Research Article

Audiovisual Associations in Saint-Saëns' *Carnival of the Animals*: A Cross-Cultural Investigation on the Role of Timbre

Empirical Studies of the Arts
2025, Vol. 43(2) 1162–1180

DOI: 10.1177/02762374241308810
journals.sagepub.com/home/art


Nicola Di Stefano[1], Alessandro Ansani[2], Andrea Schiavio[3], Suvi Saarikallio[2], and Charles Spence[4]

Abstract

Several studies have investigated crossmodal associations involving audiovisual stimuli. To date, however, far fewer studies have explored the relationship between musical timbre and visual features (e.g., soft/harsh timbres with blue/red colours). To fill this gap in the literature, 249 participants were invited to judge the match between different coloured images and musical excerpts. The images depicted seven characters from Saint-Saëns' "Carnival of the Animals"; the audio stimuli consisted of the music the composer created to represent each character. To test the effect of timbre and culture, the audio stimuli were presented either in the original orchestral version or in the piano transcription, while the participants were recruited from various countries, encompassing both Western and non-Western nationalities. The results demonstrate that timbre influences crossmodal associations between musical excerpts and drawings, while these associations remain consistent across cultures, languages, and levels of musical background.

[1]Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy
[2]Centre of Excellence in Music, Mind, Body and Brain – Department of Music, Art and Culture Studies, University of Jyväskylä, Jyväskylä, Finland
[3]School of Arts and Creative Technologies, University of York, York, UK
[4]Crossmodal Research Lab, University of Oxford, Oxford, UK

Corresponding Author:
Alessandro Ansani, Centre of Excellence in Music, Mind, Body and Brain – Department of Music, Art and Culture Studies, University of Jyväskylä, Jyväskylä, Finland.
Email: alessandro.a.ansani@jyu.fi

Keywords
crossmodal correspondences, cluster analysis, music perception, musical timbre

Introduction

Crossmodal associations, also known as crossmodal correspondences, have been defined as the tendency for a sensory feature, attribute, or dimension in one sensory modality, either physically present, or merely imagined, to be matched (or associated) with a sensory feature, attribute, or dimension in another modality (Motoki et al., 2023; Spence, 2011). Unlike synaesthesia, which is, by definition, idiosyncratic in terms of the inducer-concurrent mapping (see Deroy & Spence, 2013; Grossenbacher & Lovelace, 2001), crossmodal correspondences tend to be consensual (see also Sun et al., 2018). For example, numerous studies have shown that participants consistently associate round shapes with sweetness, and angular shapes with sour-tasting foods (Deroy et al., 2013; Spence, 2023).

As suggested by Daniel Stern (1999), the early constitution of intersubjectivity and a meaningful world in infants relies on vitality contours, which are affective and amodal forms of communication and expression that are present, for instance, in a caregiver's voice and touch. Cross-modality can thus be considered inherent to human communication and emotional experience from the early stages of human life.

In the audiovisual domain, studies have documented the existence of consistent correspondences between simple auditory and visual stimuli, such as pitch and size (e.g., Evans & Treisman, 2010; Gallace & Spence, 2006; Mondloch & Maurer, 2004; Walker, 1987); pitch and timbre/textural features of sound, such as roughness (e.g., Eitan & Timmers, 2010; Hamilton-Fletcher et al., 2018; see Di Stefano & Spence, 2022, for a review on roughness); pitch and lightness/brightness (e.g., Brunel et al., 2015; Hubbard, 1996; Klapetek et al., 2012; Marks, 1974, 1987) or hue (e.g., Di Stefano & Spence, 2022; Melara, 1989, for a review); pitch/sound waveform and shape/angularity (Marks, 1987; Parise & Spence, 2012; see, also, Ortmann, 1922).

While the majority of studies have focused on isolated and simple visual/auditory features, far less research has explored the associations between complex visual and auditory stimuli. Albertazzi and colleagues (2015) demonstrated consistent audiovisual associations between highly-complex stimuli (i.e., paintings) and music excerpts from the classical repertoire for guitar (or transcriptions, e.g., Villa-Lobos, Albeniz). Associations between complex stimuli have been explained using the semantic differential technique based on perceptual and emotional features (e.g., bright and calm, respectively; see also Cowles, 1935; Miller, 2021; Spence, 2020; and Iosifyan et al., 2022).

In a recent study using a set of stimuli based on Prokofiev's symphonic fairy tale *Peter and the Wolf*, Di Stefano and colleagues (2024) demonstrated that participants made highly consensual connections between musical excerpts and black-and-white

images depicting the characters of the fairy tale. Moreover, the results showed that the associations appear to be consensually perceived across cultures and languages, namely English, Italian, Spanish, and Chinese (see also Trainor & Trehub, 1992, for a similar earlier study conducted in a sample of North American children). These findings have been explained through the emotional mediation hypothesis (Spence, 2020), which suggests that stimuli are matched across sensory modalities (especially audiovisual) based on their similar emotional meaning or profile (see the recent studies by Hashim et al., 2023, and Rosi et al., 2023; see also Rigg, 1937, for an early investigation of musical meanings).

Far fewer studies have investigated the relationship between musical timbre and visual features (e.g., colour, shape). For example, participants in an online experiment by Adeli and colleagues (2014) associated soft timbres with blue, green, or light grey rounded shapes, harsh timbres with red, yellow, or dark grey sharp angular shapes, and timbres with elements of softness and harshness along with a mixture of the two previous shapes. Similar results were obtained in a replication study by Gurman et al. (2021; see also Liu et al., 2021; Wallmark et al., 2021, on the effect of timbre in visual perception).

