From music to dance: the inheritance of semantic inferences

Pritty Patel-Grosz[1], Jonah Katz[2], Patrick Georg Grosz[1],
Tejaswinee Kelkar[3] and Alexander Refsum Jensenius[1]
[1]University of Oslo, [2]West Virginia University, [3]Universal Music Norway AS

1 Overview

Music can give rise to abstract semantic inferences about music-external situations. We ask whether dance (which we define as music-accompanying body movement) also gives rise to similar abstract semantic inferences. We experimentally test whether inferences from a given musical sequence are inherited by body movement triggered by this musical sequence. Our results indicate that such an inheritance of semantic inferences may occur.

2 Theoretical underpinnings – from music to dance semantics

Recent research in formal semantics argues that music can give rise to inferences about music-external objects (so-called *virtual sources*), which allow listeners to infer descriptive or narrative meaning, Schlenker (2017, 2019a, to appear). A typical example discussed in Schlenker (2019a:52) is found in Saint–Saëns's *Carnival of the Animals*, where a low-pitched melody is mapped onto a large object, namely an elephant. By contrast, a high-pitched melody would not give rise to the inference that there is a large object in the narrative; high pitch can be mapped onto a small object instead – for example, a mouse. Such inferences are iconic in that the denotation of the meaning-bearing object – in this case the music – operates on its form. To illustrate, we apply Greenberg's (2021) formalism to the above example and posit the iconic semantics in (1) for object-denoting pitch.

For Greenberg, an iconic semantics is defined such that the form of the sign, symbolized by the bold-typed $\boldsymbol{M}$ in (1), also occurs in its denotation. When we interpret a piece of music M with regards to a narrative situation s, we can draw an inference that the pitch of M is inversely mapped onto the size of a salient object in s. The higher the pitch, the smaller the object. This can be implemented by multiplying the pitch of M with a contextually given constant k (which is smaller than 1, in order to derive the inverse mapping of pitch and size). When an inference of this type is met for a given narrative, then we can say that [[M]] is true in s (or [[M]] is satisfied by s). This means that a low-pitched melody is true of a narrative situation in which we are dealing with a large object, and false of a narrative situation in which we are dealing with a small object. Such inferences are by their very nature abstract, i.e. it does not matter whether the object is an elephant, a landscape, or, more abstractly, *a magnificent idea*.

(1) For a piece of music $\boldsymbol{M}$ and a constant k ($k < 1$) in a narrative situation s,
$[[\boldsymbol{M}]]$ is satisfied by s only if size($\iota x.x$ is an object in s) = k * pitch($\boldsymbol{M}$)

Crucially, the properties of music that give rise to such iconic inferences (pitch, loudness, speed, silence, dissonance, change of key; see Schlenker 2019b:433-436) have counterparts in music-accompanying movement, for example dance. An observation from choreomusicology suggests that musical pitch corresponds to the direction of gestures in space in body movement (Mason 2012:10); see Kelkar & Jensenius (2018) for critical discussion. We take such correspondences between music and body movement as our point of departure and present an experimental study that addresses the following questions: (i.) do abstract body movements (e.g. dancing/moving spontaneously to a piece of music) give rise to semantic inferences comparable to the inference in (1)? (ii.) are there parallels between the inferences that we draw from hearing music, and the inferences that we draw from seeing abstract body movement? (iii.) if we perceive body movement *D* that was initially performed as an interpretation of a short musical sequence *M*, is there a correspondence between our inferences from *D* and our inferences from *M*?

The experiment tests the hypothesis that body movements *D* which are performed in response to a musical sequence *M* 'inherit' properties of *M*, thus giving rise to the same or similar semantic inferences. As a concrete illustration of such inheritance, we map the musical meaning inference in (1) to a body-movement-related meaning inference as given in (2). This assumes, for purpose of illustration, that *pitch* is inversely mapped onto the *height of the hands* of a person who is moving to the music. While higher pitch was mapped to smaller size, higher gestures in body movement are intuitively mapped to larger size.

