A framework for organising audio-visual cross-modal correspondences for the Soundsketcher project

Konstantinos Giannos[1,*], Asterios Zacharakis[1], George Athanasopoulos[2] & Emilios Cambouropoulos[1]
[1]School of Music Studies, Aristotle University of Thessaloniki
[2]Humboldt-Universität zu Berlin
[*]giannosk@mus.auth.gr (υπεύθυνου)

ABSTRACT

The study of cross-modal correspondences, the systematic association between different sensory modalities, has steadily increased in recent years, with examples like the bouba-kiki effect illustrating these associations between shapes and sounds. Besides this concept, aural sonology explores how sound can be visually represented, particularly in the analysis of electro-acoustic music. In parallel, graphic score creation links visual elements with sounds, often relying on artistic intuition. The Soundsketcher project seeks to develop a tool for automatic music visualisation using established audio-visual correspondences, aiming to map musical characteristics like pitch, loudness, and timbre to visual features such as vertical position, size, angularity, and hue. The project will evaluate how well these mappings translate complex audio into graphic scores and their potential for artistic expression.

Πλαίσιο οργάνωσης οπτικο-ακουστικών αντιστοιχίσεων για το έργο Soundsketcher

ΠΕΡΙΛΗΨΗ

Η μελέτη των διαισθητηριακών αντιστοιχίσεων, δηλαδή της συστηματικής συσχέτισης μεταξύ διαφορετικών αισθητηριακών οδών, αναπτύσσεται σταθερά, με παραδείγματα όπως το φαινόμενο μπούμπα-κίκι που αναδεικνύει τη συσχέτιση σχημάτων με ήχους. Πέραν τούτου, η ακουστική ηχολογία ερευνά την οπτική αναπαράσταση του ήχου, ιδιαίτερα κατά την ανάλυση ηλεκτρακουστικής μουσικής. Παράλληλα, η δημιουργία γραφικών παρτιτούρων συνδυάζει οπτικά στοιχεία με ήχους, συχνά αξιοποιώντας καλλιτεχνικά κριτήρια. Το έργο Soundsketcher προσβλέπει στην ανάπτυξη ενός εργαλείου αυτόματης οπτικοποίηση της μουσικής βασισμένο σε καθιερωμένες οπτικο-ακουστικές αντιστοιχίσεις, επιδιώκοντας να χαρτογραφήσει μουσικά χαρακτηριστικά όπως το τονικό ύψος, την ένταση, ή τη χροιά σε οπτικά χαρακτηριστικά, όπως η κατακόρυφη θέση, το μέγεθος, η ύπαρξη γωνιών και η απόχρωση. Το εργαλείο θα αξιολογηθεί ως προς τη μετάφραση πιο περίπλοκων ήχων σε γραφική παρτιτούρα και το δυνητικό πλαίσιο καλλιτεχνικής έκφρασης.

Introduction

Over the last decades, there has been an increasing interest in cross-modal correspondences where attributes and features from different sensory modalities are systematically associated with each other. One famous example of such correspondence is the bouba-kiki, or maluma-takete effect. According to that, round and smooth objects tend to be matched with "bouba", contrary to angular and rough objects that tend to be matched with "kiki" [1]. Associations between the auditory and the visual (i.e., spatial, geometrical, etc) extend beyond simple shapes and sound pairings, such as 2- or 3-dimensional visualisations, mental imagery [2], free-hand drawings [3-4], or gestural motion [5].

Building on this idea of translating sound into visual terms, the field of aural sonology specifically investigates how sound can be visually represented and analysed. In that domain, various systems have been developed to analyse and represent electro-acoustic music, enabling users to manually interact with the systems, such as ianalyse 5 [6] or Acousmographe [7].

In parallel to these analytical systems, graphic score creation is an artistic practice that similarly pairs visual elements with sounds, offering a guide to sound production or a visual analogy of auditory experiences. Such approaches often rely more on artistic intuition than on established cross-modal mappings. The Soundsketcher project seeks to create a prototype application for automatic music visualisation through graphic scores. One of the fundamental goals is to base mappings between sonic and visual structures on associations derived from existing knowledge of audio-visual correspondences. This study presents the currently identified cross-modal correspondences between musical or acoustic characteristics and visual properties as the framework that underpins Soundsketcher's mappings.

