some of the most notable contributions have organised timbre descriptions into semantic categories, many of which involve cross-modal loans. As early as 1941, Lichte identified brightness and roughness as important dimensions in the description of timbre —qualities that draw from the visual and tactile domains, respectively [12]. Since then, the prominence of brightness and roughness dimensions (and their variations) within timbral semantics has been reinforced by numerous studies across diverse linguistic backgrounds [5,4,27,2,39,23,22]. At the same time, it should not be overlooked that research has also revealed nuanced variations within these primary dimensions, as well additional semantic categories altogether such as, for example, *full*, *woody*, *brassy/metallic*, *hollow*, *open*, *percussive*, etc. [37,20,19]. Given that a considerable portion of timbral semantics originates from cross-modal correspondences [30], it has been argued that systematically exploring such associations offers a fruitful path for future research on extra-musical meaning conveyed by sound [25]. Indeed, a growing body of recent work has focused on potential links between timbral and visual or tactile qualities [24,32,31,34,11,21] and even between timbral and olfactory qualities [35,36].

While brightness is arguably the most well-established semantic dimension of timbre, the tactile-related quality of roughness has drawn scholarly attention since the time of Helmholtz [7], and consistently emerges as the most widely agreed-upon attribute across linguistic groups and levels of musical expertise [39,37,32,38,22]. Roughness constitutes one of the principal qualities of surfaces and is primarily assessed through touch. That said, certain aspects of roughness can also be evaluated visually [6]. With this in mind, we recently explored potential relationships between various graphical manifestations of roughness (i.e., roughness, noisiness, granularity, and sharpness) and sound stimuli specifically designed to convey these qualities in the auditory domain [40] using a five-alternative forced-choice (5AFC) experimental design. Our data showed that the audio-image correspondences were above chance overall. At the same time, the rough and granular visual categories could be acoustically differentiated using an audio feature originating from the Modulation Power Spectrum (MPS), known to predict auditory roughness [3,22] and Zwicker's sharpness model [41].

Emboldened by these preliminary results, we hereby proceed to a simplified variation of this experiment that is partly inspired by the Bouba/Kiki effect [9,17,16], which has demonstrated the systematic mapping between words and shapes. A bit more than a decade ago, Adeli and colleagues [1] also expanded the Bouba/Kiki paradigm from word-shape to sound-shape associations. Their work demonstrated that softer timbres tended to be associated with rounder, Bouba-style forms, while harsher ones were linked to spikier, Kiki-style shapes.

However, this work identified associations between sound and shapes primarily at the source-cause level (e.g., piano, saxophone, gong) rather than with specific acoustic correlates. In the current experiment, participants were given the option to adjust the properties of a shape rather than choosing from a predefined set of images. Thus, using a single slider, it was possible to morph a full circle into a spikier shape composed of 20 trapezoids, which could gradually evolve into triangles by reducing the length of their shorter side (see Figure 1). The underlying assumption was that this continuous visual transformation would progress from smooth to sharp, passing through an intermediate stage perceived as rough —much like how the lexical blend *Grougi*, positioned between *Kiki* and *Bouba*, evokes a rough quality. This design provided a neat and efficient way to collect data on all three perceptual qualities (smoothness, roughness and sharpness) using a single, intuitive interface, described in greater detail in the Method section.

These investigations into audio-visual correspondences are part of a broader research project entitled Soundsketcher [29], which aims to automatically translate audio into perceptually meaningful visual representations for the creation of aural scores in both creative and educational contexts.

2 Method

2.1 Sound stimuli

The experimental dataset consisted of 30 short audio stimuli (1.5–3 seconds), drawn from various sound libraries (e.g., Freesound, MUMS [14]), adapted from past studies [13], or designed from scratch to equally represent the three sonic qualities of interest (smooth, rough, and sharp; 10 stimuli per category). To construct this set, the research team, together with undergraduate students from a music cognition module, compiled a broader pool of candidate sounds, which were then discussed and evaluated through informal listening across the three semantic categories, ultimately yielding the final 30 stimuli. The selected stimuli were subsequently loudness equalised through an iterative listening process conducted within the research team.

