Key Clarity is Blue, Relaxed, and Maluma: Machine learning used to discover cross-modal connections between sensory items and the music they spontaneously evoke

Maddalena Murari[1], Anthony Chmiel[2], Enrico Tiepolo[1], J. Diana Zhang[3], Sergio Canazza[1], Antonio Rodà[1], and Emery Schubert[3]

[1] University of Padova, CSC, Dept. of Information Engineering, Italy
[2] Western Sydney University, The MARCS Institute for Brain, Behaviour and Development, Sydney, Australia
[3] University of New South Wales, Empirical Musicology Laboratory, School of the Arts and Media, Sydney, Australia
antonio.roda@unipd.it

Abstract. Semantic differential is often used to investigate the relationship between music and other sensory modalities such as colors, tastes, vision, and odors. This work proposes an exploratory approach including open-ended responses and subsequent machine learning to study cross-modal associations, based on a recently developed sensory scale that does not use any explicit verbal description. Twenty-five participants were asked to report a piece of music they considered close to the feel/look/experience of a given sensory stimulus. Results show that the associations reported by the participants can be explained, at least in part, by a set of features related to some timbric and tonal aspects of music.

Keywords: Cross-modal correspondences, sensory scales, audio features, sound and music computing.

This is a pre-print version of the accepted manuscript.
Please cite this work as listed below:
Murari M., Chmiel, A., Tiepolo, E., Zhang, J. D., Canazza, S., Rodà, A., & Schubert, E. (2020). Key clarity is blue, relaxed, and maluma: Machine learning used to discover cross-modal connections between sensory items and the music they spontaneously evoke. In Shoji H. et al. (Eds) Proceedings of the 8th International Conference on Kansei Engineering and Emotion Research. KEER 2020. Advances in Intelligent Systems and Computing, vol 1256. Springer, Singapore.
https://doi.org/10.1007/978-981-15-7801-4_22

1 Introduction

Music, when soft voices die, / Vibrates in the memory – / Odours, when sweet violets sicken,/ Live within the sense they quicken. [18] An intriguing aspect of music is its capacity to elicit a rich number of sensations and images. Many studies have investigated the relationship between music and colors, tastes, vision, and odors, suggesting that people can exhibit consistent cross-modal responses in different sensory modalities [23,7,8,17,22,21].

According to Spence [19] [12], these connections can be explained by structural correspondences due to similarities of neural coding across modalities. However, the more complex and rich associations conveyed by sections and whole pieces of music are hardly explained by this interpretation. Correspondences may also develop through statistical learning: regularities in the environment – such as the fact that larger objects tend to create louder sounds – would cause an internal link between the senses. Other correspondences may have a semantic origin: "high" pitches and "high" elevations use the same terminology, which could lead to an association between pitch and elevation. According to Palmer [16], a mediating factor (emotion) can provide a more parsimonious explanation for the correspondences between music and color. Kansei models also investigate the connotative meaning of music: Sugihara et al. [20] characterized 12 music pieces from various repertoires by means of 40 pairs of Kansei words. Kinoshita et al. [3], using Osgood's semantic differential, investigated the implementation of a Kansei music selection system that automatically selects suitable music in car audio systems according to the external scenery.

One of the shortcomings of the semantic differential technique is related to the difficulty to grasp the denotative meaning of language. By "denotative" Osgood [11] refers to the descriptive use of signs as contrasted with their emotive or affective use. In the sentence "we set a wall between us", the word "wall" is used to suggest a physical boundary, which is its denotative meaning, but it also implies the idea of an emotional barrier. Osgood's descriptive scales are more concerned with frequency of usage rather than dictionary meaning. Another limit of Osgood's semantic differential is represented by the question of sensitivity – the ability to reflect as fine distinctions in meaning as are ordinarily made. Can the semantic differential tease out nuances in meaning which are clearly felt but hard to verbalize deliberately?

In our opinion, sensory scales can represent a valid approach to understand the relationship between music and other sensorial experiences and may reveal a useful tool to investigate perceptual aspects of synesthesia and cross-modality in a low dimensional space [2]. Evaluation based on sensorial information seems to be not (or less) mediated by verbal association. In our previous experiments we compared the results obtained through the evaluation of musical excerpts by means of sensory and verbal scales [9]. One limitation is that we employed only musical excerpts taken from the classic repertoire and, only recently, we applied sensory scales to the evaluation of jazz music investigating the relationship between bebop and cool jazz [14].

