The Dimensions and Complexities of Audio-Visual Association

Augoustinos Tsiros
Centre for Interaction Design
Edinburgh Napier University
Edinburgh, EH10 5DT
a.tsiros@napier.ac.uk

This paper draws on literature from psychology, neuroscience, linguistics, and philosophy to portray a cross-modal conception of auditory and visual phenomena focusing on the similarities in form, structure, and function and exploring the links to perception, conception, and language. Through an extensive literature review, we identify issues related to audio-visual association and explore divergences and convergences between the two modalities. The aim of this paper is to examine how recent research findings from brain science could inform the theoretical and methodological approaches used for studying similarity of auditory and visual percepts in the context of designing multimodal interaction.

Perception, Cognition, Multimodality, Structure, Embodied Metaphor, Similarity, Multimodal Interaction

1. INTRODUCTION

Although nowadays, we have the technologies required to make multisensory associations, the differences between modalities, in terms of physics, psychophysics, neurophysiology, as well as the subjective interpretation of sensory signals, introduce a serious obstacle to forming a general theory of correspondence (i.e. a set of principles that underpin both sensory modalities, and that have the potential to inform the process of designing intuitive audio-visual associations, for comprehension and interaction). As digital technology allows us to make arbitrary associations between any type of data (i.e. modal or amodal), the question of how we can approach the design of such associations in objective terms poses a challenge to a number of disciplines including computer music, information display, sensory substitution amongst others. At first glance, sensory modalities might appear to be of such disparate nature. Table 1 briefly demonstrates that the gap between the two senses in terms of physics, psychophysics, and cognition is significant. This factor often discourages further experimentation of the similarities between the two sensory modalities in objective terms. However a number of studies that have empirically investigated audio-visual feature correspondence show that a number of specific feature pair correlations have been highly rated by participants and there are also cross-study consistencies in the feature pairs that were considered as being good correlates (Walker (1987), Lipscomb et al. (2004), Giannakis (2000) (2006), Berthaut et al. (2010), Küssner et al. (2012)).

Examining the similarities between auditory and visual perception and cognition could lead to a better understanding of the multimodal nature of perception, shed light on the relationships between perception, conception and language, and expose common underlying principles that underpin all sensory modalities. Based on such principles, multisensory associations could be designed in more objective terms, leading to representations that derive their significance and expressive power from the sensory experience rather than established through learning and cultural convention i.e. representations that can be comprehended as opposed to be learned, and have an isomorphic character.

Table 1: *A brief comparison of Vision and Audition*

Source of Information	**Comparison of Vision and Audition**	
	Vision	**Audition**
Transmission	*Fast*	*Relatively slow*
Wavelength	*Very short*	*Relatively long*
Frequency	*Very high*	*Relatively low*
Primary Concern	*Surface & Objects*	*Source & Events*
Secondary Concern	*Location & Color*	*Surface*

Indeed, finding a schema of correspondence between auditory and visual senses is not an easy task. The lack of consensus as to how a percept emerges from the brain's physical signals and computational mechanism and the vague, unconstrained, and context dependent notion of similarity make any efforts to compare the two modalities and find correspondences even more difficult. However, if we aim to find answers to these questions, we should begin by identifying which domains between the two modalities could be compared, which is the focus of this paper.

2. UNDERSTANDING STRUCTURE

Emergence is a concept pertaining to a number of field including psychology, biology, linguistics, physical systems, artificial intelligence, information representation and multimodal interaction design. The emergence of something invokes/implies the idea of structure. In the context of perception, the origins of structure have been interpreted in various ways. The most commonly proposed ideas regarding the origins of structure are pre-formation, contingent creation or construction (Piaget, 1968). Piaget finds that the very idea of structure raises some interesting questions. He suggests that although there are divergent views about the concept of structure in different disciplines, we could broadly understand structures in terms of three fundamentals: wholeness, transformations and self-regulation. Piaget approaches the problem of structure from a strictly empirical angle, suggesting that in order to avoid lapsing back to transcendentalist interpretation of structure such as Platonic forms, Husserlian essence, and Kantian prior forms of synthesis, we should focus primarily on procedures, processes, and the relationships between the elements that give rise to recognizable wholes in the states of structure rather than the wholes in themselves. Wholes can undergo transformations. The idea of transformation also has its own enigmatic facets. For transformation to exist, it must be governed by laws. These laws can be both external to the structure, and implicitly generated. The very idea of wholeness and transformation makes the question of their origin and their relationship inevitable. Piaget suggests that we should first distinguish between elements of structure and the laws that are applied to them.

