Audio Engineering Society

Convention Paper

Presented at the 150th Convention
2021 May 25–28, Online

Implications of crossmodal effects and spatial cognition on producing in spatial audio

Thomas Görne[1], Kristin Kuldkepp[1], and Stefan Troschka[1]

[1]*Immersive Audio Lab, Hamburg University of Applied Sciences, Faculty of Design, Media and Information, Hamburg, Germany*

Correspondence should be addressed to Thomas Görne (thomas.goerne@haw-hamburg.de)

ABSTRACT

It is quite common to use spatial language in the description of the sensation of sound: A sound can be big or small, it can be edgy, flat or round, a tone can be high or low, a melody rising or falling – all these linguistic metaphors are apparently emerging from the crossmodal correspondences of perception. An auditory object can have a metaphorical size, shape and position in space besides its (perceived) physical size, shape and position in space. The present paper reviews research on crossmodal effects and related findings from different disciplines that might shine a light on the production and aesthetics of spatial audio. In addition, some preliminary results of experiments with complex spatial sonic structures are presented.

1 Introduction

In audio engineering, the spatial representation of sound is often regarded a solely technical issue. But then it is quite common to use spatial language in the description of the sensation of sound: A sound can be big or small, it can be edgy, flat or round, a tone can be high or low, a melody rising or falling. All these linguistic metaphors apparently are expressions of the crossmodal correspondences of perception, the linking between typically an auditory perception like pitch or intensity of a sound and a visual perception like vertical position, size or brightness of a physical object. An auditory object thus can have a metaphorical size, shape and position in space besides its (perceived) physical size, shape and position in space. And obviously, physical and metaphorical features of a spatial audio production can be more or less congruent.

The crossmodal correspondences can be related to embodied cognition and to the way the mind conceptualizes not only the physical environment but also abstract ideas in spatial imagery. A specific focus here is set on the height scheme, providing the spatial framework for some very basic conceptual metaphors of the mind, like "happy is up, sad is down".

Some questions arise from these ideas:

- Can a listener distinguish between the metaphorical and physical properties of a sound?
- Can the metaphorical properties of sound be regarded as universal?

- Do spatial cognition and the conceptual spatial metaphors of the mind have an influence on the emotional impact of actual spatial positions of auditory objects, of their spatial relations and spatial trajectories?

2 Spatial metaphors of sound

2.1 Early findings of philosophy, psychology and linguistics

In 1883, the German philosopher Carl Stumpf stated that we express the sensation of tone *"with a certain psychological necessity"* in spatial metaphors, most evident in the "height" of a tone: *"The power of spatial imagery of tones is indeed remarkable."*[1] [1: 189]. In the following he discusses other metaphorical descriptions, stating that *"we ascribe in general a ›dull, dark‹ emotional character to the low tones, a ›sharp, bright‹ emotional character to the high tones."*[2] [1: 203] And finally he notes the interdependence of pitch and the perceived "size" or "volume" of the auditory object: *"In the imagination, the low tones have a larger circumference."*[3] [1: 207].

Visual metaphors of sound identifed by Stumpf thus are height, size and brightness, the first two being spatial metaphors, besides other descriptions like sharpness.

In the early 20th century those intersensory metaphors attracted attention in the new field of gestalt psychology. Wolfgang Köhler presented in 1929 the famous "Maluma/Takete" experiment on the metaphorical shape of sound: The subjects were asked to assign the meaningless words "Maluma" and "Takete" to abstract figures, one rather round, the other rather edgy – the answers have been very consistent. This experiment has since then been successfully reproduced numerous times, the phenomenon nowadays often referred to as "Bouba/Kiki effect" [2: 224f], [3].

The idea of the shape of a sound comes naturally: students asked to draw sounds they are presented with often come to surprisingly similar results (Fig. 1).

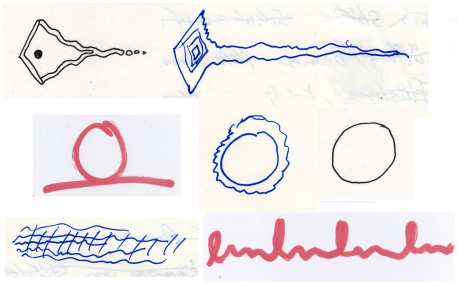

Fig. 1: Scribbles from communication design students asked to draw various sounds, top: small gong hit with a mallet, middle: large wooden tea box hit with the hand, bottom: spinning sound of a fishing reel

Parallel with Köhler but independent of his work, and probably unaware of Stumpf, the linguist Edward Sapir studied the phenomenon of "phonetic symbolism" in English language, namely the perceived "size" of vowels, finding that /a/ and /o/ are perceived "larger" as /e/ and /i/ [4].

Two years earlier, Diedrich Westermann – investigating "sound pictures" (Lautbilder) or *ideophones* in the West African languages Kpelle and Nyangbo – noted that in these languages exist pairs of similar words, composed with either "dark" vowels like /u/ and /o/ or "bright" vowels like /e/ and /i/, that refer to similar objects but of different size, thickness or brightness [5].