A methodological risk of using experimenter-selected for rating via sensory scales is that given our nascent understanding of cross-modal responses to music, the experimenter may introduce uncontrolled biases in the selection processes, such as selecting music that contrasts in emotion, rather than because of possibly contrasting sensory experiences. In the experiment we reported here that we asked participants to freely associate sensory scales to musical excerpts without any limitation of repertoire. Participants were asked to watch, touch, lift and interact with various objects and to report spontaneously which musical pieces came to mind.

The aims of the paper are: (i) to check for each sensory item what unmediated musical characteristics were set forth; (ii) to validate the sensory scale with best generalization performance; (iii) to offer new insights in the field of cross-modal correspondences.

2 Experiment

2.1 Participants

Twenty five participants completed the study. The sample contained 17 females (68%), 5 males (20%), and 2 unspecified participants (8%). Age ranged from 18 to 48 years (*M* = 25 years, *SD* = 6.8 years). Participants were asked how many years they had played an instrument for (range 0 - 25; *M* = 8.7, *SD* = 6.4), how many years they had received training on an instrument (range 0 - 15; *M* = 7.3, *SD* = 4.9), and how many hours they listened to music in a day (range 0 - 6; *M* = 2.9, *SD* = 1.4).

2.2 **Materials**

The sensory scale used in this study (see Table 1) was developed to replicate to the extent possible the sensory scale developed at the CSC of University of Padova (see [10]). Due to Ethics limitations for the present study, one pair of items from the existing scale (Bitter-Sweet) was excluded, and a second (Cold-Hot) was modified so that it entailed two visual representations of temperature instead of cups of cold or hot water.

The objects were placed on a guide sheet (Fig. 1) which had images of each of the objects. The objects and the guide sheet were contained within a box, with the guide sheet printed on A3 paper in landscape orientation. Letters were placed on each object in discrete locations (e.g. the underside of the bottle) that matched with the items on the guide sheet, and doing so enabled participants to easily identify the sensory scale items and the correct pole orientation (i.e., which bottle equated to “C”, and which equated to “D”). After pilot testing, we added a simple instruction on the guide sheet to make clear how the participant was to interact with each sensory scale item by using the text “look”, “hold”, “touch” and so on, being careful to reduce chances of the description potentially mediating the sensory experience. For example, the expression “feel” was avoided.

Table 1. List of sensory items used in the present study. Each item was part of a matched pair, labelled with two letters.

Item labels	Sensory scale poles	Description of items	Guided interaction
A - B	Cold - Hot	Images depicting a cold (A) and a hot (B) temperature. Dimensions of each: 4.3 cm x 4.3 cm	"Look"
C - D	Light - Heavy	Two plastic bottles wrapped in black tape, with one bottle empty (C – 5 g) and the other full of liquid (D - 600 g). Due to the tape, participants were not able to visually distinguish between the two bottles	"Lift"
E - F	Hard - Soft	A cylindrical piece of wood (E), and a cylindrical piece of polystyrene foam (F). Dimensions: 16 x 3 x 3 cm; 16 x 6 x 6 cm	"Hold"
G - H	Orange - Blue	Images of the two colors (NCS notation: S 1080-Y70R [G]; S 2055-B10G [H]). Dimensions of each: 4.3 cm x 4.3 cm	"Look"
I - J	Tense - Relaxed	Two plastic, lidless boxes with a piece of wire attached across the opening (I), and a rubber band attached across the opening (J). In each case, black tape was used to cover the wire/rubber band so that participants were not able to visually distinguish between the two strips	"Press the strip"
K - L	Takete - Maluma	Images containing a computer visualization of the two visual forms Takete (K) and Maluma (L); see [4]. Dimensions: each 4.3 cm x 4.3 cm	"Look"
M - N	Smooth - Rough	Two strips of sandpaper, rated at N1200 (M) and N30 (N). Dimensions: each 15 x 10.5 cm	"Touch"