2.2 Visual variable

The visual response element was a shape-morphing interface presented in the form of a continuous slider (Figure 1). The base shape was a perfect circle. As participants moved the slider to the right, the shape gradually morphed into increasingly jagged, star-like forms with higher angular modulation. This transformation was continuous and allowed for fine-grained adjustments across the full range of possible shapes. The slider corresponded to a hidden scale ranging from 0 to 20 that was not revealed to the participants. This morphing continuum was designed to provide participants with a straightforward and intuitive way to explore the transition from round to angular shapes. The basic assumption here was that the midpoint between complete roundness (i.e., smoothness) and full angularity (i.e., sharpness) would signify roughness.

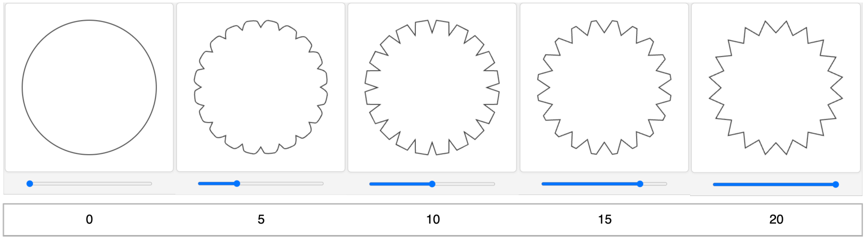


Fig. 1. Visual shape continuum reflecting textural morphing from smooth to rough and finally sharp. Participants adjusted a slider to select the shape that best represented the sound they heard. The shape gradually transformed from a smooth circle (left), to rough undulating contours (center), and finally to sharp, pointed spikes (right). The numbers shown beneath each image represent the corresponding slider value, noting that this was not visible to participants.

2.3 Participants

A total of 27 participants took part in the study. The participant pool was diverse in terms of gender and musical background. Most participants identified as male (19), followed by female (8), with one participant choosing not to disclose their gender.

Musical experience levels varied considerably: the sample included 9 professional musicians, 5 semi-professionals, 7 amateurs and 6 non-musicians. The majority of participants reported playing a musical instrument, with instrument types including piano, guitar, and electric bass, to more specialised ones such as the double bass.

Regarding listening equipment, most participants used closed-back (over-ear) headphones, while others used on-ear headphones or studio monitors. All participants reported normal hearing and normal or corrected-to-normal vision. Participation was voluntary and no compensation was provided.

2.4 Experimental procedure

The experiment was conducted online via a browser-based interactive interface. In each trial, participants were presented with an initial circle shape representing the smoothness and a slider beneath it at value 0. Their task was to adjust the shape of the image by moving the slider until they felt it best represented the quality of each sound stimulus. They could trigger audio playback by clicking a button labeled “Press "A" or click to play the sound”. At the beginning of the session, participants were presented with three practice sounds to familiarize themselves with the procedure and the interface. These practice trials were excluded from the data analysis. In the main experiment, the presentation order of the 30 audio stimuli was randomised for each participant to counterbalance potential order effects. There were no restrictions on how many times participants could listen to a sound before making their judgment. Once they were satisfied with their choice, they clicked the "Next Sound" button to proceed. This process was repeated for all 30 randomised trials following the initial practice phase.

Throughout the experiment, the system recorded the slider position (which corresponded to a specific shape configuration), the associated sound ID, and the timestamp of the selection. The full set of audio stimuli, along with the shape generation code used in the experiment, is publicly available on GitHub.

3 Results

3.1 Analysis of responses

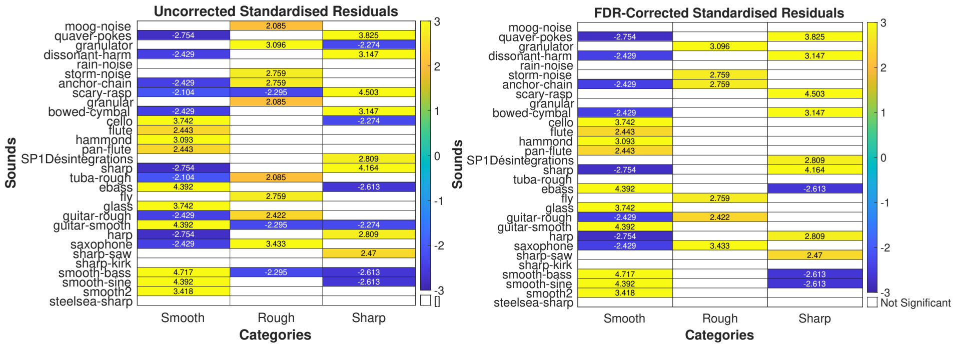
The primary goal of the present analysis was to determine whether the behavioural responses exhibited systematic trends. To this end, the original continuous variable (ranging from 0 to 20) was transformed into a categorical variable by dividing the scale into three regions: [0–5] for smooth, [6–15] for rough, and [16–20] for sharp. This categorisation was informed by both visual inspection of the stimuli's shape transformations (see Figure 1) and by the goal of optimising Cramér's V and the fit of the acoustic models (see below).