Today, in the context of linguistics it is common to describe the vowel /i/ as "high, bright, small" in contrast to the "low, dark, large" /o/ or /u/ (see e.g. Jakobson and Waugh [6])[4]. This inherent meaning of certain vocal sounds appears to be expressed specifically in a language's ideophones, rendering words metaphorically meaningful[5]. The most remarkable finding here is that a consistent "sound symbolism" or "phonetic symbolism" can be found across cultures.

Stumpf's thoughts on language, Köhler's findings of a "gestalt" of the auditory object, the linguist's findings

[1] *"Die Kraft der Raumsymbolik bei Tönen ist in der That auffallend."*

[2] *"So schreiben wir den tiefen Tönen im Allgemeinen einen ›dumpfen, dunklen‹, den hohen Tönen einen ›scharfen, hellen‹ Gefühlscharakter zu."*

[3] *"Den tiefen Tönen kommt in der Vorstellung eine grössere Ausdehnung zu."*

[4] This "sound symbolism" or "phonetic symbolism" is consistent with the "pitch/height", "pitch/size" and "pitch/brightness" metaphors identified by Stumpf, as the frequency of the second formant of the /i/ is roughly one octave above that for the /o/ and /u/.

[5] One of numerous examples in English is the naming of fictional characters according to their appearance, e.g. "Pooh" and "Piglet" in A.A. Milne's wonderful children's stories.

on phonetic symbolism all can be summed up as *crossmodal metaphors*, as the description of the sensation of sound in the terminology of other sensory modalities like vision or touch. Crossmodal metaphors render sound meaningful with visual or haptic properties of a percept that is imagined as a *thing*, as a bounded and tangible physical object in space.

Lakoff and Johnson note in their seminal work on the metaphorical structure of thought and reasoning: *"Once we can identify our experiences as entities or substances, we can refer to them, categorize them, group them, and quantify them – and, by this means, reason about them."* [7: 25]

Fig. 2: David Gibson, The Art of Mixing (Youtube video tutorial)

In fact, the thingness of the auditory object with its crossmodal properties is so inevitable that it not only works as a seemingly natural visualization of sound in mixing tutorials (Fig. 2) but even is used for objective descriptions in the audio engineering world, just one example being DELTA's "sound wheel", designed as a tool for loudspeaker assessments with terms like "brilliance" or "bass depth" [8].

2.2 A cross-cultural view

A brief look into different languages, including non-Indo-European languages, supports the assumption that the "height" of a sound is a widespread linguistic metaphor of pitch. According to a short informal sample taken by the authors, musical pitch is known to be "high" or "low" in languages as different as Arabic (Lebanon), Hebrew, Bulgarian, Czech, Polish, Russian, Danish, Dutch, German, Welsh, Latvian, Estonian, French, Italian, Portuguese, Romanian, Spanish, Armenian, Burmese (Myanmar), Chinese (Hunan), Nepalese, Japanese, Vietnamese, and Indonesian[6].

Less frequently found metaphorical descriptions of pitch are the *thickness* of sound, e.g. in Armenian, Russian, Farsi*, Kichwa (Ecuador), Latin American Spanish (e.g. Colombia, Ecuador), Turkish*, and Zapotec* (Mexico), and the related metaphor of *size*, e.g. in Bashi† (Congo), Basongye† (Congo), Jabo† (Liberia), Kpelle‡ (Liberia), and again in Armenian and Zapotec* (Mexico)[7].

Stumpf already identified width or thickness as a metaphor of pitch as well as the *sharp-heavy* opposite pair in ancient Greek and Roman cultures [1], the latter still common in the Romance languages (agudo-grave / acuto-grave / aigu-grave). Similarly, *sharpness* is used e.g. in Bhojpuri (India/Bihar), Chinese (Hunan), Farsi* and Alemannic (Germany)[8]. In some languages even *brightness* is a metaphor of pitch, e.g. in Danish, Swedish, Alemannic (Germany) and Latin American Spanish (e.g. Colombia, Ecuador)[9].

Quite often subjects mention multiple metaphors of pitch. It seems likely that similar crossmodal meatphors are used throughout cultures, referring to similar (but probably not identical) perceptual features like pitch or timbre – hence the "bright" cymbal, the "sharp" snare, the "fat" bassdrum. And of course, the perceptual boundary between pitch and timbre itself appears to be fuzzy, at least in the description of non-tonal sounds.

2.3 Crossmodal correspondences

The linguistic crossmodal metaphors obviously are expressions of the *crossmodal correspondences*[10] of perception, connections between the different sensory modalities – mainly visual and auditory –, not to be confused with the likewise individual and rare phenomenon of synaesthesia. Crossmodal correspondences are a mechanism that supports orienting in a complex environment by organizing multisensory stimuli referring

[6] Data by the authors.

[7] Data by the authors, except *after Shayan et al. [9], †after Merriam [10: 96f, 118], ‡after Stone [11].

[8] Data by the authors, except *Shayan et al. [9].

[9] Data by the authors.

[10] Terminology according to Spence [12]. Also referred to as intersense modality, cross-modality reference, synaesthetic cross-modality correspondence, inter-modal association, cross-modal equivalence, cross-modal similarity, cross-modal mapping, or cross-domain mapping.

to the same objects, in addition to semantic and spatiotemporal factors.

