Motion as the Connection Between Audio and Visuals

and how it may inform the design of new audiovisual instruments

Niall Moody
Centre for Music Technology
University of Glasgow
UK
niallmoody@yahoo.co.uk

ABSTRACT

While it is often claimed that any connection between visual art and music is strictly subjective, this paper attempts to prove that motion can be used to connect the two domains in a manner that, though not objective in the true sense of the word, will nevertheless be perceived in a similar fashion by multiple people. The paper begins with a brief history of attempts to link audio and visuals, followed by a discussion of Michel Chion's notion of synchresis in film, upon which the aforementioned claim is based. Next is an examination of motion in both audio and moving images, and a preliminary look at the kind of mappings which may be possible. As this is all done with a view to the design of new musical, or audiovisual, instruments, the paper finishes with a brief discussion of an instrument being designed based on this idea that motion can act as a connecting force.

Keywords

motion, audiovisual, instrument, mappings, visual music

1. INTRODUCTION

Though it is perhaps too disparate to be called a tradition, there is a significant precedent in attempts to link visual art and music, from the musical inspiration behind the abstract paintings of Wassily Kandinsky[8] to the relatively numerous colour organs[10] developed since the early 1700s. While it is often claimed that there is no truly objective connection between *colour* and sound, this paper aims to demonstrate that, though it is not objective in the true sense of the word, motion can act as a connection which is nevertheless perceived in much the same way by multiple people, regardless of their background. Beginning with a brief history of attempts to link audio and visuals in order to set the scene, the paper then moves onto a discussion of Michel Chion's notion of synchresis, which forms the basis for much of the discussion in the paper. This is followed by an examination of how motion may connect the two domains, and a number of example mappings between audio and visuals are presented.

2. A BRIEF HISTORY OF AUDIO-VISUAL RELATIONS

This section will present a brief overview of what the author feels are the most significant attempts to link audio and visuals, beginning with Father Louis-Bertrand Castel's original 'colour organ' from the mid-1700s. The section is subdivided into various subgenres of, or threads within, audiovisual art, though it should be noted that there is considerable overlap between these sections, and they are in no way intended as comprehensive categories.

2.1 Colour Organs

The earliest known attempt to fuse music and visuals was Father Louis-Bertrand Castel's Clavecin Oculaire[10], or Ocular Harpsichord - essentially the first example of a family of instruments that are generally called colour organs (though this term is sometimes used to refer to music visualisation devices, this section is primarily concerned with a particular group of audiovisual *instruments*). The instrument was based around a normal harpsichord, above which was mounted a large frame, with sixty small windows of coloured glass, corresponding to the notes on the keyboard. Behind these windows were mounted candles. When the performer struck a key, it would lift a curtain in the window, to allow the light of the candle to shine through. In this manner, Castel attempted to link colour and music in a very specific way, where a particular colour corresponded directly to a particular note. Indeed, this idea that colour could correspond directly to pitch would seem to be a common thread among colour organs, at least those which use a piano-style keyboard as their interface. A more recent example would be A. Wallace Rimington, whose book 'Colour-Music: The art of mobile colour' describes an instrument similar to Castel's, albeit using electric illumination rather than candles (and dispensing with the sound production mechanism altogether), as well as his own 'colour scale'[1], whereby colours are mapped to the keys on its keyboard according to the wavelength of light (i.e. red having the lowest wavelength corresponds to C, and so on).

[1]There is an interesting comparison of various 'colour scales', including Castel's and Rimington's, at the RhythmicLight.com website[6], specifically `http://rhythmiclight.com/archives/ideas/colorscales.html`

2.2 Visual Art

Visual art's (in this instance, visual art refers primarily to painting) interest in an audio-visual connection first gained prominence in the early twentieth century. With the rise of instrumental music in the nineteenth century, a number of artists began to see its 'pure' abstraction as an ideal that visual art should strive for, as opposed to the previous, representational forms. One of the major figures in this movement towards 'musical' abstraction was Wassily Kandinsky, for whom music was a major inspiration. In the book 'Concerning the Spiritual in Art'[8] for example, he states his claim that "the various arts are drawing together. They are finding in music the best teacher. With few exceptions music has been for some centuries the art which has devoted itself not to the reproduction of natural phenomena, but rather the expression of the artist's soul, in musical sound"[2]. It was this expression of the artist's soul that Kandinsky sought to achieve through his painting, seeking to create visually the kind of abstraction that had existed within music for a number of years, and which he saw as a pure expression of the artist's 'inner life'[3]. While Kandinsky is perhaps one of the most prominent painters for whom music was a significant inspiration, he is by no means alone. Morgan Russell and Stanton Macdonald-Wright, for example, derived a system whereby colours were seen as "intimately related chromatic waves"[4], and complex harmonies and rhythms were built up around this idea. Interestingly, the artists also believed that the rhythms they created infused their paintings with a notion of time, that they suggested motion. This interest in motion further led them to attempt to build a "kinetic light machine"[5] (though it was never completed) which would let them compose with actual, as opposed to suggested, motion.