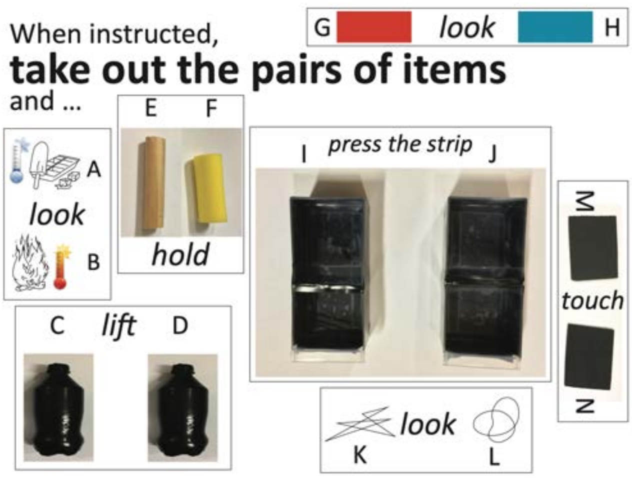

Fig. 1. Guide sheet for sensory scale (see https://www.dei.unipd.it/~canazza/ItemMapHD.pdf to download the HD photo).

2.3 **Procedure**

Participants were greeted by a lab assistant, completed an Ethics consent form, and were asked to wash their hands before entering the laboratory. They were seated at one of six workstations where they were asked to follow instructions on a computer, as presented by a survey that was created in Java programming language to randomize stimulus order and to collect responses. The room was quiet, and when the survey began participants were asked to try to work in silence, raising their hand if they had any questions. Questions were dealt with quietly and discreetly. Participants were initially presented with the following instructions via the computer survey:

> This study is about your spontaneous musical response to a variety of stimuli. You will be asked to feel, lift, look at and/or experience a number of objects with the aim to see if any piece of music was spontaneously evoked in your mind. There are no right or wrong answers, and you can report whatever music comes to mind. You do not need to justify any of your answers. We do not expect you to have any reason at all for a piece of music coming into your mind, other than it occurred when you felt or experienced the object. Please do not use your hand-held device during the experiment.

The subsequent experiment was split into two sections. For Section 1, participants were given the below instructions. These instructions were repeated for each pair of items with labels as shown in Fig. 1 (in a randomized order). In this example we use the item pair A and B:

> Take out of the box objects A and B. Feel them/look at them/experience them for a while. Imagine a piece of music that is as close to the feel/look/experience of the object A (in comparison to the object B). You may need to wait for a few moments until something comes to mind, or you may hear something that comes to mind instantly. Either way, try to be as relaxed and spontaneous as possible. Write down as much about the piece as you can, filling the following form. Leave blank if you have absolutely no idea. When you have finished, return the objects to their original location in the box and click Next/Submit.

For each item pair, response boxes were supplied for “Title”, “Section”, “Composer”, “Artist/Performer”, and “Any other general information”. Once responses had been made for each item pair, participants progressed to Section 2. In Section 2 participants were asked for further details on each of the responses that they made in Section 1, to gain a better understanding of why the pieces came to mind (if any did). For each item pair, the responses the participant had provided in Section 1 were displayed on the screen to ensure accuracy (e.g., “For stimulus A [in reference to B] you reported the following details”). Participants were then asked “Which of the following best described the process of music coming into your mind”, and selected one of the following responses:

1. No music came into my mind;
2. Music came into my mind spontaneously in response to the item and I cannot explain why;
3. Music came into my mind spontaneously in response to the item and I think I can explain why (Brief explanation);
4. Music came into my mind after some thought but I cannot explain what the thinking process was;
5. Music came into my mind, and after some thought I could explain what the thinking process was (Brief explanation);
6. If none of the above, please describe what you recall happened in your mind.

If response 3, 5, or 6 was selected, an extended response text box was provided. Following this, participants were asked “To what extent would you say the piece that came to your mind ‘felt’ like the feeling/experience of the object?” Participants responded with an 11-point scale, with 0 labelled as “Not at all”, and 10 labelled as “Completely/Perfectly”. Finally, participants were asked to enter any additional details for the piece of music they had entered in Section 1, including a Youtube link if possible. Headphones (Sennheiser HD280 Pro) were provided for participants to verify any links that they provided. This study received ethics approval (UNSW Human Ethics Approval HC190152).