Piaget recognises two types of structures, mathematical / logical structure and psychological/ linguistic structures. He points out that an important difference between mathematical/ logical structures, and psychological/ linguistic structures is that the former are regulated by rules and operations which could be considered as perfect as their operations have their inverse, while psychological, linguistic, and possibly sociological structures are imperfect in the sense that they are not in a strict sense reversible, or reducible to the elements which the structure consists of. The third fundamental aspect of structure is self-regulation. Self-regulation refers to the rules and/or operations that define the affordances between wholeness and transformation. Piaget's conception of self-regulation refers to the rules or logic by which the structure is governed or bound.

2.1 An Example of Psychological Structures

An example of psychological structure can be found in gestalt perception. Gestalts are the organisational principles of cognition that allow humans to perceive structure, patterns and wholes in nature. Gestalt theory introduced the concept of Gestalten which literally means configuration and can be understood as recognisable wholes in the states of the structure. Gestaltens are multidimensional entities, patterns, qualities consisting of a variable number of attributes, and signify concepts which are not fully described or reduced to their constituent parts (North, 1995). Gestalt principles could be then seen as involuntary, active and coercive processes of equilibration and self-regulation that define both the transformation afforded by the whole and the selection of normative/ordinary forms. Piaget states that the structure never leads beyond the system by which it is governed.

According to Piaget, the first gestalt law determines the relationship between the wholes in structure and the elements the structure consists of. The law states that the elements of a structure are in a sense subordinate to the whole. For example, any transformation in the configuration of the elements of the structure would primarily affect the perception of the whole and not that of the individual elements. The second law is that of good form, which states that perceptual totalities usually are simple, symmetrical and regular. Good forms are the results of equilibration which enables us to organise the stimulation patterns in ways which conform to meaningful forms based on prior perceptual knowledge. This knowledge derives from wholes commonly experienced in the past. Gestalt laws are cognitive heuristics that provide an interface between logical/ mathematical structures and psychological structures. Gestalt laws apply to auditory, visual, and haptic perception and determine how structure is perceived, so it could be argued that gestaltens and gestalt laws are an integrative force between modalities. Therefore gestalt perception is potential area were structural, and functional isomorphisms between sensory modalities can be identified. Examining empirically similarities between the rules and the phenomenology of gestalt is a promising direction for the design of digital multimodal association and interaction. In the following sections, we will discuss a number of studies that aim to shed light on the relationships between auditory and visual modalities.

3. CONVERGENCE AND DIVERGENCE BETWEEN AUDITORY AND VISUAL MODALITIES

It could be argued that physiological and cognitive mechanisms involved in perceiving visual and auditory information appear to diverge in a number of ways. Each modality occupies different parts of the brain, handles different types of signals, and produces a rather distinct phenomenological experience. For example, light consists of electromagnetic waves, travels extremely fast and has extremely small wave length. In contrast, sound energy is mechanical, travels relatively slowly, and the size of the wavelength can be rather large (Handel, 2006). In the case of hearing, mechanical waves are broken down into frequency components by the hair cells in the inner ear which bend depending on the variation of the intensity of specific frequencies. While in the case of seeing, cells fire to the intensity variations in small regions of the retina and moreover fire maximally to intensity variations that occur alongside specific directions, (ibid: p.7).