Several studies have demonstrated the emotional impact of timbre. For example, Hailstone and colleagues (2009) found that timbre alone influences the perception of emotions in music, independent of other acoustic, cognitive, or performance factors (see also Eerola et al., 2012). Given the role that emotions play in mediating audiovisual associations between complex stimuli (Di Stefano et al., 2024; Spence, 2020), we hypothesised that timbre also influences cross-modal associations in complex audiovisual stimuli. To test this hypothesis, 249 participants were invited to judge the match between different images and music in terms of piano and orchestral timbre. The images depicted seven of the characters from Saint Saëns' *Carnival of Animals* (1886/1922), while the audio stimuli reproduced the music the composer created to represent each character. To test the effect of timbre, the audio stimuli were presented either in the original, orchestral version, or in the piano transcription. To test the effect of cultural background, participants were recruited from different countries and with different Western and non-Western nationalities, including Chinese and Hindi. The results demonstrate that timbre influences crossmodal associations between musical excerpts and drawings, while these associations remain consistent across cultures, languages, and levels of musical expertise.

Materials and Method

Participants

A sample of 249 adult participants ($M_{age} = 30.57$ years, $SD = 10.4$, 59.8% females) was recruited using Prolific Academic (app.prolific.com). 123 participants filled in the questionnaire in English (49.4%), 48 in Italian (19.2%), 38 in Hindi (15.2%), and 40 in Chinese (16.0%).

Participants were grouped in the following two ways:

- Non-Western (NW) ($N = 146$, 58.6%, $M_{age} = 31.10$ years, $SD = 9.4$, 58.2% females) and Western (W) ($N = 103$, 41.3%, $M_{age} = 29.83$ years, $SD = 11.6$, 62.1% females). The groups didn't differ in terms of their age ($p = .344$) or gender distribution ($X^2 = 0.70$, $df = 2$, $p = .703$).
- Nationality (Nat): Chinese: $N = 54$, 21.6%, $M_{age} = 29.5$, $SD = 7.5$, 72.2% females; Indian: $N = 92$, 36.9%, $M_{age} = 32.0$, $SD = 10.3$, 50.0% females; Western: $N = 103$, 41.3%, $M_{age} = 29.8$, $SD = 11.6$, 62.1% females. Once again, the groups did not differ in terms of their age ($p = .249$); however, a Chi-Squared test revealed that they were unbalanced in their gender distributions ($X^2 = 14.82$, $df = 4$, $p = .005$).

Stimuli and Procedure

The stimuli consisted of musical excerpts and coloured images. The audio stimuli consisted of 7 musical excerpts from Saint Saëns's musical suite *Carnival of the Animals* (Saint Saëns, 1962). All of the excerpts were saved as .wav files (stereo, 16-bit, 44.1 KHz). For the orchestra version, we used the recording by Leonard Bernstein with the New York Philharmonic Orchestra (1962) and for the piano version we used the reduction by Lucien Garban (available here: https://www.youtube.com/watch?v=I0KT4iXmeUs&t=27s). The titles and durations of the orchestral and piano excerpts are as follows: Royal March of the Lion (orchestral: 84 s, piano: 83 s), Hens and Roosters (44 s, 44 s), Tortoises (90 s, 90 s), The Elephant (71 s, 77 s), Characters with Long Ears (35 s, 35 s), The Cuckoo (122 s, 131 s), and The Swan (167 s, 191 s). The average duration of the orchestral version was 83.88 s, while the piano version averaged 87.88 s.

The drawings were extracted from the Multipic dataset, a standardized set of 750 drawings with multilingual norms (Duñabeitia et al., 2018). The correctness rate of associations, as measured by the Multipic's authors in several languages (i.e., British English, Spanish, French, Dutch, Italian, and German), was very high for all the selected drawings: Lion (100%), Chicken (89.76%), Turtle (100%), Elephant (100%), Donkey (98.08%), Swan (99.51%). This guaranteed that participants could easily identify the animals depicted. Given that the same database lacks an image of the cuckoo, we selected an image whose drawing trait was similar to that used in the Multipic dataset.

During the experiment, participants listened to all of the musical excerpts in random order and were invited to indicate the extent to which the audio stimulus matched (i.e., fit) each of the coloured representations of the animals. Participants were exposed to the following instruction: "Listen to the musical excerpt and look at the drawings. To what extent do the excerpt and the drawings match one another?". The fit was measured using a 100-point slider. Participants could freely adjust the slider to any point without being able to see the corresponding numerical value. There were no time constraints for completing the procedure.

Statistical Analyses

The participants' musical expertise and their emotional usage of music were measured by means of the Musical Training and Emotions factors of the Goldsmiths Musical Sophistication Index (Gold-MSI: Müllensiefen et al., 2014). Both the Musical Training ($\alpha = .90$, 95% CI [.88, .92]; $\omega = .91$, 95% CI [.89, .92]) and the Emotion ($\alpha = .72$, 95% CI [.66, .77]; $\omega = .72$, 95% CI [.64, .78]) scales had satisfactory internal consistency.

The Role of Timbre, Cultural Group, and Musical Sophistication: Model Comparison. The statistical analyses have been carried out in the R environment (RStudio version 2023.12.1, build 402). To model the fit score, we used a Linear Mixed Model approach via the *lme4* package (Bates et al., 2015). We used a model comparison technique (Rodgers, 2010) to assess the roles of timbre, cultural group, nationality, and musical sophistication (i.e., musical training and emotion). This procedure was deemed more appropriate than mere significance testing because of the high number of observations ($N = 12{,}201$) and involved interacting levels among the variables, due to which significance levels could exhibit statistically significant but not meaningful results. On the contrary, model comparison aims to verify the extent to which the model's predictions improve when adding a given predictor or interaction term, the underlying assumption being that the more accurate the estimates (or the lower the error), the more important the predictors.