1. Sound properties - Visual correspondences

1.1 Pitch

One of the most extensively studied cross-modal relationships is that between musical pitch—commonly represented by the fundamental frequency (F0)—and its visual counterparts. A key association in this relationship is *vertical elevation*: 'high' corresponds to high pitch and 'low' to low pitch (e.g., [8]), which is observed in many languages of the world [9-10] and is reflected in the Western music notation system. Additional findings have revealed that the *horizontal dimension* has been mapped onto pitch with 'left' corresponding to lower pitches and 'right' to higher pitches [5], which appears to be a comparatively weaker connection as it is found mainly in pianists [11]. *Size* is another visual attribute that has been related to pitch. Specifically, large objects have been linked to low pitches, while small objects to high pitches [12], with the association getting stronger with age [13]. Also, *thickness* is consistently mapped to low pitches and thinness to high pitches [14]. Finally, an association between *brightness* and pitch height has been observed in individual pitches, meaning that brighter colours correspond to higher pitches [15], as well as

ascending melodic motions correspond to bright colours and descending motions are mapped to darker colours [16].

1.2 Loudness

Besides pitch height, loudness is a percept accompanied by a long list of studied associations and has been modelled in various ways [17-18]. Regarding *spatial location*, louder sounds were found to be related to higher elevation [18], or horizontal movement [5]. Albeit both pitch and loudness have been associated with vertical elevation, in simultaneous alteration of both, the pitch association appears to be stronger than the loudness one [5]. Additionally, large objects are typically matched to louder sounds, while small objects are matched to quieter sounds [20]. Lastly, strong associations have been observed between loudness and *brightness* in adults [21] and infants [22].

1.3 Time

Visual representations of time in music notation tend to be influenced by culture and language. Written language is linear, and some possible directions are left-to-right (e.g., Latin, Greek, Cyrillic alphabets, etc.), right-to-left (e.g., Arabic, Hebrew), or top-to-bottom (e.g., Japanese kanji), thus, informing the way these populations may represent time [9]. For example, traditional Western musical notation is arranged linearly in a left-to-right fashion, whereas a portion of traditional Japanese musicians opted for vertical, top-to-bottom representations. At the same time, nonliterate participants largely unexposed to Western culture opted for iconic representations of sound without necessarily representing time [4], though line segment length had been found to be proportional to sound duration for multiple populations [9].

1.4 Timbre

Timbre is one of the least explored sound elements for its potential visual analogues. A handful of studies have examined the relationship between timbre and shape, where instruments producing soft sounds such as the piano or the cello were associated with rounded shapes and instruments such as crash cymbals were associated with angular shapes [23]. At the same time, listeners have linked auditory roughness with jagged and spiky 2- and 3-dimensional shapes [24-25].

Despite the scarcity of studies on direct timbral-visual associations, recent works have identified some salient semantic dimensions of timbre such as brightness, roughness or mass [26] and their various nuances [27-28]. Interestingly, most of these semantic concepts can be visually represented. However, timbre is considerably more complex to model perceptually compared to pitch and loudness. Identifying perceptually relevant timbre features is already a complex task, with the identification of semantically relevant ones posing an even greater challenge. Nevertheless, evidence on the acoustic correlates of semantic concepts continues to accumulate. For example, sound sources characterised by an energy distribution that is skewed towards higher partials [26], strong temporal modulations or prominent noisy components [29-31] are more likely to be described as rough. Similarly,

high-frequency components, usually quantified by the spectral centroid, are typically associated with higher auditory brightness (e.g., [32]. The concept of auditory mass is more elusive in acoustic terms, with some evidence suggesting a positive link with spectrotemporal variation and inharmonicity [26] or with loudness, spectral saturation and low registers [28]. Notably, the majority of these studies focus on isolated musical sounds, while mapping concurrent audio streams to graphic representations introduces an additional level of complexity. This challenge can be at least partially addressed through the advancing capabilities of AI-powered sound source separation applications for musical signals.

1.5 Tonality-Harmony

An increasing amount of research focuses on higher-level musical features such as tonality or harmony. Listeners regard the lack of functional tonal progressions reflected in diatonic modes as an element of dissonance and, in turn, map such musical stimuli to rough images of pixel noise [33]. Similarly, the dissonance of chords in musical excerpts is matched with both visual and tactile representations of roughness [34].