Subsequently, a Chi-square on the contingency table of responses (Table 1) test revealed a significant deviation from uniformity in category choices across sound stimuli, $\chi^2(58) = 478.89$, $p < 0.00001$, indicating a systematic, non-random pattern in how sounds were attributed to the three categories. The effect size, measured by Cramér's V ($V = 0.544$), reflects a relatively strong structure in these category assignments [18].

Table 1. Contingency table showing participant selections (counts) by semantic category for each stimulus.

Stimulus	Smooth	Rough	Sharp	Stimulus	Smooth	Rough	Sharp
moog-noise	4	15	8	sharp	1	5	21
quaver-pokes	1	6	20	tuba-rough	3	15	9
granulator	7	18	2	ebass	23	3	1
dissonant-harm	2	7	18	fly	4	17	6
rain-noise	6	10	10	glass	21	3	3
storm-noise	4	17	6	guitar-rough	2	16	9
anchor-chain	2	17	8	guitar-smooth	23	2	2
scary-rasp	3	2	22	harp	1	9	17
granular	6	15	6	saxophone	2	19	6
bowed-cymbal	2	7	18	sharp-saw	6	5	16
cello	21	4	2	sharp-kirk	8	8	11
flute	17	6	4	smooth-bass	24	2	1
hammond	19	5	3	smooth-sine	23	3	1
pan-flute	17	7	3	smooth2	20	3	4
SP1Désintégrations	5	5	17	steelsea-sharp	7	13	7

At the individual level, Figure 2 illustrates the strength of associations between each sound and the three response categories. Subfigure 2a presents the significant uncorrected standardised residuals from the Chi-square test of independence, while Subfigure 2b shows the significant standardised residuals after False Discovery Rate (Benjamini-Hochberg) correction, which enhances sensitivity to meaningful relationships while controlling false discoveries. The uncorrected coefficients reveal 27 out of 30 significant associations, while even the stricter corrected coefficients still indicate significant associations for 24 out of 30 sounds with one of the visual categories.



(a) (b)

Fig. 2. Heatmap of the standardised residuals from the Chi-square test of independence, with blue indicating negative values and yellow representing positive values, where non-significant values are shown in white. (a) Significant matchings for the uncorrected standardised residuals. (b) Significant matchings after False Discovery Rate (Benjamini-Hochberg) correction.

3.2 Auditory modelling

After establishing a strong systematic relationship between sounds and visual categories, we sought to interpret these relationships through acoustic predictors. To this end, we computed conditional probabilities reflecting the likelihood of each visual category being chosen given a specific sound stimulus, P(Visual | Auditory). Responses were aggregated into a 30×3 contingency table, capturing the distribution of selections across conditions. A multinomial regression was then performed on this contingency structure to assess the extent to which acoustic predictors systematically influenced visual choices.

The modelling explored some audio features that have been linked with auditory roughness and sharpness. The roughness models by Vassilakis [28] and Sethares [26] were implemented using the MIR Toolbox [10] (window size: 40ms, 50% overlap). Additionally, the roughness and sharpness models based on the work by Fastl & Zwicker [41] were computed using the MATLAB functions acousticRoughness and acousticSharpness. Furthermore, a metric based on the

Modulation Power Spectrum (MPS), known to predict auditory roughness [3,22], was also calculated. The MPS is the 2D-Fourier transform of a spectrogram that quantifies the distribution of modulation energy across the temporal and spectral dimensions. In this study, the 2D Fourier transform was applied to a loudness spectrogram estimated using MATLAB's acousticLoudness function (window size: 50 ms, hop size: 2ms), which implements Zwicker's method for loudness estimation (ISO 532-1) [41], rather than a spectrogram derived from a standard Fourier transform. MPS-roughness was calculated as the ratio of energy within the 30–150 Hz range of the temporal modulation axis to the total positive modulation energy, using a slight variation of the metric proposed by Rosi et al. [22].