As a first step, a baseline model was constructed analysing the interaction between the musical stimuli and drawings. To take into account inter-subject variability and consistently with the repeated-measure design, the participants were modelled as random intercepts, as follows:

$$baseline\ model: fit\ = musical\ piece\ \times drawing\ + (1 \mid ID)$$

Subsequently, to test whether the interaction changed depending on the timbre (i.e., Orchestral vs. Piano), a second model with a three-way interaction was built, as follows:

$$model\ 1: fit\ = musical\ piece\ \times drawing\ \times timbre\ + (1 \mid ID)$$

Similarly, the following four models were built to test the influence of culture, nationality, musical training and music emotion:

$$model\ 2: fit\ = musical\ piece\ \times drawing\ \times cultural\ group\ + (1 \mid ID)$$

$$model\ 3: fit\ = musical\ piece\ \times drawing\ \times nationality\ + (1 \mid ID)$$

$$model\ 4: fit\ = musical\ piece\ \times drawing\ \times musical\ training\ + (1 \mid ID)$$

$$model\ 5: fit\ = musical\ piece\ \times drawing\ \times music\ emotion\ + (1 \mid ID)$$

These models were compared using information criteria, Bayes Factors (BF), and

cross-validation. In greater detail, after inspecting the Akaike and Bayesian Information Criteria (AIC and BIC) values, we directly compared the BF as a proportion of how much one model fitted the data better than the baseline model. In model comparison, Bayes Factors indicate the ratio of evidence supporting one model over the evidence supporting the alternative model. For example, a BF = 7 indicates that the evidence for one model is seven times greater than the evidence for the other model (Fife, 2022).

Lastly, out-of-sample predicted accuracy (McElreath, 2020) was analysed through a 5-fold cross-validation procedure. Each model was trained and validated across five unique balanced subsets of the dataset. We compared the models based on three performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Root Mean Squared Logarithmic Error (RMSLE). These metrics provide a detailed insight into model accuracy, error distribution, and the ability to handle outliers and proportionate errors.

The associations have been inspected through the model's Estimated Marginal Means (EMMs) via the *emmeans* package (Lenth, 2024). Given the massive number of pairwise comparisons, the significance was assessed by using Bonferroni-corrected deviation contrasts. In greater detail, the fit score assigned to a given drawing with a musical piece was compared against the average fit score assigned to all the drawings when the same musical piece was presented. In this way, a positive contrast estimate indicated that, given a musical stimulus, a specific drawing reported a score higher than the average fit score for that music.

Power Analysis. A sensitivity analysis via simulation approach ($N = 3000$) was performed using the *simr* package in R (Green & MacLeod, 2016) to assess the minimum necessary sample size to detect the highest order interaction of interest (i.e., "musical piece × drawing × timbre") with 80% power. This procedure was preferred over an a priori power analysis since it was practically impossible to hypothesise the effect size for the three-way interaction beforehand, especially due to the massive number of pairwise comparisons involved. The analysis was based on the three-way interaction because higher-order interactions have been found to require a larger amount of participants to achieve power (Heo & Leon, 2010). The results indicated that 43 participants were enough to detect the three-way interaction with power = 82.40%, 95% CI [79.90,84.71]. With 52 participants, the power was 91.20%, 95% CI [89.27,92.88].

Association Profiles: K-Means Clustering. The association profiles (i.e., the way in which the participants clustered the animals together) were inspected through a k-means Cluster Analysis. Due to the repeated-measures structure of the data, we normalised the fit scores within each participant so that all participants had equal weight in the analysis. To select the optimal number of clusters, we resorted to *NbClust* package (Charrad et al., 2014). NbClust computes 23 different indices for computing the number of clusters (e.g., KL: Krzanowski & Lai, 1988; Hartigan: Hartigan, 1975;

Silhouette: Rousseeuw, 1987; Gap statistic: Tibshirani et al., 2001) and indicates the solution with the highest number of preferences across all indices. Both Euclidean and Manhattan distances were used to compute the number of clusters, leading to the same 4-cluster solution. Manhattan distances are preferred for cluster scoring estimation due to their lower sensitivity to outliers (Kumar, 2017) and their better performance in cases of high dimensionality (Aggarwal et al., 2001).

Results

Before proceeding to the first modelling phase, participants familiar with *The Carnival of the Animals* (N = 34) were excluded from the analysis due to potential bias in their associations. This exclusion was based on their response to the question, "How familiar are you with *The Carnival of the Animals* (Saint-Saëns)?", administered at the end of the experimental procedure. Participants who answered "I can recognize it" were excluded from the analysis.

Model Comparison

Compared to the baseline model, the model which included the timbre (Model 1) exhibited a better fit to the data. In particular, the BF indicated clear evidence in favour of Model 1. Conversely, the model which included the cultural group (Model 2) reported a worse fit (see Table 1), thus indicating that the cultural group does not tangibly affect the associations. This result is corroborated by the cross-validation, wherein Model 1 showed decreased Root Mean Squared Error, Mean Absolute Error, and Root Mean Squared Logarithmic Error.

Music-Drawings Associations

In what follows, we present the details of how each musical stimulus was associated with the visual stimuli. See Figure 1 for the overall results of the

Table 1. Model comparison.

n.	Model	AIC	BIC	$BF_{(b-1)}$	RMSE	MAE	RMSLE	R^2
b	Basic	98,946	99,315		31.2	27.2	1.61	0.12
1	**+timbre**	**98,560**	**99,284**	**>100**	**31.0**	**26.8**	**1.60**	**0.14**
2	+Nat	98,471	99,550	<.001	31.3	27.1	1.61	0.13
3	+WNN	98,741	99,465	<.001	31.3	27.2	1.61	0.13
4	+MT	98,849	99,573	<.001	31.3	27.2	1.61	0.13
5	+Emo	**98,813**	**99,537**	<.001	31.3	27.3	1.62	0.13

In the BF column, the BF compares the model at hand against the baseline model. BF > 0 indicates evidence in favour of the model at hand; BF < 0 indicates evidence in favour of the baseline model. The best model is in bold.