2. Visual mapping of Soundsketcher

2.1 Selecting a visual representation

Here, we propose a series of visual analogues to the auditory concepts reviewed to support the Soundsketcher system. Certainly, this task is far from trivial, and some overlaps may exist; for instance, Smalley [35] highlights the existence of noises with high pitch content or pitches with high noise content (e.g., whistling, breathy speech, screeching, etc.). Nonetheless, we opted for the following pairs acknowledging the empirical research regarding common cross-modal correspondences and the Western classical paradigm. As summarised in the table, pitch is assigned to vertical elevation, and time to the left-to-right direction. Duration is matched to length, and loudness is mapped to area/object size. Timbral concepts such as fullness, thickness, or density, have been theoretically connected to the spectral flux, inharmonicity and loudness, possibly extending the idea of size to solidness/hollowness. Regarding other prominent timbral semantics, roughness can be mapped to angularity and brightness to hue and/or colour. In the figure below, it is demonstrated how a rectangle/line segment is manipulated along these visual aspects to convert audio samples into a computer-generated graphic score. Acknowledging the cultural and artistic dimensions of these correspondences, users will be free to choose the correspondences of their preference.

Table 2.1 Proposed visual concepts as analogues to auditory/acoustic concepts

Auditory concept	**Acoustic feature**	**Visual concept**
pitch	F0	elevation
loudness	loudness models	object size/elevation

time direction	-	left-to-right
duration	onset detection, segmentation	length
roughness	modulation power spectrum, harmonic to noise ratio	angularity
brightness	spectral centroid	hue/colour, brightness
thickness	spectral flux, inharmonicity, loudness	thickness

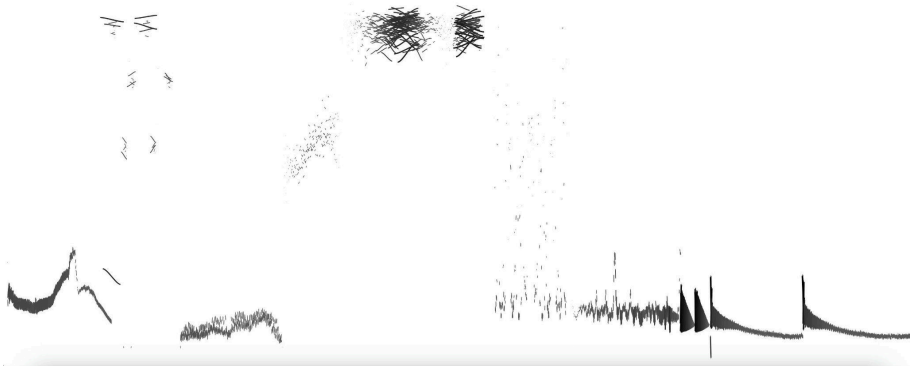

Figure 2.1 Screenshot of the Soundsketcher system displaying a graphic representation of an audio file

2.2 Evaluation

The initial application of the mappings as they appear in Figure 2.1, will be tested in simpler case studies. The rising gestures can be interpreted as rising tones, the larger objects as louder, the shapes with more angles as sonically rougher and so on. Starting from simple audio samples where certain audio features (e.g., pitch and loudness) alter distinctly, we plan to assess whether users will understand the graphic score as a translation of the audio and if alternative correspondences are preferred. Subsequently, more complex examples involving combinatory alterations of acoustic features will be introduced. Furthermore, this way we aim to examine if other types of representations are useful, considering prescriptive aspects such as the origin of a sound or the necessary action to produce it. In addition, we expect to explore whether these correspondences are restrictive or not in artistic endeavours and how the generated graphic scores are open to interpretations. Later, more complex audio with diverse timbral characteristics and simultaneously changing audio features will be investigated, along with how this complexity is reflected in the multivariate manipulations of the visual features.

3. Conclusion

In conclusion, this paper presents the foundation of the Soundsketcher system's graphic score generation by proposing mappings between sound characteristics such as pitch, loudness, and timbre to visual properties like vertical elevation, size, and shape. The proposed correspondences are informed by empirical research, while also allowing flexibility for users to reflect their individual preferences. The evaluation phase will further investigate the clarity and interpretability of these mappings, particularly in more complex auditory scenarios. Ultimately, Soundsketcher aims to purposefully combine the auditory and visual experiences, enhancing both artistic creativity and music interpretation.