A multinomial regression model was then employed to predict visual category probabilities through the audio features. A log transformation was applied to the predictor table X, $X_{\log} = \log_{10}(X + \varepsilon), \quad \text{with } \varepsilon = 10^{-4}$ to suppress positive outliers in the audio features and improve the stability and interpretability of regression coefficients. Standardisation (z-scoring) was subsequently performed to ensure comparability across features. Model evaluation relied on log-likelihood-based metrics, including McFadden's R^2 and Nagelkerke's R^2, to assess explanatory power. Rather than relying on discrete classifications, predictions were derived from estimated probability distributions over the visual categories, providing a more continuous interpretation of the relationship between predictors and response tendencies.

Table 2. Multinomial regression results predicting the selection of the shape categories smooth and rough (vs. sharp) based on acoustic features. The sign and magnitude of the beta coefficients indicate how the probability of selecting a given category is influenced by each feature, accompanied by 95% confidence intervals and significance levels. Model fit: Log-Likelihood ($\mathcal{L}$) = -12.77; McFadden's R^2 = 0.61, Nagelkerke's R^2 = 0.83; Prediction Accuracy = 100%; Log Loss = 11.18. These metrics reflect the model's goodness-of-fit, classification performance, and uncertainty in predictions. Likelihood Ratio Test (LRT) p-values: Vassilakis' roughness = 0.0001, Zwicker's sharpness = 0.0000, indicating that both predictors are statistically significant.

Predictors	**Betas**	**p-Value**	**95% CI**
Cat. 1: [Smooth]			
Intercept	-337.7	**<.0001**	[-343.05 , -332.38]
Vassilakis' roughness	-845.3	**<.0001**	[-849.5, -841.09]
Zwicker's sharpness	-2608	**<.0001**	[-2616.3, -2599.6]
Cat. 2: [Rough]			
Intercept	775.7	**<.0001**	[770.67 , 780.76]
Vassilakis' roughness	189.6	**<.0001**	[185.66 , 193.6]
Zwicker's sharpness	-1120.5	**<.0001**	[-1127 , -1114]

Several pairwise combinations of the five examined descriptors resulted in significant and comparable models, all of which featured one of the roughness descriptors as the first predictor and *Zwicker's sharpness* as the second. The model reported here (Table 2) yielded the highest McFadden's R^2 (.612) and features Vassilakis roughness as the first predictor and *Zwicker's sharpness* as the second. This model also demonstrated meaningful interpretability, as all of the betas were statistically significant. The beta coefficients presented in Table 2 reflect how the two auditory predictors influence the likelihood of a shape falling into the first two categories (i.e., smooth and rough), relative to the reference category (sharp).

According to the model, the likelihood of a stimulus falling into the smooth category over the sharp category increases substantially as Vassilakis' roughness decreases, and even more so as Zwicker's sharpness decreases. In contrast, while higher Vassilakis' roughness increases the likelihood of the rough category relative to sharp, lower Zwicker's sharpness continues to increase the likelihood of rough over sharp.

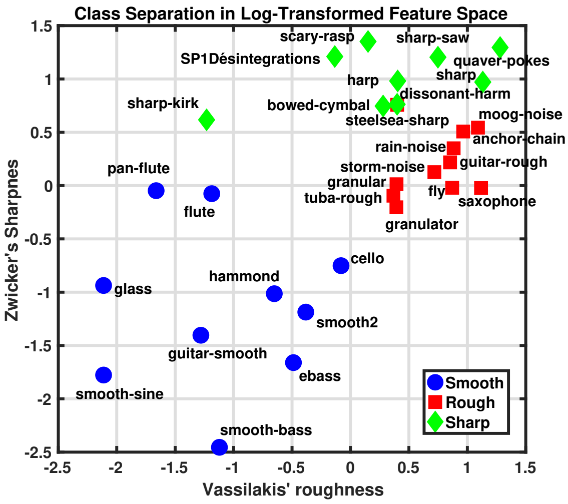


Fig. 3. Separation of visual categories in the feature space defined by Vassilakis' roughness and Zwicker's sharpness (log-transformed). Each point represents a stimulus, colored and shaped according to the most probable visual category (smooth, rough, or sharp) based on participant responses. The clear separation between regions illustrates the strong contribution of these auditory features to the categorisation of shape responses.