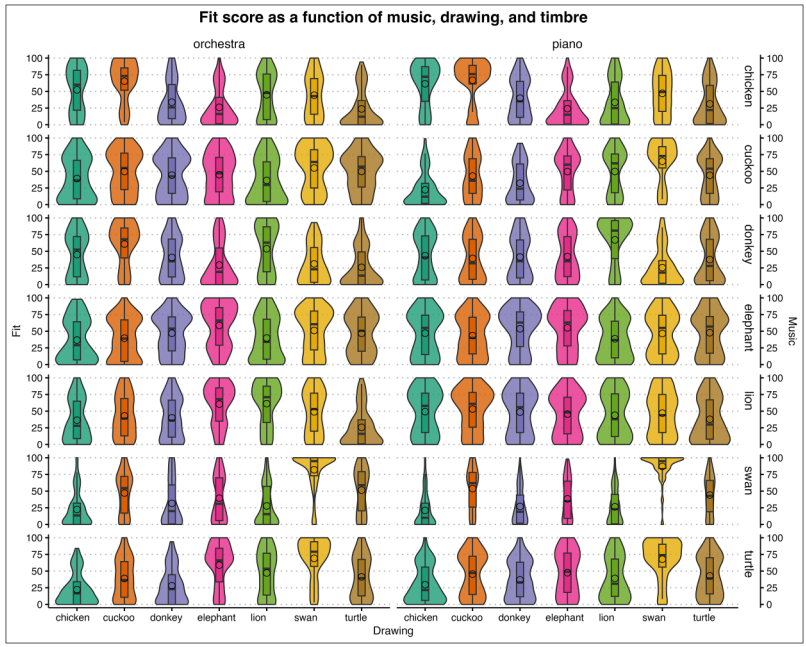

Figure 1. Fit score as a function of music, drawing, and timbre (orchestral vs piano). In the y-axis, for each Music-Drawing coupling, 100 means that participants rated the coupling as 'very good', while 0 indicates that the two stimuli 'do not match'. The form of the violin plots indicates the distribution curve. The boxplots within each violin represent interquartile ranges (IQRs). Black horizontal lines within the boxplots indicate median values. Black circles within the boxplots indicate mean values. Horizontal layers represent the different musical excerpts, while columns correspond to drawings.

associations. For the sake of brevity, we discuss just the Bonferroni-corrected significant contrasts (i.e., $p < .05$). For the complete list of the post-hoc comparisons, see the Online Supplementary Appendix.

Chicken. With the orchestral version of the chicken music, the cuckoo drawing ($Est = 20.03$, $SE = 2.89$, $p < .001$) exhibited the highest fit (compared to the mean fit), whereas Elephant ($Est = -16.96$, $SE = 2.91$, $p < .001$) and Donkey ($Est = -8.07$, $SE = 2.89$, $p = .011$) were significantly below the mean. The chicken drawing just approached statistical significance ($Est = 6.03$, $SE = 2.89$, $p = .066$).

Several differences were found in the piano version. The Cuckoo drawing's fit remained the highest ($Est = 22.24$, $SE = 2.88$, $p < .001$). However, the Chicken drawing was also significantly higher than the mean ($Est = 15.94$, $SE = 2.88$, $p < .001$). Elephant, Lion, and Turtle reported fit scores significantly below the mean ($p < .001$).

Cuckoo. In the orchestral version, the Swan drawing had the highest fit ($Est = 11.63$, $SE = 2.91$, $p < .001$), followed by the Turtle ($Est = 6.51$, $SE = 2.91$, $p = .048$). The Lion's fit score was the lowest ($Est = -7.34$, $SE = 2.91$, $p = .024$). The Swan maintained its position in the piano version ($Est = 19.95$, $SE = 2.98$, $p < .001$). Donkey ($Est = -10.94$, $SE = 2.98$, $p < .001$) and Chicken ($Est = -19.92$, $SE = 2.88$, $p < .001$) had the lowest associations.

Donkey. The Donkey music was highly associated with the Cuckoo ($Est = 18.97$, $SE = 2.92$, $p < .001$) and Lion ($Est = 8.76$, $SE = 2.92$, $p = .006$) drawings. Strong negative associations were reported with Turtle ($Est = -17.81$, $SE = 2.93$, $p < .001$), Elephant ($Est = -14.94$, $SE = 2.92$, $p < .001$), and Swan ($Est = -13.46$, $SE = 2.92$, $p < .001$). The piano version of the Donkey music was mostly associated with the Lion ($Est = 21.53$, $SE = 2.88$, $p < .001$), whereas Swan ($Est = -18.33$, $SE = 2.88$, $p < .001$) and Cuckoo ($Est = -7.02$, $SE = 2.88$, $p = .031$) had negative associations.

Elephant. In the orchestral version, the Elephant drawing reported the highest score ($Est = 11.75$, $SE = 2.91$, $p < .001$) followed by the Swan ($Est = 10.22$, $SE = 2.89$, $p = .001$). The Elephant drawing had the highest fit in the piano version too ($Est = 9.97$, $SE = 2.88$, $p = .001$), very close to the Donkey ($Est = 8.52$, $SE = 2.88$, $p = .007$).

Lion. Lion ($Est = 16.15$, $SE = 2.91$, $p < .001$) and Elephant ($Est = 16.64$, $SE = 2.91$, $p < .001$) drawings were highly matched with the Lion music. Turtle ($Est = -18.45$, $SE = 2.91$, $p < .001$) and Chicken ($Est = -6.80$, $SE = 2.89$, $p = .038$) had the lowest fit scores. This pattern was very different in the piano version, where the Cuckoo drawing was the only one highly associated with the Lion music ($Est = 9.70$, $SE = 2.88$, $p = .002$).

Swan. The Swan drawing had the highest fit score ($Est = 37.28$, $SE = 2.89$, $p < .001$); whereas Chicken ($Est = -20.20$, $SE = 2.89$, $p < .001$), Lion ($Est = -15.64$, $SE = 2.89$, $p < .001$), and Donkey ($Est = -10.40$, $SE = 2.89$, $p < .001$) had the lowest ratings. The pattern was consistent with the piano version, where the Swan drawing reported the highest fit score ($Est = 42.06$, $SE = 2.88$, $p < .001$). Once again, Chicken ($Est = -22.24$, $SE = 2.88$, $p < .001$), Lion ($Est = -18.62$, $SE = 2.88$, $p < .001$), and Donkey ($Est = -16.55$, $SE = 2.88$, $p < .001$) had negative associations. In contrast to the orchestral version, in the piano version, the Cuckoo drawing had a positive fit ($Est = 8.71$, $SE = 2.88$, $p = .006$); while the Elephant score was negative ($Est = -6.58$, $SE = 2.88$, $p = .044$).