4. Acknowledgement

This project is carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union – NextGenerationEU (Implementation body: HFRI).

5. References

[1] V. S. Ramachandran and E. M. Hubbard, "Psychophysical investigations into the neural basis of synaesthesia," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **268**(1470), pp. 979–983, (2001). doi:10.1098/rspb.2000.1576

[2] Z. Eitan and R. Y. Granot, "How music moves," *Music Perception*, **23**(3), pp. 221–248, (2006). doi:10.1525/mp.2006.23.3.221

[3] L. Engeln and R. Groh, "Cohearence of audible shapes—a qualitative user study for coherent visual audio design with resynthesized shapes," *Personal and Ubiquitous Computing*, **25**(4), pp. 651–661, (2020). doi:10.1007/s00779-020-01392-5

[4] G. Athanasopoulos and N. Moran, "Cross-cultural representations of musical shape," *Empirical Musicology Review*, pp. 185–199, (2013). doi:10.18061/emr.v8i3-4.3940

[5] M. B. Küssner, D. Tidhar, H. M. Prior, and D. Leech-Wilkinson, "Musicians are more consistent: Gestural cross-modal mappings of Pitch, Loudness and tempo in real-time," *Frontiers in Psychology*, **5**, (2014). doi:10.3389/fpsyg.2014.00789

[6] P. Couprie, "EAnalysis: Developing a sound-based music analytical tool," *Expanding the Horizon of Electroacoustic Music Analysis*, pp. 170–194, (2016). doi:10.1017/cbo9781316339633.009

[7] Y. Geslin and A. Lefèvre. Sound and musical representation: the Acousmographe software. In *Proceedings in International Conference on Mathematics and Computing (ICMC2004), San Francisco, California, USA*. (2004).

[8] K. Uno and K. Yokosawa, "Cross-modal correspondence between Auditory Pitch and visual elevation modulates audiovisual temporal recalibration," *Scientific Reports*, **12**(1), Dec. 2022. doi:10.1038/s41598-022-25614-3

[9] G. Athanasopoulos, S.-L. Tan, and N. Moran, "Influence of literacy on representation of time in musical stimuli: An exploratory cross-cultural study in the

UK, Japan, and Papua New Guinea," *Psychology of Music*, **44**(5), pp. 1126–1144, (2016). doi:10.1177/0305735615613427

[10] E. Rusconi, B. Kwan, B. Giordano, C. Umiltà, and B. Butterworth, "Spatial representation of pitch height: The SMARC effect," *Cognition*, **99**(2), pp. 113–129, (2006). doi:10.1016/j.cognition.2005.01.004

[11] L. Stewart, V. Walsh, and U. Frith, "Reading music modifies spatial mapping in pianists," *Perception & Psychophysics*, **66**(2), pp. 183–195, (2004). doi:10.3758/bf03194871

[12] L. J. Speed, I. Croijmans, S. Dolscheid, and A. Majid, "Crossmodal associations with olfactory, auditory, and tactile stimuli in children and adults," *i-Perception*, **12**(6), (2021). doi:10.1177/20416695211048513

[13] L. F. Cuturi, A. Tonelli, G. Cappagli, and M. Gori, "Coarse to fine audio-visual size correspondences develop during primary school age," *Frontiers in Psychology*, **10**, (2019). doi:10.3389/fpsyg.2019.02068

[14] S. Dolscheid, S. Shayan, A. Majid, and D. Casasanto, "The thickness of musical pitch," *Psychological Science*, **24**(5), pp. 613–621, (2013). doi:10.1177/0956797612457374

[15] J. Ward, B. Huckstep, and E. Tsakanikos, "Sound-colour synaesthesia: To what extent does it use cross-modal mechanisms common to us all?," *Cortex*, **42**(2), pp. 264–280, (2006). doi:10.1016/s0010-9452(08)70352-6

[16] W. G. Collier and T. L. Hubbard, "Judgments of happiness, brightness, speed and tempo change of auditory stimuli varying in pitch and tempo.," *Psychomusicology: A Journal of Research in Music Cognition*, **17**(1–2), pp. 36–55, (1998). doi:10.1037/h0094060

[17] P. Boersma and D. Weenink. Praat: doing phonetics by computer. Version 5.3.15. (2012).