They appear to be based mainly either on innate neural connections, or on infant development following the most likely properties of the physical environment [12]. There might also exist "semantically mediated" correspondences, assuming that arbitrarily the same words are used for different sensory modalities, but as such connections would be of lesser interest for the topic they shall be skipped here.

Crossmodal correspondences have been investigated mainly since the 1980's and then extensively since the 2010's [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24]. They are examined typically through "speeded discrimination" or "speeded choice reaction" tests, where the subject's reaction time to multisensory stimuli is measured. It appears that with congruent stimuli – e.g. a figure with a high position on the screen combined with a high-pitched sound on the headphones – the task (e.g. identifying the position on the screen) is performed significantly faster in comparison to the presentation of incongruent stimuli, e.g. a figure with a high position on the screen combined with a low-pitched sound on the headphones.

Correspondences have been examined and verified between pitch or spectral weight on the one hand and spatial height, brightness, shape (rather edgy or rather round, cf. "Bouba/Kiki"), size (cf. phonetic symbolism), surface structure (in form of spatial frequency), and vertical direction of movement on the other hand, as well as between sound intensity and brightness.

Summing up the research on crossmodal correspondences, a sound with dominant high frequencies matches with a small, bright, edgy object with a position high in space. A sound with dominant low frequencies matches with a large, dark, round object with a position low in space. A sound with increasing frequency represents an upwards directed movement and vice versa – according to Stumpf, Berlioz was mocking a fellow composer who utilized this as figurative spatial movement [1: 191].

In recent years there has been a discussion if specifically the pitch/height mapping can be considered universal – an important question bearing in mind the relevance of the height metaphor for spatial audio. Spence assumed it to be semantically mediated and therewith language dependent [12: 978ff]. Since then, more research discussed the topic controversially, focusing specifically on preverbal infants [20, 23, 25]. However, the matter can be regarded as settled: In 2018, Walker et al. provided evidence of pitch/height mapping in newborn children [26], and in 2019 the pitch/height correspondence was even found in dogs [27].

Notably, not all crossmodal correspondences of perception find their expression in the linguistic metaphors of a particular language. And conversely, language might affect the spatial conceptualization of auditory perception [28, 29]. Dolscheid et al. state: *"People who use different linguistic space-pitch metaphors also think about pitch differently."* [28]

Crossmodal correspondences not only facilitate the construction of audiovisual objects from auditory and visual percepts, they can basically affect auditory perception. Metaphorical properties and perceived physical properties of the auditory object can be indistinguishable.

The interdependence of pitch and perceived height of the auditory object was first described in 1930 by Caroll Pratt ("pitch height effect" or "Pratt's effect"). Under the impression of the works of Stumpf and Köhler, and specifically discussing Stumpf in-depth, Pratt investigated this most obvious intersensory connection. In his experiment, Pratt set up a test with pure tones in octave relations. He concluded: *"High tones are phenomenologically higher in space than low ones."* [30]

Pratt's findings have later been reproduced repeatedly, both with tone bursts and band-limited noise [31, 32]. Likewise, it has been shown that the perceived size of the auditory object depends of pitch [33, 34], as it depends of e.g. spatially distributed decorrelated signal content known to induce the "apparent source width".

A different effect of the crossmodal correspondence was recently investigated, showing that the judgment of musical pitch is biased by vertical movement of the listener [35] – a direct expression of embodied cognition (see below, section 3.3).

3 Spatial cognition

3.1 Body-related perception of space

Olson and Bialystok quote Darwin with the question: *"Has the oyster a necessary notion of space?"*[36] –

presumably not, as space is experienced through movement and interaction with the environment, and as mental representations of the spatial environment at least for humans are grounded in bodily experience [36, 37].

Olson and Bialystok argue that spatial relations and spatial trajectories are not conceptualized in a uniform and homogeneous three-dimensional space, but that spatial cognition is based on three basically different body related Cartesian axes front/back, left/right, and up/down or high/low. Spatial relations and movements along these axes are easier to perceive and to conceptualize than other orientations [36]. The vertical axis appears specifically meaningful due to gravity [36, 37].

More complex spatial relations, namely the orientation along oblique axes, are not as easy to perceive. The recognition of oblique structures is more time-consuming as of vertical or horizontal structures [36: 182ff, 205ff]. The media theorist Rudolf Arnheim, pupil of Köhler, points out the *"dynamics of obliqueness"* in visual perception: *"obliqueness always [...] is seen as a gradually increasing deviation from, or approximation of, the stable positions of the vertical and horizontal"* [38: 88].

3.2 Conceptual metaphors referring to space and movement

> *"I've looked at clouds from both sides now / From up and down, and still somehow / It's cloud illusions I recall / I really don't know clouds at all"* (Joni Mitchell)

Spatial thinking is not only defining the way we perceive physical objects in our environment, their spatial relations and movements. In order to be able to reason about abstract concepts (and of course about immaterial physical phenomena like sound), we cannot but imagine them similarly as things in space, subjected to physical forces. Barbara Tversky formulates as one of the "Laws of Cognition": *"Spatial thinking is the foundation of abstract thought."* [37]

The metaphorical nature of cognition is verbalized in expressions like "coming up in the weeks ahead", "she's at the peak of her career", "that was a low trick", "he is in a dark mood", "that warmed my spirits", "they're very close", "he was overcome by emotion", "she was moved" [7, 39].