Turtle. In the orchestral version, the Turtle music was highly associated with the Swan ($Est = 27.72$, $SE = 2.89$, $p < .001$) and Elephant drawings ($Est = 16.08$, $SE = 2.89$, $p < .001$). By contrast, the fit scores related to the Chicken ($Est = -22.73$, $SE = 2.92$, $p < .001$) and Donkey ($Est = -17.06$, $SE = 2.91$, $p < .001$) drawings were significantly lower than the mean. Some of these findings remained consistent in the piano version.

The Swan drawing was again the one with the highest fit score ($Est = 23.54$, $SE = 2.89$, $p < .001$), and the Chicken ($Est = -13.45$, $SE = 2.89$, $p < .001$) and Donkey drawings ($Est = -6.87$, $SE = 2.89$, $p = .036$) reported the lowest scores.

Association Profiles

Participants' responses were further investigated by running a k-means cluster analysis to identify consistent patterns in how they rated the fit of the animals. For example, we looked at whether a high fit score for the Elephant drawing co-occurred with a high score for the Lion drawing, regardless of the music piece being played. This analysis allowed us to speculate on the existence of higher-order mental categories that can account for the participants' responses beyond the animal level. These categories might include animal species (such as birds and mammals), physical properties (such as size and shape), and more abstract attributes/qualities (such as clumsy, elegant, or aggressive).

Given the influence of timbre on participants' associations, two separate cluster analyses were conducted: one for the orchestra and one for the piano. Both analyses resulted in a four-cluster solution based on the examination of 23 indices (see Table 2 Cluster Piano and Orchestra below).

A Mann-Whitney U test was then run for each of the four clusters comparing the standardised fit scores of orchestra vs. piano ($ps > .620$). These tests failed to reveal any significant difference between the considered comparisons, thus indicating that, within the same cluster, similar associative patterns existed across timbres. Therefore, we decided to delve deeper into the association profiles by grouping all observations into the same analysis.

The overall k-means cluster analysis resulted in a four-cluster solution (see Figure 2) as suggested by 10 indices, followed by a 3-cluster solution ($N = 6$), 5-cluster ($N = 4$), and 7-cluster ($N = 3$). Cluster 1 presented very high scores[1] for the Lion ($M = 0.85$) and Elephant ($M = 0.63$) and low scores for all other animals ($M < 0.32$), while Chicken ($M = 0.26$) and Turtle ($M = 0.17$) had very low scores. Cluster 2 was characterised by very high scores for the Swan ($M = 0.85$) and low scores for all other animals ($M < 0.53$). In Cluster 3, the Chicken ($M = 0.84$) and Cuckoo ($M = 0.76$) presented very high scores, whereas the Elephant had a very low score ($M = 0.19$). Finally, Cluster 4 was characterised by very high scores for the Turtle ($M = 0.83$) and Elephant ($M = 0.79$), whereas the Cuckoo had a very low score ($M = 0.38$), similar to that of the Chicken ($M = 0.43$).

To further explore participants' association strategies, we assigned each observation to a specific cluster and then analysed the music-drawings association profiles within each cluster. This method enabled us to evaluate how accurately participants matched musical excerpts with the "correct" cluster, defined as the cluster containing the corresponding character (see Figure 3).

This analysis further confirms the distinctiveness of "The Swan," which was assigned to the correct cluster in 64% of the matchings. The music of the Elephant

Table 2. K-means cluster analysis. Composition of the four clusters for both timbres.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	Orchestra	Piano	Orchestra	Piano	Orchestra	Piano	Orchestra	Piano
Lion	0.82	0.85	0.21	0.17	0.30	0.28	0.43	0.55
Chicken	0.27	0.21	0.11	0.15	0.81	0.82	0.54	0.68
Turtle	0.30	0.33	0.42	0.48	0.24	0.31	0.81	0.79
Elephant	0.72	0.69	0.34	0.34	0.12	0.14	0.63	0.75
Donkey	0.34	0.34	0.11	0.22	0.49	0.58	0.83	0.75
Cuckoo	0.30	0.22	0.47	0.49	0.82	0.80	0.48	0.53
Swan	0.32	0.31	0.84	0.84	0.42	0.40	0.59	0.52
U	24		20		23		22	
p	>.999		.620		.902		.801	

In the last two rows, the U and *p* values are reported for Mann-Whitney U tests.

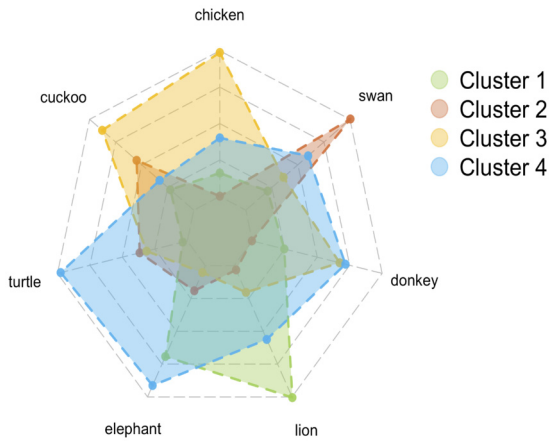

Figure 2. Representation of the four-cluster solution provided by the k-means cluster analysis.

was also relatively well-matched with the correct cluster (42%) compared to other clusters, similar to the Chicken being matched with Cluster 3 (58%). The Lion was correctly matched more often (39%) with the Big Animals cluster than with any other. However, some associations proved more ambiguous. For instance, the Cuckoo was predominantly linked with Cluster 4 (37%, "Slow Animals"), while the Turtle was more frequently assigned to the "Swan" Cluster (41%) than to the "Slow Animals" Cluster (31%). This observation might be expected, considering the shared calm and unhurried traits between turtles and swans. This analysis underscores the ambiguous nature of the music associated with the Donkey, which was nearly equally linked with the Big Animals cluster (47%) and the Feathered cluster (37%).