Further audition is viewed predominately as a temporal sense closely associated with the recognition of events. Vision is viewed predominately as a spatial sense closely associated with recognition of objects (Handel (2006), Bregman (1990)). According to the Oxford English Dictionary, the definition of an object is "something thrown in the way " or "to stand in the way so as to obstruct or obscure" while the definition for an event is "to emerge out of a temporal flow" (ibid: p.5). According to Kubovy et al. (2001), this is somewhat a misleading and an oversimplified view. Kubovy et al. support that the main obstacle to considering sound as objects and not as events is that unlike most visual objects, sound is not opaque. Sound spectrum tends to be sparse. In visual terms, sound could be thought as fence. Sounds in most cases consist of a fundamental frequency and discreet harmonics resulting in sparse structures of varied distribution, and as it is the case with most fence like structures, they do not fully prevent from seeing through them (ibid: p.98). Although one sound could mask another, it has been argued that audible sources do not have a direct corresponding attribute to opacity (Bregman, 1990). Kubovy et al. (2001) argues that if it is due to the notion opacity that we fail to identify objecthood in sound then we might need to consider redefining the term 'object' in such a way that it does not rely on the physical attribute of opacity. In audition, we are mainly concerned with sources that produce sound, as oppose to the material and structural properties of the surfaces that reflect sound, which is the case with perceiving light. Surfaces that reflect sound might provide spatial cues and alter perceptible qualities of an auditory object, but they are not what we perceive as the object. A bat for example uses sound in the way we use light to perceive what we consider as visual information. Audition in humans has evolved to perceive spectral flux rather than spectral reflectance which is the case in bat vision (Mollon, 1995).

Nevertheless the two modalities appear to converge and overlap in a number of ways. Handel supports that many of the similarities between auditory and visual perception and cognition suggest that the two senses are fundamentally the same. Both audition and vision function by partitioning and contrasting structure and noise. Both forms of perceiving involve finding structures and patterns in the energy flux. Understanding better the equivalences between the two modalities can deepen our knowledge of how we perceive and experience information in the environment. Handel makes an extensive comparison between hearing and seeing across six sensory domains. More specifically, he compares auditory and visual perception in terms of the following:
1) Transformation of sensory information to perceptual information
2) Characteristics of auditory and visual scenes
3) The transition between noise and structure
4) Perception of motion
5) Visual color and auditory timbre
6) Auditory and visual segmentation

As the first domain, Handel suggests that there are striking similarities in the way auditory and visual perception encodes information received by the sensory organs (ibid: 2006, p.95). Handel made an extensive survey of research on the receptive fields, and the electrical firing spikes caused when encoding sensory into perceptual information. The results suggest that audition and vision are perceptually identical in this respect. In the second domain, Handel explains that there are a number of reasons to argue against similarity between auditory and visual perception at an object level, because auditory objects are more temporally bound while visual objects are more spatially bound. Moreover visual objects tend to occlude other objects behind them while auditory objects sum common frequencies components. However Handel supports that there are a number of good reasons to suspect that there is a concrete set of principles that unify perceptual processing and experience. Similarities can be found in the tuning of sensory receptors to sensory energy, in the hierarchical organization of cognitive function, and in the interactions and integration of sensory specific information (ibid: p.150).

For the third domain, Handel examines similarities between two concepts: (i) the way low level segmentation of perceptual scenes is accomplished in auditory and visual systems, and (ii) the transformation between visual noise to texture and auditory noise to pitch perception. Handel suggests that finding strict structural correspondences in the way patterns emerge in space-time between the two modalities is a difficult task. There are limitations due to the difference in the resolution of the two systems, the organisation of pathways, the difficulties to match physical properties and even levels of a single property (ibid: p.149). Moreover, there are differences in the spatial and temporal aspects between auditory and visual scene analysis in terms of segmentation and grouping, which is essential for the identification of patterns and structures (Handel (2006), Bregman (1990)). In the fourth domain, Handel points out that in both auditory and visual perception of motion, motion is a direct result of the ability to detect changes in the configuration of texture over time, hence motion is tightly linked to texture segregation in both modalities.