Among the dominant conceptual metaphors, Lakoff and Johnson specify the *orientational metaphors* [7: 14ff], spatial metaphors recurring to verticality, like

- happy is up; sad is down,
- conscious is up; unconscious is down,
- health and life are up; sickness and death are down,
- having control or force is up; being subject to control or force is down,
- good is up; bad is down,
- rational is up; emotional is down

It is striking that among all possible spatial orientations verticality appears to be unique. An explanation is the experiential basis of the verticality metaphor, for the posture of the body resembles the mood or emotional state of a person as well as health or illness [7]. Casasanto and Bottini further explicate: *"Across cultures, people spontaneously elevate the chest or raise the arms above the head to express pride, and hang the head or slump the shoulders to express shame. Accordingly, upward- and downward-directed bodily actions can influence the retrieval of emotional memories."* [40: 140]

The front/back axis denotes the direction of sight and of movement, and is used e.g. for one-dimensional spatial conceptualizations of time, where time is considered as movement on a line in space, following either the "moving ego" ("we're marching through time") or the "moving time" ("time marches on") metaphor [37: 163ff]: "the past is behind us", "there are obstacles in the way", "the best is yet to come".

The left/right axis appears least meaningful among the body related spatial axes, as the meaning of the lateral position is not only dependent of handedness, but also of reading and writing habits, constituting a cultural code [36, 37]. Furthermore, left and right are reversed in gestural communication, as speakers gesture from their own point of view [37]. And besides handedness and cultural habits, left and right are basically interchangeable due to the bi-lateral symmetry of the body.

Casasanto exemplifies the dominant left/right metaphor: *"Across languages and cultures, good things are often associated with the right side of space and bad things with the left. This association is evident in positive and negative idioms like my "right-hand" man and "two left feet", and in the meanings of English words derived from the Latin for "right" (dexter) and "left" (sinister)"* [41: 110-111]. The obvious experiential basis for this conceptual metaphor is the person's handedness – in spatial preference experiments right- and left-handers respond differently [41].

3.3 Embodiment and space in music cognition

The philosophers Robert A. Wilson and Lucia Foglia define the *Embodiment Thesis*: *"Many features of cognition are embodied in that they are deeply dependent upon characteristics of the physical body of an agent, such that the agent's beyond-the-brain body plays a significant causal role, or a physically constitutive role, in that agent's cognitive processing"*, and they introduce the main principle of embodied cognitive science: *"Without the involvement of the body in both sensing and acting, thoughts would be empty, and mental affairs would not exhibit the characteristics and properties they do."* [42]

Different theories and findings from neuro- and cognitive science advocate the body's influential role in cognition, perception and action, one being the famous mirror neuron theory [43, 44, 45, 46]. Following this conception, the mirror system produces knowledge of an action, even when not performed but merely observed: *"We understand an action because the motor representation of that action is activated in our brain"* [44: 661]. Based on the mirror system research, Gentilucci and Corballis propose the evolution of language from manual gestures [47]. This theory explains the normative behavior of gesturing while speaking, and it explains that gesturing and using our bodies in various ways help us perform cognitive tasks better. Tversky comments: *"[Gestures] map thought directly. They represent thought [...] as actions in space"* [37: 130].

And, to add just one recent finding of music cognition research, gestures and bodily movements not just express and facilitate thought and emotion, they even influence the way we perceive sound: An upwards or downwards directed gesture or bodily movement (walking upstairs vs. downstairs; reaching up vs. down) influences the perceived pitch of a tone according to the height metaphor [35].

Embodied cognition theories link the mere spatial perception with bodily experiences, rendering spatial structures meaningful through conceptual spatial metaphors.

Stumpf already asked *"if [...] everytime we hear tones, when we abandon ourselves speech- and thoughtless to them, images of the spatial deep, high, raising and falling link with this, self-acting"*[11] [1: 200f].

[11] *"[...] ob also immer, wenn wir Töne hören, sprach- und gedankenlos ihnen hingegeben, Bilder des räumlich Tiefen, Hohen, Auf- und Absteigenden von selbst sich damit verknüpfen."*

The philosopher Gernot Böhme explicates the emotional impact of "musical space", the metaphorical aural space spanned by high and low tones, that is *"experienced affectively"*, namely *"[low pitched elements] as heavy and burdensome, ascending elements as alleviative and joyful"*[12] [48: 266].