(2) For a body movement ***D*** and a constant *k* (*k* > 1) in a narrative situation *s*,
[[***D***]] is satisfied by *s* only if size(ι*x*.*x* is an object in *s*) = *k* * height(hands(***D***))

3 Experimental design

We depart from toy inferences of the type in (1) and (2). Instead, we use six combinations of short musical sequences (between 1.45 seconds and 5.0 seconds in length) and motion capture renderings of movements carried out to accompany those sequences by participants in the study of Kelkar & Jensenius (2018), where participants were asked to trace the music that they heard with their hands. Illustrations of the motion capture renderings are given in Figure 1. The complete set of stimuli can be found in the following folder: https://www.dropbox.com/sh/do0p22rs85kusr0/AABt1SPmbZLtMpbg8REGP1Pha?dl=0

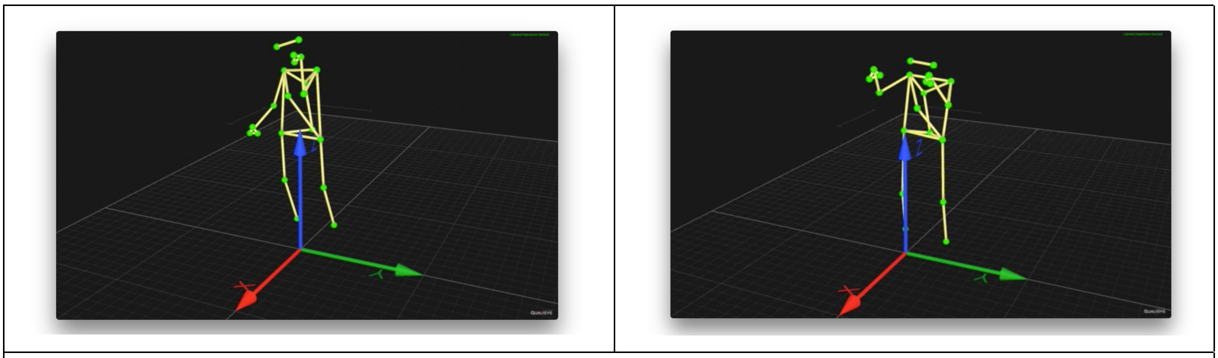

Figure 1: stills taken from the videos used as visual stimuli for the experiment

All participants listened to the six sound files and separately watched the six silent videos; they did not watch combinations of videos and sound files. Stimuli were clustered in that participants were either presented with all audio stimuli before all video stimuli, or the other way around. Participants were prompted to rate on a slider scale from 0 to 100 how well the sound file / silent video expressed one of the following emotions: *Angry*, *Bored*, *Calm* and *Excited* – with 1 trial for each emotion (2x6x4 trials in total). These emotion terms are based on the four quadrants of Russell's (1980) circumplex model of emotion, where *Angry* is [Valence: negative, Arousal: positive], *Excited* is [Valence: positive, Arousal: positive], *Calm* is [Valence: positive, Arousal: negative], and *Bored* is [Valence: negative, Arousal: negative]. We used emotion terms as opposed to concrete properties such as size (cf. the toy example in (1)-(2)), to avoid participants directly interpreting properties of the music or movement; furthermore, there is a precedent of probing emotive meanings in music and movement in the findings of Sievers et al. (2013).

Participants were recruited via announcements on social media and various online fora devoted to music, dance, and linguistics. The experiment was carried out online in the PCIbex environment (Zehr & Schwarz 2018). Before the experiment, participants filled out a questionnaire on their demographic, linguistic, and musical background. The instructions for the experiment asked participants to use a slider to indicate how well they thought the given sound or video expressed the given emotion. Participants were able to play the sounds and videos as many times as they desired. The experiment took about 15 minutes to complete. Both native and non-native speakers of English participated in the study; only participants who reported being native speakers of English are analysed here, since emotion words were provided in English, and cross-linguistic variation cannot be excluded.

Our experiment tests several hypotheses related to (i-iii) above. In particular, we examine whether: (a.) participants draw consistent inferences about particular stimuli, i.e. if stimuli with high ratings for *Angry* received low ratings for *Calm*, and so forth. (b.) some of the information that auditory stimuli convey can be recovered from movement stimuli that were created as a response to those sounds. Positive answers to these questions would support the idea that music and music-accompanying movement encode descriptive information in comparable ways, i.e. that participants draw the same types of inferences about musical and movement stimuli.

4 Results

The first question we examined is whether listeners respond to audio and video stimuli in a broadly comparable way, bearing on (i) and (ii) above. Table 1 shows the mean responses and standard deviations to each of the four emotion descriptors for audio and video files. Overall rating levels are similar for the two modalities, as are the relative patterns amongst descriptors. Participants exhibit a tendency to assign higher scores for high-arousal descriptors (*angry, excited*) than low-arousal ones (*bored, calm*). There are no gross differences between the two modalities here, suggesting that participants are as likely to infer emotional content from movement as they are from music.