2.3 Abstract Film

While the visual art described above attempted to link music to visuals by means of static colours and forms (*suggesting* music), the rise of cinema and the various technological advancements that came with it paved the way for an artform able to make use of film's time-based, dynamic nature. One of the most prominent abstract filmmakers was Oskar Fischinger. Fischinger created a number of abstract animations tightly synchronised to music, often using geometric shapes moving in increasingly complex patterns. While Fischinger mainly used existing music by other composers in these films, in 1930 he devised a method of recording geometric shapes directly onto the film soundtrack[7], as a way connecting the visual directly to the aural. This connection was taken further by John and James Whitney in the early 1940s, with their Five Film Exercises[6]. These films were created using two devices built by John: an optical printer to create the visuals, and a device which used the motion of pendulums to create sound. Through the use of these devices, the Whitneys were able to create some of the most striking early audiovisual works of art, where sound and image are inextricably linked. Both Whitney brothers continued to work in the field of abstract film, with James producing such films as Yantra and Lapis[7], and John producing films such as Permutations and Arabesque[8]. John also developed a theory, upon which his later works are based, that patterned motion was the link between music and image, as described in his book: 'Digital Harmony: On the Complementarity of Music and Visual Art'[4].

2.4 Recent Developments

One of the most significant recent developments with regard to an audiovisual artform (certainly when considering audiovisual instruments) is Golan Levin's Audio Visual Environment Suite[9]. The suite is a collection of five (computer-based) audiovisual instruments, which enable the performer to create dynamic, moving images and sounds simultaneously, using a so-called 'painterly' interface. What is significant about these instruments is that they represent an examination of audiovisual interfaces, and an attempt to design an interface more suited to audiovisual performance. While the colour organs described previously typically produce both sound and visual, the fact that the interface is a standard piano-style keyboard surely limits their expressive possibilities, particularly in the visual domain. By choosing a painterly interface (essentially the instruments are controlled via a stylus), Levin arguably allows for a greater range of expression, and certainly much tighter control over the visual aspect of the instruments. In addition to Levin's instruments, there have been numerous installation works in recent years which have attempted to expand upon those audiovisual works before them by moving away from the perceived limitations of film, where images are projected onto a single, fixed, flat surface, and towards a more immersive experience. The book 'Visual Music: Synaesthesia in Art and Music since 1900'[4] details a number of these, such as Jennifer Steinkamp's 'SWELL', and Cindy Bernard and Joseph Hammer's 'projections+sound'.

2.5 Popular Culture

Popular culture has in some respects been particularly receptive of the audiovisual ideas outlined so far. An early example would be Disney's Fantasia, upon which Oskar Fischinger worked before falling out with Walt Disney over artistic differences. Though some of Fischinger's ideas were rejected or altered before the film was released, it is significant that a number of them made it into the final film relatively unscathed[9]. Another film commonly cited in audiovisual writings is Stanley Kubrick's '2001: A Space Odyssey', specifically the stargate sequence, which was created using slit-scan photography, a technique John Whitney had used himself in his 'Catalog' reel of visual effects. At around the same time, rock concerts would commonly incorporate light shows created by artists such as Mark Boyle and Joan Hills in London, and Glenn McKay in San Francisco[10] (among others) alongside the music. Indeed, it would seem that popular music has driven a number of developments in the field of audiovisual art. The music video is the most obvious example of this, with a huge variety of approaches taken towards the audio-visual relationship, and visuals often very tightly synchronised with the music they are accompanying. An example that is particularly relevant to

[2] [8], p.19
[3] Ibid, p.1
[4] [4], p.43
[5] Ibid, p.46
[6] Ibid, p.125. See also [13], p.138-143.
[7] [4], p.125, p.130-139
[8] Ibid, p.144-147. See also [13], p.97-113
[9] [4], p.89
[10] Ibid, p.161-169

the ideas outlined in this paper is Alex Rutterford's video for Autechre's 'Gantz Graf'[1]. Somewhat related to the music video, though with perhaps a more technical bias, is the demoscene[11], a computer-based subculture, whereby programmers demonstrate their skill by creating software demonstrations of elaborate visual effects synchronised to music, all calculated in real-time. These computer-generated visual effects are themselves related to another, essentially audiovisual, phenomenon, that of music visualisation. Perhaps most popular from iTunes, though common to nearly all recent software music players, music visualisation produces computer-generated visuals and attempts to synchronise them to the music in some way (though often this is reduced to little more than an oscilloscope view of the audio playing).