Regarding the fifth domain, visual colour and auditory timbre also appear to have similarities. Handel suggests an analogy between colour constancy and the ability to predict sound quality at one pitch and loudness of the different frequency components. Likewise colour constancy also depends on the spectral distribution of the different wavelengths of the components of the light and their intensities. Both colour and timbre are speculative in nature in the sense that they are estimates of likeness that could be categorised under one of the major classes of the colour scale (e.g. red, blue, green), which also is the case with timbre. Of course it should be noted that both the perception of colour and that of timbre are not extremely well understood phenomena in cognitive and psychological terms. For the sixth and final domain, Handel argues that the organisation principles, as well as functional similarities in object identification and the receptive fields of auditory and visual objects, suggest that both modalities are underpinned by principles which are identical, and could be though as generalised Gestalts. For example, according to Handel, the eyes and the ears each receive two signals that are slightly displaced and different; the problem which cognition has to solve is establishing correspondences between displaced stimuli both at unimodal and multimodal level. In vision, the difference is mainly spatial while in audition it is mainly temporal. The task in both cases is to match the signals received by each eye and ear in order to construct a coherent mental representation. Perceiving is not merely about attending to the parts of the sensory stimuli. Perceiving is an active process, which is not modular, but involves interactions and integrations amongst sensory stimuli. A coherent mental representation of the environment is not constructed and experienced independently for each modality. Instead, we perceive the external environment as a unified phenomenon. According to Handel the processing of sensory information "occurs both simultaneously, in parallel at different neural locations, and successively, serially, as firing patterns converge from this locations" (ibid: p.7).

3.1 Cross-modal binding based on the two streams hypothesis

It has been suggested that there are two subsystems in perception named the dorsal stream which is concerned with identification of objects and the ventral stream which is concerned with the spatial location of object in relation to an individual (Kubovy 2010). The dorsal stream is also known as the *what* subsystems and the ventral stream also known as the *where* subsystems. It has been argued that audition and vision share the two stream hypothesis. Initially the theory was concerned only with vision, but not long after the hypothesis was applied to auditory perception Rauschecker et al. (2000), and later was further adapted to account for multimodal aspects of perception such as audio-vision (Kubovy et al. (2001), Kubovy et al. (2010)). As Kubovy explains, the senses have evolved to receive and recognise information, with the ultimate goal to aid the organisms that possess the apparatus to survive. The senses should be flexible and adaptable so that they can respond to a rapidly evolving dynamic environment. The stimulus we receive from the environment is only rarely uni-modal. Many sensory phenomena in the nature can be experienced through multiple senses. Therefore it has been argued that studying sensory modalities in an isolated manner could only be justifiable if sensory stimulation of one sensory modality was interpreted independently to the signals received by other sensory modalities (Shimojo et al., 2001).The interaction and integration between sensory stimuli raises questions regarding the mechanisms and the rules that underpin cross-modal perception.

Kubovy et al. (2010) provided a plausible explanation to the question of how sensory information might interact and integrate. Their theory is based on two concepts: the first is the idea of indispensable attributes and the second sensory integration based on the *what/where* subsystems. They support that auditory and visual *what* and *where* subsystems have complex relationship and interact at multimodal level. In order to explain how this might happen, Kubovy devised two thought experiments, one for vision and one for audition. His thought experiments had as objective to explain how the perception of numerosity might be affected by coinciding synchronous stimuli, and investigate which attributes are indispensable for the discrimination of numerosity between objects (ibid: p.56). In the first thought experiment, they considered two visual features light wavelengths and spatial location. Through their thought experiment, they demonstrated that when two colored light sources collapse in space and time, our ability to discriminate between the two is compromised. However, in audition, our ability to discriminate between two sound sources collapses when the frequency and the time are identical. They conclude that the indispensable attributes in the case of sound are frequency and time while in the case of visual information indispensable attributes are space and time. A particularly interesting point they make which goes beyond the individual attributes and their ability to aid in discrimination of numerosity (i.e. indispensable attribute), is the idea of collapse of numerosity. One particularly fascinating aspect of Kubovy's and Valkenburg's theory is the suggestion that, when spatial and temporal alignment and a plausible causal relationship exist between an auditory and a visual object, then the two phenomena collapse into a single percept. They argue that this happens because the auditory and visual *what* and *where* subsystems coincide. This explanation has the strength to demonstrate how shared attributes between the two sources have the ability to affect our perception of numerosity, which has the potential to explain some facets of multisensory binding.