While the magnitude of the model's beta coefficients and the perfect classification accuracy might raise concerns about potential overfitting, the two-dimensional feature space defined by the log-transformed Vassilakis' roughness and Zwicker's sharpness (Figure 3) shows a clear separation between the visual categories. The fact that stimuli associated with each category —based on the most frequent visual assignment by participants— occupy distinct regions in the feature space supports the conclusion that these auditory features play a meaningful role in shaping perceptual categorisation. Nonetheless, the robustness of this model should ideally be confirmed through independent validation.

Across multiple 5-fold cross-validation runs, the multinomial regression model trained on auditory features achieved mean classification accuracies ranging from 76.7% to 86.7%, with an overall average of approximately 80.0% (±11.4%). The corresponding mean log loss values averaged around 9.6 (±1.9), indicating relatively low predictive confidence despite high classification accuracy. This suggests that the auditory features offer meaningful discriminatory power for modelling sound–shape correspondences, although the elevated log loss points to limited certainty in the probability estimates, likely a consequence of the small dataset size.

4 Discussion

These preliminary findings offer promising insights into cross-modal correspondences between vision and audition. In particular, our hypothesis of a visual continuum linking smoothness to sharpness via roughness appears to be supported by the categorisation and analysis of our visual variable. As mentioned in the Results section, the discretisation of the continuous visual variable into three distinct regions (*smooth*, *rough*, and *sharp*) was guided by a combination of visual inspection of the generated shapes and iterative refinement based on optimisation of the Cramér's V values (here quantifying the strength of association between visual categories and the sound stimuli) and goodness-of-fit measures from the subsequent acoustic modelling. A Chi-square test of independence identified audio-visual associations, which exceeded chance levels both overall and across individual stimuli. Indeed, the distribution of our 30 audio stimuli across the three visual categories was approximately even (see Figure 2) and broadly aligned with our initial assumptions on perceived smoothness, roughness and sharpness. Thus, our results extend the well-known Bouba/Kiki effect from speech-based sounds to musical timbres, in agreement with previous research [1], while also revealing an intermediate perceptual category linked to roughness.

Based on this finding, we subsequently sought to achieve an acoustic modelling of the visual categories employing a number of established models of auditory roughness and sharpness. The modelling further strengthened our assumptions on the distinctions achieved between smoothness, roughness and sharpness. Sounds with higher roughness values, as measured by Vassilakis' model [28], were more likely to be assigned to the *rough* category and less likely to be assigned to *smooth*, relative to *sharp*. Additionally, higher values of Zwicker's sharpness [41] reduced the likelihood of assignment to both *smooth* and *rough*, again relative to *sharp*. This pattern is clearly illustrated in the scatterplot of Figure 3, where stimuli with higher probabilities of being rated as sharp (green diamonds) tend to cluster in the upper-right region of the plot, characterised by both high Zwicker's sharpness and high Vassilakis' roughness. Stimuli with a higher likelihood of being categorised as rough (red squares) appear just below and further to the right, defined by lower acoustic sharpness but even higher acoustic roughness values on average. Finally, stimuli most likely to be rated as smooth (blue circles) occupy the bottom-left region, marked by both low acoustic roughness and low sharpness. Note that each of the three categories contains the 10 stimuli we initially assumed to represent them, with the sole exception of *steelsea-sharp*. Although acoustically it belonged to the sharpness group (see Fig. 3), which was also our assumption, participants rated it with a higher probability of being rough.

The multinomial regression used for acoustic modelling produced a markedly stronger outcome compared to a similar model in our recent work [40]. This improvement is expected, given that here we limited the visual categories to three (rather than five previously) and increased the number of audio stimuli to 30 (compared to 18 before). These changes already create more favourable conditions for modelling. Moreover, the improvement may also stem from a more careful selection of audio stimuli that better represent the qualities under examination, as well as from the fact that the continuous visual morphing paradigm constitutes a more efficient method for acquiring visual information.