[18] B. R. Glasberg and B. C. Moore. A model of loudness applicable to time-varying sounds. Journal of the Audio Engineering Society, **50**(5), 331-342. (2002).

[19] D. Kohn and Z. Eitan. Seeing Sound Moving: Congruence of Pitch and Loudness with Human Movement. In *12th International Conference on Music Perception and Cognition/8th Triennial Conference of the European Society for the Cognitive Sciences of Music*. Thessaloniki: The School of Music Studies, Aristotle University of Thessaloniki, p. 541 (2012).

[20] Z. Eitan. How pitch and loudness shape musical space and motion: New findings and persisting questions. In *The psychology of music in multimedia*, edited by S.-L. Tan, A. Cohen, S. Lipscomb, and R. Kendall, pp. 161-187. Oxford: Oxford University Press. (2013).

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

[22] D. J. Lewkowicz and G. Turkewitz, "Cross-modal equivalence in early infancy: Auditory–visual intensity matching.," *Developmental Psychology*, **16**(6), pp. 597–607, (1980). doi:10.1037/0012-1649.16.6.597

[23] M. Adeli, J. Rouat, and S. Molotchnikoff, "Audiovisual correspondence between musical timbre and visual shapes," *Frontiers in Human Neuroscience*, **8**, (2014). doi:10.3389/fnhum.2014.00352

[24] K. Liew, S. J. Styles, and P. Lindborg, Dissonance and roughness in cross-modal perception. In *Proceedings of the 6th Conference of the Asia Pacific Society for the Cognitive Sciences of Music*. (2017).

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

[26] A. Zacharakis, K. Pastiadis, and J. D. Reiss, "An interlanguage study of musical timbre semantic dimensions and their acoustic correlates," *Music Perception*, **31**(4), pp. 339–358, (2014). doi:10.1525/mp.2014.31.4.339

[27] L. Reymore, "Characterizing prototypical musical instrument timbres with timbre trait profiles," *Musicae Scientiae*, **26**(3), pp. 648–674, (2021). doi:10.1177/10298649211001523

[28] J. Noble, E. Thoret, M. Henry, and S. McAdams, "Semantic dimensions of sound mass music," *Music Perception*, **38**(2), pp. 214–242, (2020). doi:10.1525/mp.2020.38.2.214

[29] J. Rozé, M. Aramaki, R. Kronland-Martinet, and S. Ystad, "Exploring the perceived harshness of cello sounds by morphing and synthesis techniques," *The Journal of the Acoustical Society of America*, **141**(3), pp. 2121–2136, (2017). doi:10.1121/1.4978522

[30] L. Reymore, E. Beauvais-Lacasse, B. K. Smith, and S. McAdams, "Modeling noise-related timbre semantic categories of orchestral instrument sounds with audio features, pitch register, and Instrument Family," *Frontiers in Psychology*, **13**, (2022). doi:10.3389/fpsyg.2022.796422

[31] V. Rosi, P. Arias Sarah, O. Houix, N. Misdariis, and P. Susini, "Shared mental representations underlie metaphorical sound concepts," *Scientific Reports*, **13**(1), (2023). doi:10.1038/s41598-023-32214-2

[32] A. Almeida, E. Schubert, J. Smith, and J. Wolfe, "Brightness scaling of periodic tones," *Attention, Perception, & Psychophysics*, **79**(7), pp. 1892–1896, (2017). doi:10.3758/s13414-017-1394-6

[33] K. Giannos, G. Athanasopoulos, and E. Cambouropoulos, "Cross-modal associations between Harmonic Dissonance and visual roughness," *Music & Science*, **4**, (2021). doi:10.1177/20592043211055484

[34] K. Giannos, G. Athanasopoulos, and M. Küssner. Cross-modal associations between auditory and tactile roughness across Western and non-Western harmonisations. Music Perception: An Interdisciplinary Journal. (in press).

[35] D. Smalley, "Spectromorphology: Explaining sound-shapes," *Organised Sound*, **2**(2), pp. 107–126, (1997). doi:10.1017/s1355771897009059