		Audio	Video
Angry	Mean	37	47
	SD	32	34
Bored	Mean	36	32
	SD	35	30
Calm	Mean	25	29
	SD	30	29
Excited	Mean	50	50
	SD	37	31

Table 1. Mean slider ratings and standard deviations for audio and video stimuli on each of the four descriptors in the study.

Next, we ask if individual audio and visual stimuli are subject to consistent inferences from participants. Figure 2 shows two attempts to validate the response space.

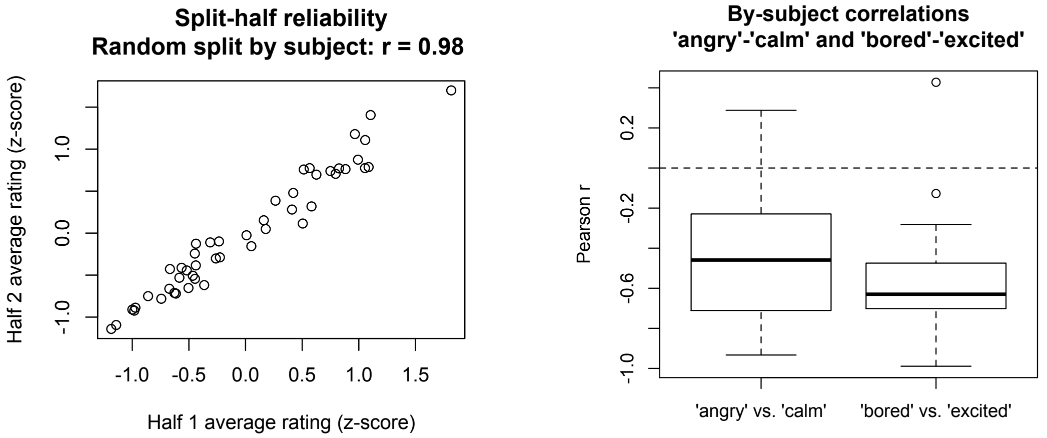

Figure 2. Left: correlation between two randomly-selected halves of the participant pool on the slider scores assigned to each combination of stimulus and descriptor. Right: Correlations between 'opposite' descriptors, computed across all stimuli within each participant.

The left plot in Figure 2 tests a form of split-half reliability, where the thing being split into random halves is the participant pool. The question is whether, for each stimulus, when a randomly-selected half of participants infer high levels of some descriptor from that stimulus, do the other half do the same? The answer is an emphatic yes (r = 0.98), showing that participants broadly agree on how much the terms *angry*, *calm, excited*, and *bored* are associated with particular stimuli. The right plot in Figure 2 summarizes within-subject correlations between 'opposite' descriptors. Our assumed theory (Russell 1980) situates the four descriptors in terms of a two-dimensional space of valence and arousal. If this is valid, we expect strong negative correlations between descriptors differing in both valence and arousal. As shown in the right plot, correlations are almost uniformly negative, some of them quite strongly so. This indicates, e.g., that when a participant infers high *angry* content from a particular stimulus, they are likely to infer low *calm* content from that stimulus.

Finally, we ask whether the inferences participants draw from video stimuli tend to resemble the inferences they draw from the sound stimuli that inspired the motion in the video. That is, do participants implicitly recover information from motion about the sound that the motion was intended to accompany? Figure 3 shows audio-visual correlations, treating each combination of stimulus and descriptor as a separate observation. The observed correlation suggests that motion can be used to encode and decode information from an auditory stimulus.

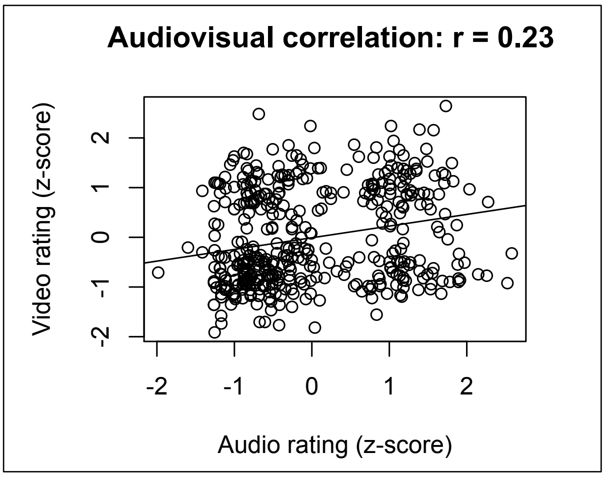

Figure 3. Correlation between slider scores for auditory stimuli and for the video stimuli that were created in response to them. Line shows general linear model of the relationship.