3 **Results**

The *Youtube* links chosen by the subjects are available at http://www.dei.unipd.it/~roda/sensory/links.pdf. Starting from this list, a thematic analysis was carried out separately by three researchers in order to highlight those musical pieces selected by participants without the mediation of visual, semantical or autobiographical aspects. For the question concerning the process of music coming into the participant's mind (see above), responses were discarded if either no answer was selected, or the response "1. No music came into my mind" was selected. This left a remaining sample of 20 participants. As a result of this selection, 61 responses of 132 (46%) were considered by experimenters as spontaneous/unmediated. In addition, in 25% of cases participants could not explain why they made the response they did (sum of items 2 and 4), and in 56% of cases participants explicitly described the experience as being spontaneous (sum of items 2 and 3). Each participant's response represents a cross-modal association between a sensory item and a musical piece. The goal of the following analysis is to understand if the pieces associated to the same sensory item share some characteristics that can explain this association. A fully automated analysis method based on machine learning techniques was implemented. This method, already used in literature (see e.g. [13]), allows the quantitative analysis of a great amount of data, using many advanced tools developed in the last few years by the machine learning scientific community. The audio signals of the selected pieces, with the data extracted from the participants' responses, were processed according to the following pipeline: a) data augmentation was required to have an adequate number of samples for the training phase of the ML algorithm; b) for each sample a set of numerical features was calculated; c) various ML models were tested to identify which was the most effective in representing the associations between music pieces and sensory scales; d) finally, the most effective features were selected, following a Sequential Forward Selection approach.

Pre-processing: data augmentation. To increase the statistical power of the results, a preliminary phase of data augmentation was followed, as suggested in all the cases where the observations are not numerous enough in relation to the number of studied dimensions [15]. This entailed (a) addition of white Gaussian noise, (b) application of low and high pass filters, and (c) splitting of the music pieces into frames. First of all, due to the addition of zero mean and 0.05 standard deviation white Gaussian noise, the number increased going from the original set of 61 music samples to the new set of 122. Then, a filtering process, in which an optimal order high pass elliptic filter with a cut-off frequency of 2000Hz and an optimal order low pass elliptic filter with a cut-off frequency of 200Hz (both with 95% ripple in Bandpass and 5% ripple in the Bandstop) were applied. This allowed the music pieces to increase from 122 to 366. Finally, due to the splitting of each music into frames of 15 seconds, overlapping each other by 5 seconds, the number of 8376 music excerpts was reached. Frames characterized by the presence of applause, silence or other elements not belonging to the original music were deleted to avoid outliers.

Features extraction. In the feature extraction phase, the datasets relating to each pair of sensory scales were created. The 'MirFeatures' function relating to the MIR Toolbox version 1.7 [5] was applied to each 15 second length excerpt from the 8376 ones obtained as explained in the previous paragraph, obtaining the 60 features listed at http://www.dei.unipd.it/~roda/sensory/features.pdf, a set largely used in the Music Information Retrieval field [6].

Model selection. In the Model Selection phase, the performances of several machine learning algorithms, useful for the classification of the music excerpts as associated to one of the sensory scales, were assessed through the datasets coming from the experiment described in Section 2. The used classifiers were: K-Nearest Neighbor, Random Forest and Support Vector Machine (SVM). Each dataset related to a pair of sensory stimuli was divided into two complementary subsets: approximately 75% for the Training Set and 25% for the Test Set. For each of the classifiers, models with different hyperparameters were created: in the K-NN, the number of neighbours and the distance used in the metric space were changed; in the SVM, different kernels (linear, quadratic, cubic, and Gaussian) were used; in the Random Forest, two different values (30 and 100) were used as number of trees. Finally, the performances of the various models were measured for each pair of sensory scales on the remaining portion of approximately 25% of the Dataset. The accuracy of the various models in the prediction of independent observations is shown in Table 2.

Features selection. The features selection phase has been done through the use of Sequential Features Selection (SFS). As SFS nature is Wrapper type, it was necessary to select one between the previous classifiers. In agreement with Table 2, it was decided to combine the musical features selection algorithm with Linear SVM, as the one with globally better performances of classification. In fact, no pair of sensory classes was characterized by an accuracy level below 70%, obtaining the 100% for one pair, and near to 100% in the other two pairs of classes. The entire datasets were used for the application of sequential features selection. At each iteration, the algorithm added a new feature, starting from an empty set, based on the impact on the performances that the adding operation has on the calculation of the mean error of a 10-fold cross validation. At the end of the execution, all the musical characteristics prior to the one found in the elbow of the cross validation error curve were chosen. Table 3 shows the selected features for each pair of stimuli.

Table 2. Classifier performance in the Test Set.