Although the multinomial regression model incorporating Vassilakis' roughness and Zwicker's sharpness is excellent at predicting the most likely response category, its confidence remains limited, suggesting low generalisability of the results. Future research should aim to include larger audio stimulus sets, acquire larger scale empirical data and potentially explore additional acoustic factors to deepen our understanding of sound-shape relationships within the smoothness–roughness-sharpness continuum. Approaching the problem from the opposite perspective, participants could be provided with a sound synthesis engine and asked to create timbres that match a variety of distinct shapes. This would allow us to directly map sound synthesis parameters onto visual categories.

Meanwhile, within the framework of the Soundsketcher project [29], these findings already provide valuable insights that contribute to the development of perceptually meaningful mappings between sounds and images. By grounding these mappings in empirically supported cross-modal correspondences, this work helps pave the way for more intuitive audio-visual interactions, enhancing both creative exploration and user experience.

Acknowledgments. We would like to thank the undergraduate student assistants who contributed to the creation and selection of sound stimuli, as well as to participant recruitment. We would also like to thank Anastasia Vakatari for her assistance with data acquisition and preprocessing. This research was carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union – NextGenerationEU (Implementation body: HFRI).

References

1. Adeli, M., Rouat, J., Molotchnikoff, S.: Audiovisual correspondence between musical timbre and visual shapes. Frontiers in human neuroscience **8**, 352 (2014)
2. Alluri, V., Toiviainen, P.: Effect of enculturation on the semantic and acoustic correlates of polyphonic timbre. Music Perception **29**(3), 297–310 (2011)

3. Arnal, L.H., Flinker, A., Kleinschmidt, A., Giraud, A.L., Poeppel, D.: Human screams occupy a privileged niche in the communication soundscape. Current biology **25**(15), 2051–2056 (2015)
4. von Bismarck, G.: Sharpness as an attribute of the timbre of steady sounds. Acustica **30**, 17–24 (1974)
5. von Bismarck, G.: Timbre of steady tones: A factorial investigation of its verbal attributes. Acustica **30**, 146–159 (1974)
6. Di Stefano, N., Spence, C.: Roughness perception: A multisensory/crossmodal perspective. Attention, Perception, & Psychophysics **84**(7), 2087–2114 (2022)
7. von Helmholtz, H.L.F.: On the Sensations of Tone as a Physiological Basis for the Theory of Music. New York: Dover (1954), 4 edn. (1877), english translation by A. J. Ellis
8. Koelsch, S.: Towards a neural basis of processing musical semantics. Physics of Life Reviews **8**(2), 89–105 (2011)
9. Köhler, W.: Gestalt psychology. New York, NY: Liveright (1929)
10. Lartillot, O., Toiviainen, P.: A Matlab toolbox for musical feature extraction from audio. In: International conference on digital audio effects. vol. 237, p. 244. Bordeaux (2007)
11. Lembke, S.A.: Hearing triangles: Perceptual clarity, opacity, and symmetry of spectrotemporal sound shapes. The Journal of the Acoustical Society of America **144**(2), 608–619 (2018)
12. Lichte, W.: Attributes of complex tones. J. Exp. Psych **28**, 455—-480 (1941)
13. Noble, J., Thoret, E., Henry, M., McAdams, S.: Semantic dimensions of sound mass music: mappings between perceptual and acoustic domains. Music Perception **38**(2), 214–242 (2020)
14. Opolko, F., Wapnick, J.: McGill University master samples collection on DVD (2006)
15. Patel, A.D.: Music, language, and the brain., chap. Meaning, pp. 299–254. Oxford University Press (2008)
16. Ramachandran, V.S., Hubbard, E.M.: Hearing colors, tasting shapes. Scientific American **288**(5), 52–59 (2003)
17. Ramachandran, V.S., Marcus, Z., Chunharas, C.: Bouba-kiki: Cross-domain resonance and the origins of synesthesia, metaphor, and words in the human mind. In: Multisensory perception, pp. 3–40. Elsevier (2020)
18. Rea, L.M., Parker, R.A.: Designing and conducting survey research: A comprehensive guide. John Wiley & Sons (2014)
19. Reymore, L.: Characterizing prototypical musical instrument timbres with timbre trait profiles. Musicae Scientiae **26**(3), 648–674 (2022)
20. Reymore, L., Huron, D.: Using auditory imagery tasks to map the cognitive linguistic dimensions of musical instrument timbre qualia. Psychomusicology: Music, Mind, and Brain **30**(3), 124 (2020)
21. Reymore, L., Lindsey, D.T.: Color and tone color: audiovisual crossmodal correspondences with musical instrument timbre. Frontiers in Psychology **15** (2025). `https://doi.org/10.3389/fpsyg.2024.1520131`
22. Rosi, V., Arias Sarah, P., Houix, O., Misdariis, N., Susini, P.: Shared mental representations underlie metaphorical sound concepts. Scientific Reports **13**(1), 5180 (2023)
23. Rosi, V., Houix, O., Misdariis, N., Susini, P.: Investigating the shared meaning of metaphorical sound attributes: bright, warm, round, and rough. Music Perception **39**(5), 468–483 (2022)