For example in audio-visual speech perception, the visual stimulus is not perceived separately from the auditory stimuli, instead the two are perceived as one phenomenon. Moreover they explain that for a successful binding to occur, the causal relationship between the two objects and their attributes must be plausible in terms of: (i) prior experience of similar events and phenomena, (ii) in terms of time (i.e. synchrony), and (iii) in terms of space (i.e. collocation). Plausible common cause is a concept that deserves more consideration. It could be argued that it is of great importance in the context of designing multimodal systems, and information displays. Plausible common cause implies that the phenomenon of binding can occur as long as the association between two modal phenomena appears realistic according to prior knowledge. Conversely, by enacting this knowledge and applying it to digital multimodal mappings, it should be possible to create intuitive associations, associations that give the impression of collapse of numerosity between the modal elements involved. Therefore for a multimodal association to be considered as intuitive, it does not necessarily have to accomplish an absolute structural isomorphism, but rather be persuasive (i.e. create the illusion of realism by conforming to prior perceptual knowledge). Hence it will be necessary to explore these concepts further to create sensory representation that derive their significance from sensory experience and require minimal learning.

4. AUDITORY AND VISUAL PHENOMENA IN THOUGHT AND LANGUAGE

In order to emphasise the fundamental role of perception and the parallels that can be drawn between perception and conception, Talmy (1996) proposed to think of them under a single term “ception”. Talmy explains that there are many disagreements in psychology with regards to where perception ends and conception begins. Drawing a line between phenomena that are purely perceptual and phenomena that are purely conceptual has been proven a difficult task, (ibid: p.139). For example when one sees a visual object such as a bicycle, does the recognition of the object reside in perception or in cognition? Kant, argued that conception without percept would be empty while perception without concepts would be blind, (Masih, 1993). The interrelationships between perception and conception are many so the point is that it is impossible to have a pure perception or pure conception. Perceptual information is shaped, formed and divided by the concepts that have formed in the past, and concepts have to be filled in with perceptual information to have any coherence of content or substance. Language on the contrary is greatly dependent on perception and conception but the dependence is not reciprocal, as the later is less dependent on the former.

As Barsalou (1999) discusses, due to recent developments in mathematics, linguistics and computer science, we have shifted our attention away from phenomenal feature views of mental representation and strived towards more conceptual/amodal views of perceptual processing. Amodal accounts of mental representation favour the view that a percept’s structural features have arbitrary relationships with the physical input and neurophysiological mechanisms that produced them. For example, the word 'chair' has no systematic or structural correspondence to the sensory-motor neural pattern stimulation that occurs when a person perceives and/or interacts with a chair. Amodal symbol systems form complex structures such as feature lists, schemata, and semantic clusters. Amodal theories lead to a unified, integrative view of perceptual processing where a single symbolic system supports and underpins all higher cognitive functions such as memory, knowledge, language and thought (ibid : p.578). However evidence suggests that both modal and amodal information are in a dialectic relationship and interact in a number of ways. A number of studies have explored building mental images using linguistic input (Denis, (1996) (2002), Taylor and Tversky, (1996)). Evidence suggests on the one hand that semantic processing heavily relies on linguistic functions and descriptions, (Burgess et al. (1997), Landauer et al. (1997)), and on the other hand that affordances derived from sensory-motor simulations underpin semantic processing (Glenberg et al., 1998). For example many linguistic metaphors are expressed using perceptual or corporeal related feature, (e.g., is in over my head, grasp a concept, I felt rough, sweet voice, etc.). The use of such linguistic metaphors proves the salient role of embodied representations and their significance in aiding expression of abstract thought through language. According to Zaltman (2002), it is not surprising that we often reappropriate concepts and properties we have learned through the sensory-motor systems and we use them as metaphors to express abstract thought, emotions, and intentions. Due to the fact that language also constitutes a central element of abstract thought, therefore sensory-motor representations and language must be in a dialectic relationship. This flexibility humans demonstrate in appropriating sensory related metaphors to express abstract thought in language shows that conceptual mapping and blending across modal and amodal information is central to human thinking and consequently we are affluent at creatively combining embodied knowledge to use for expression and communication of feelings, thought, ideas and intentions (Fauconnier (1997), Fauconnier et al. (2002)).