Now, considering the pitch/height (and pitch/size) mappings of the auditory object, one might ask if the related conceptual metaphors of space are evoked: Do the embodied conceptions of space have an impact on the sensation of sound? Eitan and Timmers (2010) investigated the connection of pitch and metaphoric height with conceptual metaphors and related associations, finding e.g. *"that ›dark‹ and ›winter‹ are appropriate metaphors for low pitch"*. They concluded: *"In the complexity of its relationships with non-auditory domains, auditory pitch demonstrates how a basic percept may intricately connote to diverse, seemingly remote realms of experience. These cross-domain mappings are often shared consensually while not explicitly expressed in the vocabulary."* [49]

4 Implications for an aesthetics of spatial audio

In music production the dominant aesthetic paradigm is the virtual frontal stage, complemented with a more or less surrounding virtual architectural space that is either recorded in the venue or created in post-production. The stage paradigm is not only the natural approach for working with stereo, but appears to be also the logical approach for the production of music that is typically performed on a stage.

In contrast, space and spatial audio have been crucial creative devices in live performances of sonic arts and electroacoustic music since the introduction of multichannel audio, utilized by artists like Schaeffer, Stockhausen, Xenakis or Boulez. As early as 1955/56, Karlheinz Stockhausen used five (later four) loudspeaker groups surrounding the audience for the performance of "Gesang der Jünglinge", working with spatial effects like static and moving auditory objects [50: 153].

However, the availability of technology seems not to be the issue. Even in modern spatial audio productions

[12] *"Breitgelagertes als schwer und bedrückend, Aufsteigendes als erleichernd und freudig"* – Böhme avoids the spatial metaphor hoch/tief by using the very uncommon term "breitgelagert", derived from ancient Greek terminology for low-pitched tones (which of course is just another metaphor, related to thickness).

the creative potential of the immersive formats is rarely used to full capacity. It appears like in the early days of quadraphonic sound production aesthetics has been more inventive and experimental, maybe inspired by the creative experiments in sonic arts. The advertisement text on the quadraphonic release of Miles Davis's *Bitches Brew* (CBS, 1971) is boasting about the creative potentials: *"The all-around-you presence of sound coupled with the ability to move elements of the program between any pair of speakers allows the utmost flexibility to artists, composers, and arrangers"* – remarkably, the recording engineers are not mentioned here, as the spatialization is obvioulsly regarded an artistic task, as a step in composition or arrangement. Needless to say, the production impressively fulfills the promises.

An aesthetic approach beyond the stage paradigm and inspired by sonic arts leads to what can be called *auditory scenography*, where sounds are arranged in static or dynamic three-dimensional scenes. The technical prerequisite is object based, scene based or channel based content rendered either for large scale loudspeaker arrays or for binaural playback, preferably with headtracking.

A scenographic approach takes advantage of physical and of metaphorical space. Some characteristics are:

- The front and specifically the center position loses its particular significance, as the front is dependent of the orientation of the body or head.
- The vertical orientation might trigger conceptual metaphors and related associations, as does the metaphoric height of the sound. Thus height and vertical movement will be experienced different from horizontal placement and movement: There is a truth behind the "Voice of God" joke.
- According to the crossmodal height metaphor, the congruency of vertical position or movement of auditory objects due to the interplay of metaphorical and physical height becomes a stylistic device.
- Spatial movement becomes likewise a stylistic device, very different from the movement induced through stereo panning on the virtual frontal stage.
- The spatial impression of a scene might depend on the utilization of irregular and in particular oblique movements and structural orientations.
- Specifically in the binaural format, proximity can become a more dominant feature than it already is in headphone-compatible pop music production.

Although one might expect the scenographic paradigm to be suitable rather for experimental musical genres, a case study with traditional acoustic music showed highest ratings for complex spatial arrangements with moving auditory objects [51].

Besides these basic properties of a scenographic production paradigm, spatial audio of course provides more creative opportunities, one being Normandeau's "timbre spatialization" proposal of spectral decomposition and spatial distribution following the pitch/height scheme [52]. Likewise, Diana Deutsch's investigation of pitch dependent lateral localization according to the listener's handedness ("Scale Illusion") [53] offers creative options for spatial effects due to incongruent Gestalt factors constituting the auditory object, not just in romantic orchestral works in traditional European seating or chromatic organ music performed on large symmetrically built instruments.

5 An artistic experiment in full 3D Audio

The perception of spatial scenes of auditory objects has been investigated in a study in artistic research on the aesthetics of spatial audio.

5.1 Test setup

In order to have access to the lower half-space the study was performed in a studio equipped with a vertically symmetric 14-channel playback system (Fig. 3).

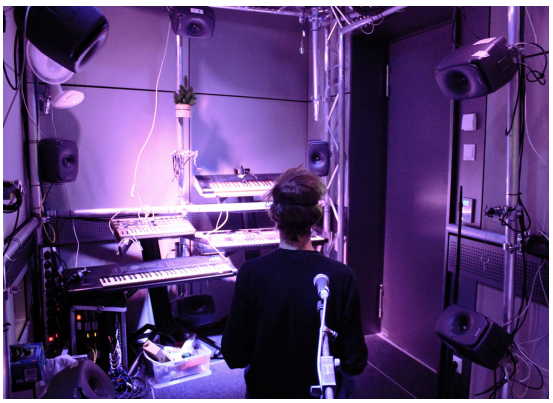

Fig. 3: Experimental setup, Immersive Audio Lab

Spatial structures under investigation were composed of 3, 4, and 5 distinctly different auditory objects (high pitched constant hiss / irregular clicks / kalimba-like

sound / deep impulsive noise / bass drone). The test subjects were asked to reconstruct the perceived spatial structures by means of a hardware controller (method of adjustment). Azimuth and elevation angles could be set individually for each auditory object.