	A-B	C-D	E-F	G-H	I-J	K-L	M-N
NN Euclidean	0.98	0.72	0.89	0.80	0.90	0.82	0.61
10-NN Euclidean	0.75	0.55	0.79	0.61	0.97	0.77	0.54
NN Cosine	0.99	0.74	0.93	0.83	0.94	0.82	0.67
10-NN Cosine	0.83	0.54	0.80	0.66	0.98	0.77	0.55
Linear SVM	1	0.72	0.99	0.73	0.99	0.75	0.72
Quadratic SVM	0.92	0.73	0.91	0.66	1	0.83	0.68
Cubic SVM	1	0.60	0.96	0.68	0.98	0.82	0.42
Gaussian SVM	0.84	0.45	0.73	0.54	0.99	0.79	0.40
RF - 30	0.98	0.62	0.90	0.48	1	0.83	0.52
RF - 100	0.99	0.58	0.85	0.54	1	0.85	0.53

Table 3. List of selected features for each pair of sensory stimuli.

	Features selected
A-B	spectral.mfcc.Mean[4] - fluctuation.peak.PeakMagMean - tonal.hcdf.Mean - spectral.mfcc.Mean[5] - spectral.skewness.Mean - spectral.mfcc.Mean[7]
C-D	tonal.hcdf.Mean - spectral.mfcc.Mean[2] - spectral.mfcc.Mean[10] - spectral.mfcc.Mean[6] - spectral.mfcc.Mean[9] - spectral.mfcc.Mean[8] - spectral.mfcc.Mean[1] - spectral.skewness.Mean
E-F	spectral.mfcc.Mean[2] - spectral.mfcc.Mean[6] - spectral.mfcc.Mean[10] - spectral.irregularity.Mean - spectral.ddmfcc.Mean[6]
G-H	tonal.keyclarity.Mean - spectral.irregularity.Mean - tonal.chromagram.peak.PeakMagMean - fluctuation.peak.PeakMagMean - spectral.centroid.Mean - spectral.mfcc.Mean[12] - spectral.mfcc.Mean[7] - spectral.mfcc.Mean[2] - spectral.ddmfcc.Mean[13] - spectral.mfcc.Mean[13] - spectral.flatness.Mean - tonal.chromagram.peak.PeakPosMean
I-J	fluctuation.peak.PeakMagMean - tonal.keyclarity.Mean - spectral.mfcc.Mean[2] - spectral.mfcc.Mean[9] - timbre.spectralflux.Mean - spectral.mfcc.Mean[7] - spectral.spectentropy.Mean - spectral.mfcc.Mean[13]
K-L	tonal.keyclarity.Mean - fluctuation.peak.PeakMagMean - spectral.centroid.Mean - spectral.mfcc.Mean[5] - tonal.chromagram.centroid.Mean - spectral.dmfcc.Mean[6] - tonal.chromagram.peak.PeakPosMean - spectral.roughness.Mean
M-N	tonal.hcdf.Mean - spectral.mfcc.Mean[2] - spectral.irregularity.Mean - timbre.spectralflux.Mean - timbre.zerocross.Mean

4 Conclusions

A perceptual experiment was carried out to study cross-associations between music and other sensory modalities, such as touch and vision. The experiment used a recently developed sensory scale that does not use any explicit verbal description. Participants' responses were analyzed following an approach based on machine learning techniques. Results show that algorithms trained on the experimental data are able to predict, with an accuracy greater than 70% (see Table 2), the associations between music and other sensory stimuli, showing that such associations can be explained, at least in part, by a set of quantitative features directly extracted by the music excerpts. In particular, according to Table 3, Mel Frequency Cepstral Coefficients (MFCC), which are a set of features related to the spectral envelope, are involved in all the classification tasks, implying a relevant effect of timbre in mediating the cross-modal associations with music. Moreover, the Harmonic Change Detection Function (HCDF), related to more or less rapid changes of the tonal harmony, appears to be involved in the association with Cold-Hot, Light-Heavy, and Smooth-Rough; whereas Key Clarity is involved in Orange-Blue, Tense-

Relaxed, and Takete-Maluma. Future work will include the comparison with other verbal and non verbal scales and an analysis of the influence of factors such as musical training and personality[1].