In order to explain how the processes that enable conceptual blending operate Lackoff and Johnson (1980) introduced the theory of image schemas, which has been a cornerstone in cognitive linguistic thought. The theory suggests that image schemas are pre-conceptual and central to the formation of conceptual knowledge and linguistic abilities. According to Hampe (2005), image schemas are highly schematic gestalts which capture the structural coherence of sensory-motor experience integrating information from multiple modalities, beneath conscious awareness. As gestalts image schemas are both active and flexible, enabling to map the phenomenal structures of modal experience to abstract and amodal structures, similar to the way the word chair is related to the experience of the object it signifies. If the knowledge we have gained from embodied experience can be used so creatively and flexibly to express abstract thought, then it could be argued that by enacting/tapping on this knowledge, we could inform the design of multimodal systems, for organisation, comprehension and interaction with modal and amodal information. In order to enact embodied knowledge, painstaking empirical work is required, and further integration of methodological approaches will be necessary (Leman (2008), Ware (1993)).

4.1 Some considerations about studying similarity between auditory and visual percepts

When Hume introduced the term *resemblance* in order to discuss similarity, he recognised similarity as a process of great importance that has a central role in the formation of categories, and in making comparison between objects and concepts (Gamboa, 2007). According to Gamboa, Hume also noticed that not all properties have equal weight when assessing similarity between two items. He argued that when a quality becomes very common amongst many objects or concepts, the property loose its significance and power to establish links between two or more objects. So according to Hume, these properties are given less weight when making a similarity judgement. By and large, humans are inclined to allocate less attention to these features because the possibilities of choice become immense. It could then be said that having common properties might not be adequate to explain similarity. Another approach is to consider similarity through the concept of likeness e.g. object (a) might be considered to be more like object (b) rather than object (c), (Gamboa, 2007).

Beyond shared property similarity, resemblances can occur on higher-level attributes such as relational and semantic similarities, (Gentner et al., 1997). In semantic and relational similarity, we have emergent or contextual property that establishes similarity, by determining which features become salient for a given context. For example empirical research findings show that subjects judged a raccoon and a snake to be more similar when the word pet was presented above the two representations, than when no context was provided (Barsalou, 1982). Goodman (1972), who did a lot of theoretical work in analysing the concept of similarity, suggests that similarity or likeness between two units such as X and Y can not be established until a third contextual/ psychological property Z defines in which respect X and Y are compared. Consequently, when a person is asked to make a similarity judgement between two units without defining from the outset the Z property, guessing what Z property/ies will be selected (i.e. selection criteria) by the person for making the judgement is hard to predict (Goodman, 1972). Therefore, it is reasonable to say that similarity judgements are not simply a predefined comparative judgement by a subject between relevant properties across a set of objects, but rather a complex and flexible process, weighting the importance of relevant properties (Kriegeskorte, (2012), Tversky (1977) (1978)). For example when comparing two objects, the features considered relevant for the comparison represents only a small subset of all the features that the objects consists of. The selection criteria based on which a subset is selected is in a dialectic relationship with the interests and intentions of the subject who makes the similarity judgement. Interests and intentions that define selection criteria might vary depending on context. Hence ambiguous relationships in weighting similarity and dissimilarity judgements, asymmetry in relationships between object pairs, and dependence on contexts pose serious challenges for studying similarity between perceptual phenomena empirically. So when studying similarity, extra caution is required to account for such problems (Goldstone, 1994).