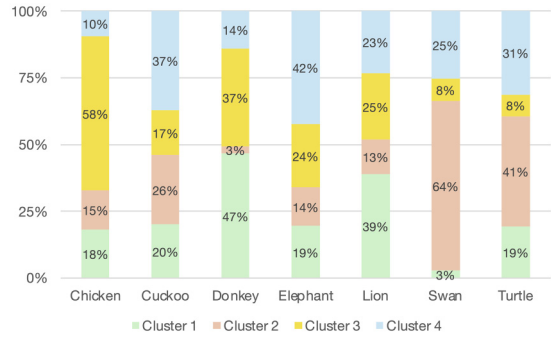

Figure 3. Representation of the distribution of observations within each cluster. For example, the music of the Chicken is matched with Cluster 1 18% of the time, with Cluster 2 15%, with Cluster 3 58%, and with Cluster 4 10%.

Discussion

This study aimed to investigate whether and how timbre affects crossmodal associations across cultures and languages using complex audiovisual stimuli. The findings demonstrated that participants' matchings of musical excerpts from Saint-Saëns' *Carnival of the Animals* with coloured drawings were consistent across cultures, languages, and levels of musical expertise. However, these associations were significantly affected by the timbre of the musical stimuli, namely, orchestral versus piano.

The musical excerpt of the Swan exhibited extraordinarily consistent association profiles, with participants most likely associating it with the drawing of the Swan across all conditions and samples. Similarly, the musical excerpt of the Elephant showed consistent associative profiles, with participants matching it preferentially with the drawing of the Elephant across all conditions and samples. Other characters, such as the Lion and the Donkey, exhibited less consistent, timbre-dependent matchings.

The k-means cluster analysis offers insights into the factors influencing associations. Cluster 1 appears to group animals primarily by size (large) and majesty, specifically the Lion and Elephant (while the Chicken and Turtle receive notably lower scores in this cluster). This aligns with association results showing that the Lion's musical excerpt is significantly more frequently linked to the drawing of the Elephant than to any other drawings, particularly in orchestral timbre.

Cluster 2 consists solely of the Swan, forming a single-item cluster. This can be attributed to the (represented or imagined) Swan's unique elegance, smooth movement, and calm demeanour, as supported by associations where the Swan's musical excerpt consistently correlates with drawings of the Swan in both orchestral and piano timbres.

Cluster 3 includes the Chicken and Cuckoo, characterised by their small size, jerky movements, and feathers. This similarity is reflected in associations where the

Chicken's musical excerpt is most frequently associated with the drawing of the Cuckoo across both orchestral and piano timbres.

Finally, Cluster 4 groups the Turtle and Elephant, categorised as slow, calm animals, though less elegant and graceful compared to the Swan. Association results confirm similarities between these animals, showing that the Turtle music is significantly more associated with the Elephant's drawing, particularly in orchestral timbre.

Overall, the cluster analysis revealed that various higher-order semantic categories likely influenced participants' associations (see also Cohen, 1993, for associations involving audiovisual complex dynamic stimuli, i.e., films). These categories include physical properties such as size (large/small) and features like feathers, dynamic properties such as velocity and fluidity, as well as more abstract features such as elegance and calmness.

Investigating the reasons behind this, one might consider that timbre differences are more ambiguous and elusive than those based on melody and rhythm, making them difficult to account for in cognitivist theories of resemblance. For example, in Davies' account (e.g., 1994, 2008), music is emotionally expressive as it resembles bodily movements or behaviours that manifest emotional or inner states (see also Hubbard, 2017; Kivy, 1989; Larson, 2012). However, while appearance emotionalism seemingly works well in explaining the expressiveness of musical features associated with movement, primarily melody (Sievers et al., 2013) and rhythm, it fails to account for the expressive quality of timbre, which lacks any phenomenologically evident connection with (human) movement (Reymore et al., 2023). However, timbre may be reflective of the vitality contours that have been proposed to constitute the affective intersubjectivity of infants (Stern, 1999). Alternative explanations to account for timbre expressivity have been provided in the recent debate based, for example, on the notion of 'atmosphere' (Di Stefano, 2023; see also Ravasio, 2017).

Finally, gaining insights from the perspective of association profiles and musical compositions reveals some interesting findings. One might assume that incorporating elements in musical compositions that directly evoke the acoustic features (like voice) of characters could aid participants in correctly identifying associations. However, our findings indicate that this assumption does not always hold true. It is noteworthy that although the musical composition for the Donkey includes elements in both its orchestral and piano versions meant to explicitly mimic the animal's distinctive "hee-haw" sound, these elements do not significantly enhance its association with the Donkey over other animals. In contrast, the music linked with the Swan, which is frequently associated with the Swan animal, lacks literal elements that reproduce the animal's characteristic acoustic features. This suggests that higher-order, abstract qualities such as elegance and grace can be conveyed more effectively through music than specific acoustic properties that are associated with animal vocalizations (at least partially in line with the emotional mediation hypothesis, see Spence, 2020, and with findings from Di Stefano et al., 2024; see also the 'visual imagery' and 'musical expectation' mechanisms of the BRECVEMA framework elaborated by Juslin, 2013).

Author Contribution

Conceptualisation, N.D.S., A.A., A.S., C.S.; Methodology, N.D.S, A.A., A.S., C.S.; Investigation, N.D.S. and A.A.; Formal Analysis, A.A.; Writing – Original Draft, N.D.S. and A.A.; Writing – Review & Editing, A.S., S.S., and C.S.; Funding Acquisition, N.D.S.; S.S.; Supervision, N.D.S. and C.S.

Data Availability

The datasets generated during the current study are available from the corresponding author upon reasonable request.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethics

The protocol was approved by the Research Ethics and Integrity Committee of the National Research Council of Italy.

Funding

This research received financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union - NextGenerationEU- Project Title "The multisensory and interactional nature of music experience. Merging aesthetics with bioengineering to investigate the multidimensional structure of musical sound." Grant Assignment Decree No. 1079 adopted on 19/07/2023 by the Italian Ministry of Ministry of University and Research (MUR). Financial support for the research carried out in this work was also provided by the Research Council of Finland [grant number 346210].

ORCID iDs

Nicola Di Stefano https://orcid.org/0000-0002-9286-0395
Alessandro Ansani https://orcid.org/0000-0002-0657-5732

Supplemental Material

Supplemental material for this article is available online.