3. SYNCHRESIS

From here we will take a slight sidestep into the world of film theory, specifically Michel Chion's notion of Synchresis. According to Chion, synchresis (created out of a combination of the words synchronism and synthesis) is "the spontaneous and irresistible weld produced between a particular auditory phenomenon and visual phenomenon when they occur at the same time"[12]. What he is referring to is an effect that has been commonplace in film for a number of years now, and (to borrow one of Chion's examples) can be particularly noticeable when looking at the way punches are often represented in film. In real life, punches rarely make much sound - whether they inflict pain or not - yet in film it is relatively rare that punches are portrayed in a naturalistic way (where the sound heard is exactly what would be heard in real life). Instead, we are accustomed to hearing assorted exaggerated whacks and thumps when a punch connects, the punch almost seeming unreal, and somehow false, if such sound is absent. The point is that the sound, though not actually related to the images we're seeing in any physical way, somehow enhances the image (Chion terms this enhancement 'added value'[13]), and makes it seem more real (or hyper-real). Our brain recognises the synchrony between image and sound and intuitively creates a connection.

While it could perhaps be argued that, in the case of the punch, the exaggerated sound is necessary to convey the physical nature of the action (in that it may not necessarily be conveyed entirely successfully via image alone), Chion claims that it is nevertheless possible for the visuals and sound to be entirely unconnected (i.e. the only thing that links them is their synchresis). Indeed, there are numerous examples of synchresis in films where there is no connection between audio and visual other than their synchrony. For example, a relatively common instance is that of a visual shot of someone walking and, instead of hearing footsteps, orchestral hits are played in sync with each step (this is particularly common in, for example, looney tunes cartoons). While the sound and visual are entirely unrelated if viewed separately, the tight synchrony encourages the viewer's brain to make a connection, to the extent that these two unrelated sequences come to be viewed as a single object.

Chion notes that not all sounds and images may be connected as simply as this however, and states that synchresis "is also a function of meaning, and is organised according to gestaltist laws and contextual determinations"[14]. The result is that certain sounds will 'adhere' to a particular visual better than others, and that this will often rely significantly on the context within which the connection is made. To return to the footsteps example previously, Chion views this as having an "unstoppable" [15] synchresis, such that it would be possible to attach any kind of sound to the image without breaking the connection between the two. This is due to our experience of the world - we learn from experience that footsteps make a sound, and, in viewing a sequence of someone walking, we expect to hear a corresponding sound - what that sound actually is, is less important than the fact that the sound occurs when we're expecting it.

4. MOTION AS THE CONNECTION

With synchresis we can see that it is possible to create a connection between audio and visual that, though not strictly objective, will nevertheless be perceived in much the same way by anyone who experiences it, regardless of their background. Looking more closely, however, what is actually happening when we experience the synchresis of the footsteps, for example? We see the foot moving in a particular way, and we hear sound accompanying (or from a different perspective, reacting to) that motion. Indeed, it is the author's contention that synchresis of this form is based on the fact that the motions of the two domains (visual and aural) are related in some way. With the footsteps, we see the foot moving downward, then coming to a sudden stop, at which point a sound event is initiated. This sound event imparts various pieces of information, but looking at its motion, we can see that its amplitude envelope, for a start, is closely related to the visual motion of the foot. When the foot comes to a halt on the ground, the amplitude envelope of the sound in a sense reacts, in that there is a sudden sharp increase in the sound's level, following which the amplitude decays, as the foot is no longer in the process of colliding with the ground. There is of course additional motion in the sound (the spectrum for example), but from looking at the amplitude envelope alone we can already see a clear link between image and sound in terms of motion.