5. DISCUSSION AND CONCLUSIONS

This paper provides a comparison between auditory and visual perception and cognition. Through a survey of relevant literature, the relationship between the two modalities was examined. Similarities and differences between the two modalities were identified examining a wide array of domains ranging from physical to psychological structures. Although there are differences between auditory and visual perception, there are also striking similarities. While the physiology and the phenomenology of the two sensory inputs are distinct, the similarities suggest that in many respects the two modalities are identical. Interactions and integrations between sensory stimuli in the brain suggests that the auditory and the visual senses have a lot in common, and that perceptual experience extends well beyond the information supplied by each individual sensory system. So it can be argued that sense perception cannot be understood by studying the different modalities in isolation. Some form of image schemas (i.e. cross-modal gestalts) should be available to regulate the interaction and integration of perceptual information. Identifying such cross-modal schemas will be crucial for understanding better the multimodal nature of perception, and informing the design of multimodal systems. We currently know that the spatial and temporal alignment of two modal signals is necessary for interaction and integration of information. Moreover a plausible causal relationship between auditory and visual stimuli is also necessary for an intermodal binding to take place. For these causal relationships to be plausible, they need to conform to prior perceptual knowledge.

This paper showed that in the context of digital information representation and multimodal interaction, it is essential that more empirical work is conducted to enact prior embodied knowledge and deploy this knowledge to represent, organise, and interact with sensory representations and their parameters intuitively, and with minimal training and learning. Following Kubovy's et al. argument of the collapse of numerosity when the dorsal and ventral auditory and visual streams coincide, it could be argued that a successful multimodal mapping should aim to achieve such collapse of numerosity between inputs and outputs of the modalities involved. A multimodal mapping that successfully aligns with prior knowledge should give the impression of isomorphism in term of part-wholes, transformation and self regulating rules. Moreover the flexibility which humans demonstrate in the use of embodied metaphors in language shows the central role of perception in language, and that our ability for conceptual blending based on relational/semantic similarity could also provide a channel where intuitive association can be constructed. The need for systematic empirical work to shed light on the multimodal nature of perception is evident, as well as the requirement to identify underlying principles based on which multisensory associations can be designed, which tap into intuitions developed from a lifetime of experiencing phenomena and causal interaction in the environment. However studying similarity between perceptual phenomena is not easy, as similarity is highly context dependent and unconstrained. Research efforts should be made to investigate similarity between auditory and visual phenomena covering domains such structural, semantic similarity, and the use of embodied metaphors. These efforts should be accompanied by research in feature extraction and pattern recognition, computational and physical modelling.

6. REFERENCES

Barsalou L. W. (1999). Perceptual Symbol Systems. *Behavioural and Brain Science* vol. 22, pp. 577–660.

Barsalou, L. (1987). The Instability of Graded Structure: Implications for the Nature of Concepts. In: *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge University Press.

Berthaut F., Desainte-Catherine M. (2010). Combining Audiovisual Mappings for 3D Musical Interaction. *International Computer Music Conference*.

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press.

Burgess, C., Lund, K. (1997). Modelling parsing constraints with high dimensional context space. *Language and Cognitive Processes,* vol. 12, pp. 177–210.

Denis M. (1996). Imagery and the description of spatial configuration. In: *Models of visuospatial cognition*. Oxford University Press.

Denis M. (2002). Can the Human brain Construct Visual Mental Images from Linguistic Input? In: *The languages of the brain.* Harvard University press.

Fauconnier G., Turner M. (2002). *The way we think: Conceptual Blending and the Minds Hidden Complexities.* Basic Books.

Fauconnier G. (1997). *Mapping in Thought and Language.* Cambridge University Press.

Gamboa S. (2007). Hume on resemblance, relevance, and representation. *Hume Studies* vol. 33, no. 1, pp. 21–40.