References

1. De Poli, G., Canazza, S., Rodà, A., Schubert, E.: The role of individual di erence in judging expressiveness of computer-assisted music performances by experts. ACM Transactions on Applied Perception 11(4) (2014)
2. De Poli, G., Murari, M., Canazza, S., Rodà, A., Schubert, E.: Beyond emotion: Multi-sensory responses to musical expression. In: The Routledge Companion to Embodied Music Interaction, pp. 78-86. Routledge (2017)
3. Kinoshita, Y., Masaki, Y., Muto, T., Ozawa, K., Ise, T.: Scenery based kansei music selection for car audio systems. In: 2009 IEEE 13th International Symposium on Consumer Electronics. pp. 94-98. IEEE (2009)
4. Kohler, W.: Gestalt psychology. Liveright, New York, 2nd ed. 1947 edn. (1929)
5. Lartillot, O., Toiviainen, P.: A matlab toolbox for musical feature extraction from audio. In: International conference on digital audio e ects. pp. 237-244. Bordeaux (2007)
6. Moffat, D., Ronan, D., Reiss, J.D.: An evaluation of audio feature extraction toolboxes (2015)
7. Murari, M., Rodà, A., Da Pos, O., Canazza, S., De Poli, G., Sandri, M.: How blue is Mozart? Non verbal sensory scales for describing music qualities. In: Proceedings of the 11th Sound and Music Computing Conference, SMC 2014. pp. 209-216 (2014)
8. Murari, M., Schubert, E., Rodà, A., Da Pos, O., De Poli, G.: How >:(is Bizet? Icon ratings of music. Psychology of Music 46(5), 749-760 (2018)
9. Murari, M., Rodà, A., Canazza, S., De Poli, G., Da Pos, O.: Is vivaldi smooth and takete? non-verbal sensory scales for describing music qualities. Journal of New Music Research 44(4), 359{372 (2015)
10. Murari, M., Rodà, A., Da Pos, O., Schubert, E., Canazza, S., De Poli, G.: Mozart is still blue: A comparison of sensory and verbal scales to describe qualities in music. In: The 12th Sound and Music Computing Conference. pp. 351-358. Maynooth University Maynooth, Ireland (2015)
11. Osgood, C.E.: Semantic differential technique in the comparative study of cultures. American Anthropologist 66(3), 171-200 (1964)
12. Parise, C., Spence, C.: Audiovisual cross-modal correspondences in the general population. The Oxford handbook of synaesthesia 790-815 (2013)
13. Rodà, A., Canazza, S., De Poli, G.: Clustering a ective qualities of classical music: beyond the valence-arousal plane. IEEE Transactions on A ective Computing 5(4), 364-376 (2014)
14. Rodà, A., Da Lio, E., Murari, M., Canazza, S.: How cool is bebop jazz: Clustering and decoding of jazz music. In: KEER2018, Go Green with Emotion. 7th International Conference on Kansei Engineering & Emotion Research 2018, 19-22 March, 2018, Kuching, Malaysia. pp. 298-308. No. 146, Linkoping University Electronic Press (2018)
15. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmenta-tion for environmental sound classi cation. IEEE Signal Processing Letters 24(3), 279-283 (2017)
16. Schloss, K.B., Lawler, P., Palmer, S.E.: The color of music. Stress 2(r2), 0{95 (2008)
17. Schubert, E., Murari, M., Rodà, A., Canazza, S., Da Pos, O., De Poli, G.: Verbal and Cross-Modal Ratings of Music: Validation and Application of an Icon-Based Rating Scale. i-Perception 10(3) (2019)
18. Shelley, P.B.: Posthumous poems. John and Henry L. Hunt (1824)

19. Spence, C.: Crossmodal correspondences: A tutorial review. Attention, Perception, & Psychophysics 73(4), 971-995 (2011)
20. Sugihara, T., Morimoto, K., Kurokawa, T.: An improved kansei-based music retrieval system with a new distance in a kansei space. In: RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759). pp. 141-146. IEEE (2004)
21. Turchet, L., Rodà, A.: Emotion Rendering in Auditory Simulations of Imagined Walking Styles. IEEE Transactions on A ective Computing 8(2), 241-253 (2017)
22. Turchet, L., Zanotto, D., Minto, S., Rodà, A., Agrawal, S.: Emotion Rendering in Plantar Vibro-Tactile Simulations of Imagined Walking Styles. IEEE Transactions on A ective Computing 8(3), 340-354 (2017)
23. Ward, J., Mattingley, J.B.: Synaesthesia: an overview of contemporary ndings and controversies. Cortex 42(2), 129-136 (2006)