This form of 'collision-based' motion - where the primary stimulus is the sudden collision of two visual objects - is not the only form of motion which can act as a connection between the two domains however. Another form of motion is the unhindered motion of an object in a linear trajectory across the screen. The author is of the opinion that this kind of motion can prove just as powerful a connection as the collision-based form, provided it is accompanied by a related motion in the audio realm. To understand how this may be, we need to look at our experience of the sound made by objects thrown through the air. A filmic example of this would be in period movies with battle scenes where arrows are used (see for example Gladiator, or any other such film) - when the arrows are flying through the air, there is an accompanying 'whoosh' sound. Another example would be the sound of a jet plane in flight. The point is that experience tells us that objects that move through the air tend to make a sound (albeit provided they are moving relatively

[11] See [2], and [3]
[12] [5], p.63
[13] Ibid, p.5
[14] Ibid, p.63
[15] Ibid, p.64

fast).

Indeed, the idea that motion can act as a connection between audio and visuals is based on our experience of the world. Our experience of the physical world tells us that objects which we can see are in motion will tend to emit sound, as a consequence of this motion (from this perspective, the visual stimuli telling us the object is in motion is also a consequence of the motion). This experience is what allows synchresis to work - our experience tells us to expect some kind of sound in conjunction with certain visual stimuli (and vice-versa, depending on the situation), and our brain, expecting this aural 'event', will connect almost any sound to the visual, *assuming* there is some kind of related motion between the two.

Looking at this from the point of view of gesture could also prove interesting. Gesture is essentially directed motion, which adds another element to our audiovisual connection. If motion connects audio and visual, it also connects gesture, and any audiovisual instrument therefore requires a clear relationship to be established between the three motions if it is to be successful not only from an audience' point of view, but also from the perspective of the performer.

At this point a caveat should be made regarding the kind of motion which may be put to use as this kind of connection. It is important that it is *perceivable* motion. To elaborate, a constant sine tone in the audio realm could be considered as possessing a certain motion, in that it relies on the vibration of particles in air in order to be audible. To the human ear however, the sound produced is fundamentally static (we are assuming the amplitude is constant), as the perception is of a single tone which does not possess any motion of its own. The same applies to the visual realm - if motion is occurring too fast for the eye to perceive, it is hard to see how it could be useful in establishing a connection with an audio stimulus (though realistically, this may be harder to achieve anyway with current monitors/projectors, since aliasing will come into play before the point where motion becomes blurred).

5. HOW TO USE THE CONNECTION

Having examined the ways in which motion may connect audio and visuals then, how can this be put to use in the design of new instruments? The first step is to define the various types of motion available to us, with a view to creating some simple audiovisual mappings as a starting point for further work. With motion, the author would make a distinction between forms of motion, and domains in which motion can occur. To elaborate, forms of motion would refer to a high level description of how something moves, where the something could be anything, whether visual or audio (for example, one form of motion would be periodic motion). Domains in which motion can occur, on the other hand, refers to parts of the audio and visual realms where motion can be perceived (a visual example would be the position of an object of some kind).

5.1 Forms of Motion

Table 1 shows a list of some forms of motion, according to the above definition, though this is by no means intended as a complete list.

Table 1: Forms of Motion

Constant Velocity
Collision-Based Motion
Periodic Motion
Gravity-Based Motion
Discontinuous Motion

Table 2: Some Domains in Which Motion May Occur

Visual		Aural
Position (of an object)		Amplitude
Size (of an object)		Pitch
Rotation		Spectral Content
Smoothness		Spatial Position
Articulation (of an object)		Noise-Pitch
Pattern		

- **Constant Velocity:** This should be fairly self-explanatory. Compared to the other forms of motion, this could perhaps be seen as providing a weaker connection between audio and visuals, since there are no discrete temporal events. This does not mean it cannot prove useful in certain situations, however. An example could be based on the experience of a stationary viewpoint watching objects (e.g. cars) moving at a high speed past it - the related sounds would pan and be subjected to the doppler effect accordingly.

- **Collision-Based Motion:** This is primarily derived from the footstep example previously - in the visual realm, it refers to objects colliding with each other and then reacting. In the audio realm, however, it refers to the kind of sound associated with collisions, referring to the way that, while the visuals are in motion before and after the collision, sound will only be instigated at the point of collision (assuming it is not already in motion from a previous collision).

- **Periodic Motion:** Again this should be fairly self-explanatory, referring to motion that repeats itself in a perceivable fashion.

- **Gravity-Based Motion:** Related to collision-based motion in that it is based on physical phenomena, this essentially refers to attraction/repulsion forces such as gravity. This is probably most easily perceived visually, though aurally it would refer to motion that gradually decreases or increases in range.