Gentner D., Markman A. B. (1997). Structure mapping in analogy an Similarity. *American Psychologist* vol. 52, no.1, pp. 45-56.

Giannakis K. (2006). A comparative evaluation of auditory-visual mappings for sound visualisation.

Organised Sound Journal, vol. 11, no. 3, pp. 297–307.

Giannakis K., Smith M. (2000). Towards a theoretical framework for sound synthesis based on auditory-visual associations. *AISB'00 Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, pp. 87–92.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences* vol. 20 pp. 1– 55. In: see Barsalou 1999.

Goldstone, R. L . (1994). The role of similarity in categorization: Providing groundwork. *Cognition*, vol. 52, pp. 125-157.

Goodman, N. (1972). Seven strictures on similarity. In: *Problems and projects*,. Indianapolis/ New York: Bobbs-Merrill pp. 437–446.

Hampe B. (2005). Image schemas in Cognitive Linguistics: introduction. In: *From perception to meaning.* Walter de Gruyter press.

Handel S. (2006). *Perceptual Coherence: Hearing and Seeing.* Oxford University Press.

Holtzman S. R. (1994). *Digital Mantras*. MIT Press Cambridge MA.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology 3.*

Kubovy, M. and Van Valkenburg D. (2001). Auditory and visual objects. *Cognition* vol. 80, pp. 97-126.

Kubovy M.and Schutz M., (2010). Audio-visual objects. *Review of Philosophy and Psychology*, vol. 1, pp. 41–61.

Küssner M. B., Prior H. M., Gold N. E., Leech-Wilkinson D. (2012). Getting the shapes "right" at the expense of creativity? How musicians' and non-musicians' visualizations of sound differ. *International Conference on Music Perception and Cognition, European Society for the Cognitive Science of Music*.

Lakoff, G., Johnson, M. (1980). *Metaphors we live by.* University of Chicago Press.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* University of Chicago Press.

Landauer, T. K., Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, vol 40, pp. 104-211.

Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive Science*, vol 10, pp. 1–40. In, see Barsalou 1999.

Leman M. (2008). Systematic musicology at the crossroads of modern music research. In: Schneider, A., *Systematic and Comparative Musicology: Concepts, Methods, Findings,* vol. 24, pp. 89–115. Hamburger Jahrbuch für Musikwissenscha, Peter Lang.

Mollon, J. (1995). Seeing colour. In: Lamb T., Bourriau J., *Colour: art & science,* Cambridge University Press, pp. 127-151.

Lipscomb S. D., Kim E. M. (2004). Perceived match between visual parameters and auditory correlates: an experimental multimedia investigation. *8th International Conference on Music Perception and Cognition.*

Masih Y. (1993). *A Critical History of Western Philosophy*. Motilal Banarsidass.

North W. (1995). *The Handbook of Semiotics.* Indiana University Press.

Piaget J. (1968). *Structuralism.* Routledge and Kegan Paul.

Rauschecker J.P., Tian B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *National Academy of Science, USA*.

Shimojo S., Shams L. (2001). Sensory modalities are not separate modalities: plasticity and interactions. Elsevier Science.

Talmy L. (1996). Fictive motion in language and "ception." In: *Language and space,* ed. P. Bloom, Peterson M., Nadel L., & Garrett M. MIT Press.

Tversky, A. (1978). Study of similarity. In: Rosch E., and Lloyd B. B., *Cognition and categorization*. Hillsdale NJ: Lawrence Erlbaum Associates, Inc.

Tversky, A. (1977). Features of similarity. *Psychological Review,* vol. 84, pp. 327-352.

Walker R. (1987). The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception and Psychophysics, v*ol 42, no. 5, pp. 491–502.

Ware C. (1993) The Foundation of Experimental Semiotics: a Theory of Sensory and Conventional Representation. J*ournal of Visual Languages and Computing.* vol 4, pp. 91-100.

Zaltman G. (2002) Eliciting Mental Models through Imagery. In: *The languages of the brain*. Harvard University Press.