25 persons, mainly graduate students with focus on sound production, artists, and composers, participated in the test. The subjects were individually situated in the sweet spot, and they were blindfolded to eliminate the visual factor of the loudspeaker array. Head movements were expressly allowed to localize the sounds.

The loudspeaker array was set up as follows:

- top center
- upper layer four channels $0°$ / $\pm 90°$ / $180°$ elevation $32...37°$
- middle layer four channels $\pm 35°$ / $\pm 145°$
- lower layer four channels $0°$ / $\pm 90°$ / $180°$ elevation $-30...-37°$
- bottom center

The signals were rendered in Max/MSP with Ircam Spat5 in 7th order Ambisonics and decoded for the loudspeaker array with the IEM AllRADecoder.

5.2 Spatial compositions under test

The different sound elements were arranged using three regular geometric structures:

- linear arrangement of three equidistant elements (“line”)
- surrounding arrangement of four equidistant elements (“square”)
- surrounding arrangement of five equidistant elements (“pentagon”)

These geometric structures were presented to the test subjects in different spatial orientations. For example, a horizontal, an oblique and a vertical variant was derived from the linear arrangement. In this way, only the position of the overall structure in relation to the listener was changed, while its internal geometric relationship was retained.

The individual sounds used for the spatial compositions differ in their temporal and spectral properties considerably from each other. When listening to all the sounds each fights for attention. Especially the differences in the temporal and spectral properties display the spatial structure more clearly.

5.3 Preliminary results

According to the test subjects, listening to and restoring configurations of five sound elements required a high level of concentration, while configurations of three elements were relatively easy to master. Accordingly, the arrangements of three elements were processed much faster than those with five elements, and the accuracy of the restored positions was higher.

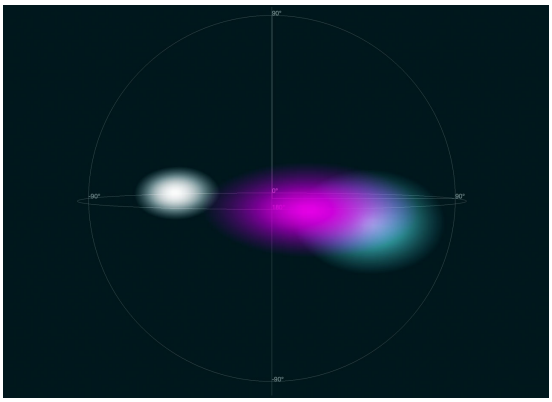

Fig. 4: Reconstruction result example: frontal horizontal “line” arrangement of 3 sounds (left to right: high pitched constant noise / irregular clicks / deep impulsive noise), 25 test subjects, size depicts the angular standard deviation

Systematic vertical deviations of the reconstructed auditory objects appear to be consistent with the height metaphor, the spatial scattering consistent with the size and thickness metaphors (Fig. 4).

The majority of the test persons stated that they perceive an overall spatial impression as well as being able to identify individual sounds one after the other, thus switching between the scene and the individual auditory objects.

During the experiment, clear differences between vertical and horizontal variants of a geometric structure were found. In the reconstruction, vertical deviations of a horizontal arrangement were considerably larger than horizontal deviations of a vertical arrangement, most obviously in the rather complex “pentagon” stucture. This suggests that vertical sonic structures are more easily identified and more precisely reconstructed than is the case with horizontal arrangements.

More results will be presented in [54].

6 Summary

Although crossmodal effects and the role of embodiment in the perception of sound have been studied extensively during the past decades, and although their impact not just on language, but also on the cognition of sound is known, spatial audio is often regarded a solely technical issue, and spatial production a task comparable to traditional stereo or surround mixing.

The scenographical production paradigm proposed here takes some of the interdependencies and interferences of physical and metaphorical properties of complex auditory scenes into account, adopting spatial composition principles of sonic arts and electroacoustic/acousmatic music, and thus attempting to take advantage of the rarely explored aesthetic potential of spatial audio, including complex spatial scenes of moving auditory objects. In such a production approach the spatialization is rather a crucial step in the composition process than a mere mixing task.

Assuming the widespread availability of binaural rendering including headtracking in computers and digital mobile devices, such a novel production aesthetics for spatial audio beyond the frontal stage paradigm of stereo, surround, and traditional 3D productions appears reasonable and worthwile.

Experiments and case studies with both electracoustic compositions and traditional music productions support this aesthetic approach.

Acknowledgement

A special thanks goes to Melina Stephan and Benjamin Yat-Fung Wong for assistance in the experiments performed in the Immersive Audio Lab.

This work was partly funded by Hamburg Ministry of Science, Research and Equality (BWFG) grant number LFF-WKGK-05 within the project KiSS Kinetics in Sound & Space.