- **Discontinuous Motion:** This refers to sudden discrete jumps, as opposed to the mostly smooth motions described previously. An example would be the rapid cutting common in music videos and also seen in certain films.

5.2 Domains in Which Motion May Occur

As mentioned previously, 'domains in which motion can occur' refers to aspects of the visual or aural realms in which motion of the forms discussed above is perceivable. Most of the entries in Table 2 (again, this is by no means an exhaustive list) should be self-explanatory, so rather than go through each one in turn, only the less obvious entries will be discussed here.

- **Smoothness:** This refers to how smooth or coarse a particular part of the visual is. That part could the

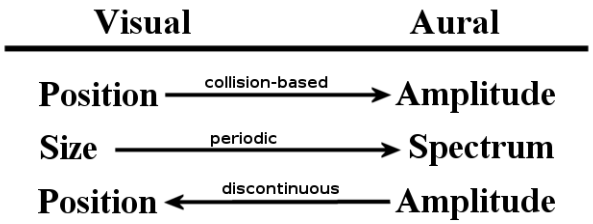

Figure 1: Some example audiovisual mappings

shape of an object, or a more general impression of how colours (and particularly patterns) contrast with each other.

- **Articulation (of an object):** This refers to particular visual objects which may articulate their shape, in much the same way as humans and animals do with their arms, legs etc.
- **Pattern:** Refers to a visual motif which is recognisably periodic, whether it is viewed statically or in motion. This is particularly relevant to John Whitney's notion of Visual Harmony[16].
- **Spectral Content:** Intended as a kind of catch-all in this context, to represent the multitude of ways in which a sound's spectrum may change over time.
- **Noise-Pitch:** Refers to the continuum between pitch and noise.

5.3 Audiovisual Latency

Before we conclude with some example mappings based on the previous discussion, there is one more area which seems ripe for exploration when looking at motion as a connection between audio and visuals: latency. In creating the tight temporal connection between audio and visuals which comes with the use of motion as a connection, we also create the possibility for works which explore this very temporal connection. For example, the visuals could move out of sync with the audio and vice versa, in a similar fashion to Steve Reich's experiments with phase. Indeed, some research has already been conducted in this area, with Yoichi Nagashima having found that the human brain is prepared to accept (and in fact, compensate for) a fairly large degree of time difference between audio and accompanying visuals[11], and that certain time differences can substantially affect how the viewer(/listener) perceives the work.

5.4 Example Mappings

Figure 1 shows some simple mappings between visuals and audio, keeping to those mappings which are less subjective, and more based on common perceptions of the world, and the audiovisual properties of physical objects (there are of course many more possible combinations). The first one describes our now familiar footstep example (at least in part - strictly speaking the spectral content of the audio should be considered as well). The amplitude of the audio is controlled by the collision of objects in the visual realm. Although it is marked as a one-way process, it could be interesting to then map the audio back to visuals in some way, so that a kind of co-operative feedback can be developed. The second example is based around the idea of a throbbing, pulsating object (one could imagine a beating heart), with the size of the visual object periodically growing and shrinking, and this controlling the spectrum of the audio (this could take the form of a low pass filter, where the cutoff frequency is controlled by the size of the visual object). The third example uses the amplitude envelope of the audio (specifically, its transients) to jump a visual object around the screen accordingly, the idea being to create a visual accompaniment to particular audio cues.

[16][13], p.5

6. A PRELIMINARY DESIGN

Though still in the early stages, an instrument is being designed based on the aforementioned principles. The instrument is intended as a sort of 'musical block of clay', visually represented as a 3d, amorphous blob which responds both to the user's input and the instrument's audio output, which will be based on physical modelling synthesis, courtesy of the Tao physical modelling language [12]. A physical user interface is also being designed, intended to allow for the kind of gestures possible or common with clay. Compared to Golan Levin's 'painterly' interface metaphor, it could be seen as a 'sculptorly' interface. The aim is to connect the motions of the audio and the visuals (as well as that of the performer's gesture) as tightly as possible, so that the instrument's output can be viewed as an audiovisual whole, where audio and visual are not easily separated. To do this, a number of mappings based on the aforementioned clay-based interface metaphor will be used (i.e. a 'squeezing' gesture could reduce the size of the visual object, and alter the pitch of the audio accordingly).