To examine whether these audio-visual correlations are robust across stimuli, descriptors, and subjects, we fit a linear mixed-effects regression model using the lme4 package in R (Bates *et al.* 2015). The dependent variable was the slider-score for video stimuli, using the slider score for the corresponding audio stimulus as a predictor. The model also included fixed effects of the order in which the two tasks were performed (audio first vs. video first), as well as its interaction with audio scores. The model included random intercepts for item, subject, and descriptor. We tested random slopes for model improvement using the likelihood-ratio test; only the by-item random slope of audio score was retained. All slider scores were centred around the midpoint of the scale, to aid interpretation of fixed effects. The significance of fixed effects was gauged by dropping parameters and using the likelihood-ratio test.

In the audio-first order, audio score was a significant (positive) predictor of video score: $\beta = 0.29$, $\chi^2 = 4.73$, $p = 0.030$. Video scores were somewhat lower when the video condition was completed first: $\beta = -13.7$, $\chi^2 = 11.6$, $p < 0.001$. And the correlation between video and audio scores was substantially lower when the video condition was completed first: $\beta = -0.19$, $\chi^2 = 5.33$, $p = 0.021$. It appears, then, that participants draw inferences from videos of movements that mirror inferences from the auditory stimuli that inspired those movements, but they do so much more reliably when the auditory stimuli are presented first than when the video stimuli are.

5 Discussion

Our results suggest that inferences from music and inferences from body movement are coherent, consistent, and mutually informative. This is in line with a view where (i.) body movement gives rise to similar inferences to what we find in music, (ii.) there are parallels between the inferences from music and the inferences from body movement, and (iii.) listeners can recover information about inferences from music just from viewing body movement based on the music.

The finding that correlations are more robust when the auditory condition occurs before the visual condition was not expected. We had anticipated that there might be some effect of order, but had no particular hypothesis about what that would be. A post-hoc hypothesis that might explain this finding involves the fact that, according to Schlenker's (2017, 2019a) theory, musical stimuli license inferences on the physical movement of virtual sources / objects (among other things). The inferred semantics of the auditory stimuli, when presented first, could activate various kinds of movement schemata; that would facilitate further processing of actual visual representations of movement. Because the motion-capture videos are straightforward representations of people moving, the effect of order could reflect such a facilitation in the auditory-first condition. On the other hand, there is no reason to think that viewing movements activates musical or auditory schemata, so the auditory condition would not benefit from this facilitation after viewing movements. This fundamental asymmetry, if replicated in future work, could thus be seen as support for Schlenker's hypothesis that musical stimuli are interpreted in terms of physical, spatial movements.

References

Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**(1). 1-48.

Greenberg, Gabriel. 2021. The Iconic-Symbolic Spectrum. Manuscript, UCLA. https://ling.auf.net/lingbuzz/005787

Kelkar, Tejaswinee, and Alexander R. Jensenius. 2018. Analyzing Free-Hand Sound-Tracings of Melodic Phrases. *Applied Sciences* 8, 135. https://doi.org/10.3390/app8010135

Mason, Paul H. 2012. Music, dance and the total art work: choreomusicology in theory and practice. *Research in Dance Education* 13, 5-24. https://doi.org/10.1080/14647893.2011.651116

Russell, James A. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psyschology* 39, 1161–1178. https://doi.org/10.1037/h0077714

Schlenker, Philippe. 2017. Outline of Music Semantics. *Music Perception* 35, 3–37. https://doi.org/10.1525/mp.2017.35.1.3

Schlenker, Philippe. 2019a. Prolegomena to Music Semantics. *Review of Philosophy & Psychology* 10, 35–111. https://doi.org/10.1007/s13164-018-0384-5

Schlenker, Philippe. 2019b. What is Super Semantics? *Philosophical Perspective* 32, 365-453. https://doi.org/10.1111/phpe.12122

Schlenker, Philippe. To appear. Musical Meaning within Super Semantics. *Linguistics & Philosophy*. https://ling.auf.net/lingbuzz/004937

Sievers, Beau, Larry Polansky, Michael Casey, and Thalia Wheatley. 2013. Music and movement share a dynamic structure that supports universal expressions of emotion. *Proceedings of the National Academy of Sciences* 110, 70–75. https://doi.org/10.1073/pnas.1209023110

Zehr, Jeremy, & Florian Schwarz. 2018. PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832