References

[1] Stumpf, C., *Tonpsychologie*, 1, S. Hirzel, 1883.

[2] Köhler, W., *Gestalt Psychology – An Introduction to New Concepts in Modern Psychology* [1947], Liveright, 1970.

[3] Ramachandran, V. S. and Hubbard, E. M., "Synaesthesia - A Window Into Perception, Thought and Language," *Journal of Consciousness Studies*, 8(12), pp. 3–34, 2001.

[4] Sapir, E., "A study in phonetic symbolism," *Journal of Experimental Psychology*, 12(3), pp. 225–239, 1929.

[5] Westermann, D., "Laut, Ton und Sinn in westafrikanischen Sudansprachen," Festschrift Meinhof, 1927.

[6] Jakobson, R. and Waugh, L. R., *The Sound Shape of Language*, Harvester Press, 1979.

[7] Lakoff, G. and Johnson, M., *Metaphors We Live By* [1980], University of Chicago Press, 2003.

[8] Pedersen, T. H., "Perceptual characteristics of audio," DELTA Tech Document TN7, 2015.

[9] Shayan, S., Ozturk, O., and Sicoli, M. A., "The Thickness of Pitch: Crossmodal Metaphors in Farsi, Turkish, and Zapotec," *Senses & Society*, 6(1), pp. 96–105, 2011.

[10] Merriam, A. P., *The Anthropology of Music*, Northwestern University Press, 1964.

[11] Stone, R. M., "Toward a Kpelle Conceptualization of Music Performance," *The Journal of American Folklore*, 94(372), pp. 188–206, 1981.

[12] Spence, C., "Crossmodal correspondences: A tutorial review," *Attention Perception & Psychophysics*, 73(4), pp. 971–995, 2011.

[13] Lewkowicz, D. J. and Turkewitz, G., "Cross-Modal Equivalence in Early Infancy: Auditory-Visual Intensity Matching," *Developmental Psychology*, 16(6), pp. 597–607, 1980.

[14] Melara, R. D. and O'Brien, T. P., "Interaction between synesthetically corresponding dimensions," *Journal of Experimental Psychology: General*, 116(4), pp. 323–336, 1987.

[15] Marks, L. E., "On Cross-Modal Similarity: Auditory-Visual Interactions in Speeded Discrimination," *Journal of Experimental Psychology: Human Perception and Performance*, 13(3), pp. 384–394, 1987.

[16] Marks, L. E., "On Cross-Modal Similarity: The Perceptual Structure of Pitch, Loudness, and Brightness," *Journal of Experimental Psychology: Human Perception and Performance*, 15(3), pp. 586–602, 1989.

[17] Rusconi, E., Kwan, B., Giordano, B. L., Umiltà, C., and Butterworth, B., "Spatial representation of pitch height: the SMARC effect," *Cognition*, 99(2), 2005.

[18] Gallace, A. and Spence, C., "Multisensory synesthetic interactions in the speeded classification of visual size," *Perception & Psychophysics*, 68, pp. 1191–1203, 2006.

[19] Evans, K. K. and Treisman, A., "Natural cross-modal mappings between visual and auditory features," *Journal of Vision*, 10(1), pp. 1–12, 2010.

[20] Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., and Johnson, S. P., "Preverbal Infants' Sensitivity to Synaesthetic Cross-Modality Correspondences," *Psychological Science*, 21(1), pp. 21–25, 2010.

[21] Knöferle, K. and Spence, C., "Crossmodal correspondences between sounds and tastes," *Psychonomic Bulletin & Review*, 2012, doi:10.3758/s13423-012-0321-z.

[22] Ozturk, O., Krehm, M., and Vouloumanos, A., "Sound symbolism in infancy: Evidence for sound–shape cross-modal correspondences in 4-month-olds," *Journal of Experimental Child Psychology*, 114(2), pp. 173–186, 2013.

[23] Dolscheid, S., Hunnius, S., Casasanto, D., and Majid, A., "Prelinguistic Infants Are Sensitive to Space-Pitch Associations Found Across Cultures," *Psychological Science*, 25(6), p. 1256–1261, 2014.

[24] Akiva-Kabiri, L., Linkovski, O., Gertner, L., and Henik, A., "Musical space synesthesia: Automatic, explicit and conceptual connections between musical stimuli and space," *Consciousness and Cognition*, 28, pp. 17–29, 2014.

[25] Lewkowicz, D. J. and Minar, N. J., "Infants Are Not Sensitive to Synesthetic Cross-Modality Correspondences: A Comment on Walker et al. (2010)," *Psychological Science*, 25(3), pp. 832–834, 2014.

[26] Walker, P., Bremner, J. G., Lunghi, M., Dolscheid, S., Barba, B. D., and Simion, F., "Newborns are sensitive to the correspondence between auditory pitch and visuospatial elevation," *Developmental Psychobiology*, 60, pp. 216–223, 2018, doi:https://doi.org/10.1002/dev.21603.

[27] Korzeniowska, A. T., Root-Gutteridge, H., Simner, J., and Reby, D., "Audiovisual crossmodal correspondences in domestic dogs (Canis familiaris)," *Biology Letters*, 15(11), pp. 1–5, 2019.