7. CONCLUSIONS

Based on Michel Chion's notion of synchresis, this paper has proposed a way of using motion to connect audio and visuals. This connection is derived primarily from our experience of the world (in particular, of the audiovisual properties of physical objects), and plays upon our expectations associated with that experience. With a view to demonstrating some example audiovisual mappings, various different forms of motion were examined. Further to this, the design of an instrument currently being developed based on this idea of using motion as a connection was described.

8. REFERENCES

[1] Warp Vision The Videos 1989-2004. DVD, 27 September 2004.

[2] Demoscene entry at wikipedia. Website: `http://en.wikipedia.org/wiki/Demoscene`, 17 January 2006. Accessed 18/01/2006.

[3] Pouet.net. Website: `http://www.pouet.net/`, 18 January 2006. Accessed 18/01/2006.

[4] Kerry Brougher, Jeremy Strick, Ari Wiseman, and Judith Zilczer. *Visual Music: Synaesthesia in Art and Music Since 1900*. Thames and Hudson, April 2005.

[5] Michel Chion. *Audio-Vision: Sound on Screen*. Columbia University Press, 1994.

[6] Fred Collopy. RhythmicLight.com. Website: http://rhythmiclight.com/, 7 August 2005. Accessed 17/01/2006.

[7] Oskar Fischinger. Sounding Ornaments. *Deutsche Allgemeine Zeitung*, 8 July 1932. Website: http://www.oskarfischinger.org/Sounding.htm Accessed 25/01/06.

[8] Wassily Kandinsky. *Concerning the Spiritual in Art*. Dover Publications, Inc. New York, 1977.

[9] Golan Levin. *Painterly Interfaces for Audio Visual Performance*. PhD thesis, Massachusetts Institute of Technology, 09 2000. Website: http://acg.media.mit.edu/people/golan/thesis/index.html Accessed 18/01/2006.

[10] William Moritz. The Dream of Color Music, And Machines That Made it Possible. *Animation World Magazine*, 1 April 1997. Website: http://mag.awn.com/index.php?ltype=search&sval=Moritz&article_no=793 Accessed 17/01/2006.

[11] Yoichi Nagashima. Measurement of latency in interactive multimedia art. In *New Interfaces for Musical Expression*, 2004.

[12] Mark Pearson. Tao. Website: http://taopm.sourceforge.net, 31 August 2005. Accessed 18/01/2006.

[13] John Whitney. *Digital Harmony: On the Complementarity of Music and Visual Art*. Byte Books, 1980.

Based on the relation between audio and visual informa-tion, Owens and Efros [21], and Korbar et al. [17] concur-rently propose to learn such visual and audio representation by a proxy task, the audio-visual temporal synchronization task.Â In this problem, the audio-visual event may contain multiple actions or motionless sounding ob-jects. The audio-visual event localization problem includes three tasks in [30], i.e., supervised and weakly-supervised audio-visual event localization, and cross-modality local-ization. Tian et al.Â At time t, we denote ftA and ftV as the local feature (segment-level) of the audio segment and visual segment, respectively. Fol-lowing [30], the local feature extractor is xed, and we build our method on top of these local features. Audio-visual communication is passing information as in the form of sound and visual component. Films, Television programs, video chat etc. are some example for audio visual communication. This type of communication can provide more communication accuracy between the individuals whom make the communication. Working of new products can be effectively demonstrated through audio-visual communication. It is an effective form of communication, since it can make a successful communication between the presenter and audience. 14.2K views Â·. View 10 Upvoters.Â Today I doubt students wouldnâ€™t do well without the audio visual components of education. It is now a usual form of educating, Continue Reading. I am addicted to make as close and intimate connections between visuals and audio as I possibly can The core of what I am always trying to create with TouchDesigner is an interactive audiovisual system which I can perform with, and in which there is a deep two-way connection between visuals and sound (so sound creates visuals and visuals also create sound).Â There is no way as pure as the oscilloscope! The great thing about TouchDesigner is that it allows me to go far beyond what a normal lissajous figure can be. â€¢ We propose a novel audio-visual fusion module to as-sociate human body motion cues with the sound sig-nals. â€¢ Our system outperforms previous state-of-the-arts ap-proaches on hetero-musical separation tasks by a large margin. â€¢ We show that the keypoint-based structured repre-sentations open up new opportunities to solve harder homo-musical separation problem for piano, ute, and trumpet duets.Â Early works [5] leveraged the tight associations between audio and visual onset signal to perform audio-visual sound attribution. Recently, Zhao et al.Â Audio-visual learning. With the emergence of deep neu-ral networks, bridging signals of different modalities be-comes easier. A series of works have been published in the past few years on audio-visual learning.