24. Saitis, C., Wallmark, Z.: Timbral brightness perception investigated through multimodal interference. Attention, Perception, & Psychophysics **86**(6), 1835–1845 (2024)
25. Saitis, C., Weinzierl, S.: The semantics of timbre. In: Siedenburg, K., Saitis, C., McAdams, S., Popper, A.N., Fay, R.R. (eds.) Timbre: Acoustics, Perception, and Cognition, pp. 119–149. Springer (2019)
26. Sethares, W.A.: Tuning, timbre, spectrum, scale. Springer Science & Business Media (2005)
27. Stepánek, J.: Musical sound timbre: Verbal description and dimensions. In: Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06). pp. 121–126. Citeseer (2006)
28. Vassilakis, P.N.: Perceptual and physical properties of amplitude fluctuation and their musical significance. Ph.D. thesis, University of California, Los Angeles (2001)
29. Velenis, K., Zacharakis, A., Cambouropoulos, E.: Soundsketcher: Graphical representation of sound based on crossmodal audiovisual correspondences. In: Proceedings of the 22nd Sound and Music Computing Conference (SMC2025). University of Music and Performing Arts, Graz, Austria (2025)
30. Wallmark, Z.: A corpus analysis of timbre semantics in orchestration treatises. Psychology of Music **47**(4), 585–605 (2019)
31. Wallmark, Z.: Semantic crosstalk in timbre perception. Music & Science **2**, 2059204319846617 (2019)
32. Wallmark, Z., Allen, S.E.: Preschoolers' crossmodal mappings of timbre. Attention, Perception, & Psychophysics **82**, 2230–2236 (2020)
33. Wallmark, Z., Kendall, R.A.: Describing sound: The cognitive linguistics of timbre. In: Dolan, E.I., Rehding, A. (eds.) The Oxford Handbook of Timbre, pp. 578–608. New York, NY: Oxford University Press (2018). `https://doi.org/10.1093/oxfordhb/9780190637224.013.14`
34. Wallmark, Z., Nghiem, L., Marks, L.E.: Does timbre modulate visual perception? exploring crossmodal interactions. Music Perception **39**(1), 1–20 (2021)
35. Zacharakis, A.: Sonic bouquet: Decoding cross-modal correspondences between timbre and scent. Music & Science **7** (2024). `https://doi.org/10.1177/20592043241274258`
36. Zacharakis, A.: The poetry of senses: exploring semantic mediation in timbre-aroma correspondences. Frontiers in Psychology **16**, 1520046 (2025)
37. Zacharakis, A., Pastiadis, K.: Revisiting the luminance-texture-mass model for musical timbre semantics: A confirmatory approach and perspectives of extension. Journal of the Audio Engineering Society **64**(9), 636–645 (2016)
38. Zacharakis, A., Pastiadis, K.: Interaction between time-varying tone inharmonicity, fundamental frequency and spectral shape affects felt tension and timbral semantics. Acta Acustica **5**, 21 (2021)
39. Zacharakis, A., Pastiadis, K., Reiss, J.D.: An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. Music Perception **31**(4), 339–358 (2014)
40. Zacharakis, A., Vlaikidou, A., Velenis, K., Cambouropoulos, E.: Facets of audiovisual roughness: from granularity to sharpness. In: Proceedings of the 22nd Sound and Music Computing Conference (SMC2025). University of Music and Performing Arts, Graz, Austria (2025)
41. Zwicker, E., Fastl, H.: Psychoacoustics: Facts and models, vol. 22. Springer Science & Business Media (2013)