Note

1. 'Very high scores' stand for scores > 1SD from the mean of the cluster, while 'very low scores' indicates scores < 1SD from the mean of the cluster.

References

Adeli, M., Rouat, J., & Molotchnikoff, S. (2014). Audiovisual correspondence between musical timbre and visual shapes. *Frontiers in Human Neuroscience*, *8*, 352. https://doi.org/10.3389/fnhum.2014.00352

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In J. Van Den Bussche, & V. Vianu (Eds.), *Database theory—ICDT 2001* (Vol. 1973, pp. 420–434). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44503-X_27.

Albertazzi, L., Canal, L., & Micciolo, R. (2015). Cross-modal associations between materic painting and classical Spanish music. *Frontiers in Psychology*, *6*, 424. https://doi.org/10.3389/fpsyg.2015.00424

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brunel, L., Carvalho, P. F., & Goldstone, R. L. (2015). It does belong together: Cross-modal correspondences influence cross-modal integration during perceptual learning. *Frontiers in Psychology*, *6*, 358. https://doi.org/10.3389/fpsyg.2015.00358

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*, 1–36. https://doi.org/10.18637/jss.v061.i06

Cohen, A. J. (1993). Associationism and musical soundtrack phenomena. *Contemporary Music Review*, *9*(1-2), 163–178. https://doi.org/10.1080/07494469300640421

Cowles, J. T. (1935). An experimental study of the pairing of certain auditory and visual stimuli. *Journal of Experimental Psychology*, *18*, 461–469. https://doi.org/10.1037/h0062202

Davies, S. (1994). *Musical meaning and expression*. Cornell University Press.

Davies, S. (2008). Introduction to a philosophy of music. *Philosophy and Phenomenological Research*, *76*(1), 222–224. https://doi.org/10.1111/j.1933-1592.2007.00128.x

Deroy, O., Crisinel, A.-S., & Spence, C. (2013). Crossmodal correspondences between odors and contingent features: Odors, musical notes, and geometrical shapes. *Psychonomic Bulletin & Review*, *20*, 878–896. https://doi.org/10.3758/s13423-013-0397-0

Deroy, O., & Spence, C. (2013). Why we are not all synesthetes (not even weakly so). *Psychonomic Bulletin & Review*, *20*, 643–664. https://doi.org/10.3758/s13423-013-0387-2

Di Stefano, N. (2023). Musical emotions and timbre: From expressiveness to atmospheres. *Philosophia*, *51*(5), 2625–2637. https://doi.org/10.1007/s11406-023-00700-6

Di Stefano, N., Ansani, A., Schiavio, A., & Spence, C. (2024). Prokofiev was (almost) right: A cross-cultural investigation of auditory-conceptual associations in Peter and the Wolf. *Psychonomic Bulletin & Review*, *31*, 1735–1744. https://doi.org/10.3758/s13423-023-02435-7

Di Stefano, N., & Spence, C. (2022). Roughness perception: A multisensory/crossmodal perspective. *Attention, Perception, & Psychophysics*, *84*(7), 2087–2114. https://doi.org/10.3758/s13414-022-02550-y

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). Multipic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816. https://doi.org/10.1080/17470218.2017.1310261

Eerola, T., Ferrer, R., & Alluri, V. (2012). Timbre and affect dimensions: Evidence from affect and similarity ratings and acoustic correlates of isolated instrument sounds. *Music Perception*, *30*(1), 49–70. https://doi.org/10.1525/mp.2012.30.1.49

Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition*, *114*(3), 405–422. https://doi.org/10.1016/j.cognition.2009.10.013

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 6. https://doi.org/10.1167/10.1.6

Fife, D. (2022). Flexplot: Graphically-based data analysis. *Psychological Methods*, *27*(4), 477–496. https://doi.org/10.1037/met0000424

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

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Grossenbacher, P. G., & Lovelace, C. T. (2001). Mechanisms of synesthesia: cognitive and physiological constraints. *Trends in Cognitive Sciences*, *5*(1), 36–41. https://doi.org/10.1016/s1364-6613(00)01571-0

Gurman, D., McCormick, C. R., & Klein, R. M. (2021). Crossmodal correspondence between auditory timbre and visual shape. *Multisensory Research*, *35*(3), 221–241. https://doi.org/10.1163/22134808-bja10067

Hailstone, J. C., Omar, R., Henley, S. M., Frost, C., Kenward, M. G., & Warren, J. D. (2009). It's not what you play, it's how you play it: Timbre affects perception of emotion in music. *Quarterly Journal of Experimental Psychology*, *62*(11), 2141–2155. https://doi.org/10.1080/17470210902765957

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

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley and Sons, Inc.

Hashim, S., Stewart, L., Küssner, M. B., & Omigie, D. (2023). Music listening evokes story-like visual imagery with both idiosyncratic and shared content. *PLoS One*, *18*(10), e0293412. https://doi.org/10.1371/journal.pone.0293412

Heo, M., & Leon, A. C. (2010). Sample sizes required to detect two-way and three-way interactions involving slope differences in mixed-effects linear models. *Journal of Biopharmaceutical Statistics*, *20*(4), 787–802. https://doi.org/10.1080/10543401003618819

Hubbard, T. L. (1996). Synesthesia-like Mappings of Lightness, Pitch, and Melodic Interval. *The American Journal of Psychology*, *109*(2), 219–238. https://doi.org/10.2307/1423274

Hubbard, T. L. (2017). Momentum in music: Musical succession as physical motion. *Psychomusicology: Music, Mind, and Brain*, *27*(1), 14–30. https://doi.org/10.1037/pmu0000171

Iosifyan, M., Sidoroff-Dorso, A., & Wolfe, J. (2022). Cross-modal associations between paintings and sounds: Effects of embodiment. *Perception*, *51*(12), 871–888. https://doi.org/10.1177/03010066221126452

Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of Life Reviews*, *10*(3), 235–266. https://doi.org/10.1016/j.plrev.2013.05.008

Kivy, P. (1989). *Sound sentiment: An essay on the musical emotions*. Temple University Press.