[28] Dolscheid, S., Shayan, S., Majid, A., and Casasanto, D., "The Thickness of Musical Pitch: Psychophysical Evidence for Linguistic Relativity," *Psychological Science*, 24(5), pp. 613–621, 2013.

[29] Fernández-Prieto, I., Spence, C., Pons, F., and Navarra, J., "Does Language Influence the Vertical Representation of Auditory Pitch and Loudness?" *i-Perception*, I-II, 2017, doi:10.1177/2041669517716183.

[30] Pratt, C. C., "The spatial character of high and low tones," *Journ. Experimental Psychology*, 13, pp. 278–285, 1930.

[31] Roffler, S. K. and Butler, R. A., "Localization of Tonal Stimuli in the Vertical Plane," *Journal of the Acoustical Society of America*, 43(6), pp. 1260–1266, 1968.

[32] Ferguson, S. and Cabrera, D., "Vertical Localization of Sound from Multiway Loudspeakers," *Journ. Audio Eng. Soc.*, 53(3), pp. 163–173, 2005.

[33] Cabrera, D. and Tilley, S., "Vertical Localization and Image Size Effects in Loudspeaker Reproduction," AES 24th International Conference on Multichannel Audio: The New Reality, Banff, 2003.

[34] Subkey, A., Cabrera, D., and Ferguson, S., "Localization and Image Size Effects for Low Frequency Sound," Convention Paper 6325, AES 118th Convention, Barcelona, 2005.

[35] Hostetter, A. B., Dandar, C. M., Shimko, G., and Grogan, C., "Reaching for the high note: judgments of auditory pitch are affected by kinesthetic position," *Cognitive Processing*, 20, pp. 495–506, 2019.

[36] Olson, D. R. and Bialystok, E., *Spatial Cognition. The Structure and Development of Mental Representations of Spatial Relations* [1983], Psychology Press, 2009.

[37] Tversky, B., *Mind in Motion. How Action Shapes Thought*, Basic Books, 2019.

[38] Arnheim, R., *Art and Visual Perception*, University of California Press, 1965.

[39] Kövecses, Z., *Metaphor and Emotion. Language, Culture, and Body in Human Feeling*, Cambridge University Press, 2010.

[40] Casasanto, D. and Bottini, R., "Spatial language and abstract concepts," *WIREs Cognitive Science*, 5, pp. 139–149, 2014.

[41] Casasanto, D., "Body Relativity," in L. Shapiro, editor, *The Routledge Handbook of Embodied Cognition*, pp. 108–117, Routledge, 2014.

[42] Wilson, R. A. and Foglia, L., "Embodied Cognition," in E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017.

[43] Jeannerod, M., "The Representing Brain: Neural Correlates of Motor Intention and Imagery," *Behavioral and Brain Sciences*, 17(2), p. 187–202, 1994, doi:10.1017/S0140525X00034026.

[44] Rizzolatti, G., Fogassi, L., and Gallese, V., "Neurophysiological Mechanisms Underlying the Understanding and Imitation of Action," *Nature Reviews Neuroscience*, 2, pp. 661–670, 2001.

[45] Rizzolatti, G. and Craighero, L., "The Mirror-Neuron System," *Annual Review of Neuroscience*, 27(1), pp. 169–192, 2004, doi:10.1146/annurev.neuro.27.070203.144230, pMID: 15217330.

[46] Craighero, L., "The Role of the Motor Systems in Cognitive Functions," in L. Shapiro, editor, *The Routledge Handbook of Embodied Cognition*, pp. 51–58, Routledge, 2014.

[47] Gentilucci, M. and Corballis, M. C., "From Manual Gesture to Speech: A Gradual Transition," *Neuroscience & Biobehavioral Reviews*, 30(7), pp. 949–960, 2006, ISSN 0149-7634, doi:https://doi.org/10.1016/j.neubiorev.2006.02.004.

[48] Böhme, G., *Atmosphäre. Essays zur neuen Ästhetik*, edition suhrkamp, 7th edition, 2017.

[49] Eitan, Z. and Timmers, R., "Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context," *Cognition*, 114(3), pp. 405–422, 2010.

[50] Stockhausen, K., *Texte zur eektronischen und instrumentalen Musik Bd. 1*, DuMont Schauberg, 1963.

[51] Karadoğan, C. and Görne, T., "Auditory Scenography in Music Production: Case Study Mixing Classical Turkish Music in Higher Order Ambisonics," Conference Paper, AES Conference on Immersive and Interactive Audio, York, 2019.

[52] Normandeau, R., "Timbre Spatialisation: The medium is the space," *Organised Sound*, 14(3), pp. 277–285, 2009.

[53] Deutsch, D., "Auditory Illusions, Handedness, and the Spatial Environment," *Journ. Audio Eng. Soc.*, 31(9), pp. 607–618, 1983.

[54] Troschka, S., Stephan, M., Wong, B. Y.-F., and Görne, T., "Perception and reconstruction of spatial sound configurations," Conference Paper, International Conference on Immersive and 3D Audio I3DA, Bologna, 2021 (accepted).