Klapetek, A., Ngo, M. K., & Spence, C. (2012). Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Attention, Perception, & Psychophysics*, *74*, 1154–1167. https://doi.org/10.3758/s13414-012-0317-9

Krzanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, *44*(1), 23–34. https://doi.org/10.2307/2531893

Kumar, R. (2017). *Analysis of shape alignment using Euclidean and Manhattan distance metrics*. *2017 International Conference on Recent Innovations in Signal Processing and Embedded Systems (RISE)* (pp. 326–331). https://doi.org/10.1109/RISE.2017.8378175

Larson, S. (2012). *Musical forces: Motion, metaphor, and meaning in music*. Indiana University Press.

Lenth, R. (2024). *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.10.1. https://CRAN.R-project.org/package=emmeans

Liu, J., Zhao, A., Wang, S., Li, Y., & Ren, H. (2021). Research on the correlation between the timbre attributes of musical sound and visual color. *IEEE Access*, *9*, 97855–97877. https://doi.org/10.1109/ACCESS.2021.3095197

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American Journal of Psychology*, *87*(1-2), 173–188. https://doi.org/10.2307/1422011

Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(3), 384–394. https://doi.org/10.1037/0096-1523.13.3.384

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press.

Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(1), 69–79. https://doi.org/10.1037/0096-1523.15.1.69

Miller, R. (2021). The semantic differential in the study of musical perception: A theoretical overview. *Visions of Research in Music Education*, *16*, 11.

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

Motoki, K., Marks, L., & Velasco, C. (2023). Reflections on cross-modal correspondences: Current understanding and issues for future research. *Multisensory Research*, *37*(1), 1–23. https://doi.org/10.1163/22134808-bja10114

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS One*, *9*, e89642. https://doi.org/10.1371/journal.pone.0089642

Ortmann, O. (1922). The sensorial basis of music appreciation. *Journal of Comparative Psychology*, *2*(3), 227–256. https://doi.org/10.1037/h0072000

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: A study using the implicit association test. *Experimental Brain Research*, *220*, 319–333. https://doi.org/10.1007/s00221-012-3140-6

Ravasio, M. (2017). Emotions in the listeners: A criterion of artistic relevance. *American Society for Aesthetics Graduate E-Journal*, *9*(1), 1–9.

Reymore, L., Noble, J., Saitis, C., Traube, C., & Wallmark, Z. (2023). Timbre semantic associations vary both between and within instruments: An empirical study incorporating register and pitch height. *Music Perception: An Interdisciplinary Journal*, *40*(3), 253–274. https://doi.org/10.1525/mp.2023.40.3.253

Rigg, M. G. (1937). An experiment to determine how accurately college students can interpret intended meanings of musical compositions. *Journal of Experimental Psychology*, *21*, 223–229. https://doi.org/10.1037/h0056146

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*(1), 1–12. https://doi.org/10.1037/a0018326

Rosi, V., Sarah, P. A., Houix, O., Misdariis, N., & Susini, P. (2023). Shared mental representations underlie metaphorical sound concepts. *Scientific Reports*, *13*, 5180. https://doi.org/10.1038/s41598-023-32214-2

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Saint Saëns. (1886/1922). *Le Carnaval Des Animaux. Grande Fantaisie Zoologique*. Durand & Fils.

Saint Saëns. (1962). *Carnival of the Animals*. [Album recorded by The New York Philharmonic Orchestra]. Columbia Broadcasting System.

Sievers, B., Polansky, L., Casey, M., & Wheatley, T. (2013). Music and movement share a dynamic structure that supports universal expressions of emotion. *Proceedings of the National Academy of Sciences*, *110*(1), 70–75. https://doi.org/10.1073/pnas.1209023110

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*, 971–995. https://doi.org/10.3758/s13414-010-0073-7

Spence, C. (2020). Assessing the role of emotional mediation in explaining crossmodal correspondences involving musical stimuli. *Multisensory Research*, *33*(1), 1–29. https://doi.org/10.1163/22134808-20191469

Spence, C. (2023). Explaining visual shape-taste crossmodal correspondences. *Multisensory Research*, *36*, 313–345. https://doi.org/10.1163/22134808-bja10096

Stern, D. N. (1999). Vitality contours: The temporal contour of feelings as a basic unit for constructing the infant's social experience. In P. Rochat (Ed.), *Early social cognition: Understanding others in the first months of life* (pp. 67–80). Lawrence Erlbaum Associates.

Sun, X., Li, X., Ji, L., Han, F., Wang, H., Liu, Y., Chen, Y., Lou, Z., & Li, Z. (2018). An extended research of crossmodal correspondence between color and sound in psychology and cognitive ergonomics. *PeerJ*, *6*, e4443. https://doi.org/10.7717/peerj.4443

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Trainor, L. J., & Trehub, S. E. (1992). The development of referential meaning in music. *Music Perception*, *9*(4), 455–470. https://doi.org/10.2307/40285565

Walker, R. (1987). The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception & Psychophysics*, *42*(5), 491–502. https://doi.org/10.3758/BF03209757

Wallmark, Z., Nghiem, L., & Marks, L. E. (2021). Does timbre modulate visual perception? Exploring crossmodal interactions. *Music Perception: An Interdisciplinary Journal*, *39*(1), 1–20. https://doi.org/10.1525/mp.2021.39.1.1

Author Biographies

Nicola Di Stefano is a researcher in cognitive science with expertise in music perception and cognition. His work focuses on consonance/dissonance and crossmodal perception involving auditory stimuli.

Alessandro Ansani is a postdoctoral researcher in music psychology, with research interests spanning audiovisual associations and empirical studies on the role of soundtracks in interpreting movie scenes.

Andrea Schiavio is a professor of music education, whose research explores music perception and cognition, with a particular emphasis on the 4E approach to understanding musical experience.

Suvi Saarikallio is a professor of music education. In her research, she examines music as a facet of human behavior, focusing on youth development, emotion regulation, learning, and wellbeing.

Charles Spence is a professor of psychology and a world-renowned expert in crossmodal perception, conducting research across modalities ranging from spatial to chemical senses.