Transient Analysis for Music and Moving Images: Considerations for Television Advertising

Andrew Rogers
University of Huddersfield
andrew.rogers@hud.ac.uk

Ian Gibson
University of Huddersfield
i.s.gibson@hud.ac.uk

ABSTRACT

In audiovisual composition, coupling montage moving images with music is common practice. Interpretation of the effect on an audioviewer's consequent interpretation of the composition is discursive and unquantified. Methodology for evaluating the audiovisual multimodal interactivity is proposed, developing an analysis procedure via the study of modality interdependent transient structures, explained as forming the foundation of perception via the concept of Basic Exposure response to the stimulus. The research has implications for analysis of all audiovisual media, with practical implications in television advertising as a discrete typology of target driven audiovisual presentation. Examples from contemporary advertising are used to explore typical transient interaction patterns and the consequences of which are discussed from the practical viewpoint of the audiovisual composer.

1. INTRODUCTION

Visual and auditory stimuli are interpreted consequent to their semiotic, emotional and functional information. Within our diversifying sensory environments, it is in the minority of occasions that they are experienced discretely.

Audiovisual materials are the prevalent cross modal media format within our society. Targeting our dominant sensory modalities, the multimodal stimulus stream is not simply concurrent, but bilaterally interdependent. The auditory and visual elements interact to form the audioviewer's unified percept by sampling and integrating information from both modalities.

Therefore when composing audiovisual material, decisions have to be made beyond the individual modality structures to the multimodal phenomenology of our interpretation. The innumerate manipulations of the multitudinous variables present in audiovisual montage structures results in this being a diverse theoretical and artistic field. The analysis procedure demonstrated in this paper consciously excludes from analysis any variability within individual modalities. This reductive analytical approach condenses the audiovisual compositional process to a singular core component, the alignment of two information streams, auditory and visual.

The methodology presented explores the audiovisual alignment by assessing and quantifying perceptually weighted transients in each modality. A Transient Interaction Pattern (TIP) is created as the artefact for interpretation of the interaction.

A TIP is easily manipulated via the temporal displacement of either auditory or visual streams. When combining specified musical and visual media into an audiovisual whole, the alignment process is the key variable determining their perceptual interactivity.

Understanding TIPs and their effect on perception enables a composer to manipulate an audience's physiological and consequent psychological response to the presentation, orienting attention at targeted features and instances within the montage. This is applicable to all audiovisual mediums, including but not exclusive to movies, television and documentary film. Television advertising is of particular analytical interest due to its characteristic reliance on visual montage and music in addition to its clear, targeted purpose – the efficacy of the advertised message.

2. MUSIC IN ADVERTISING

Associating an advertisements promotion with a musical stimulus is an uncertain venture as both the advertisement and music carry heavy semiotic connotations as individual entities with the potential to conflict or benefit the presentation in their association. For instance, both advertising theory and common sense suggests that the general reception of the music to needs to be harmonious with the personal biases of the target audience. The music will additionally have to conform to the brands advertising strategy. The affective engagement potential of the music must be considered too. Even the situational effect determined by the placement of the advertisement[1] needs to be considered[1].

Ultimately, if music is inclusive within an advertisement it is (or certainly should be) with highly considered reason. Music strives to affect the audioviewer, yet any real quantification of effect is problematic due to the innumerate perceptual variables at play. These could be weighted and modelled, and would form an estimation of their influence on target populations. However, personal preferences and associations have significant and diverse effects within the expansive demographics of television advertisement exposure. To make quantifiable inferences regarding the effect of any creative decisions, variables must be measureable against features or responses that are statistically consistent within a population. Here we

[1] Within the advertisement block, program block, audience demographic for the time of day and so forth.

consider the physiological response to transient sensory stimuli as a suitably consistent human variable to investigate TIP effects. The major benefit here is that the evaluation is exclusive from the confounding of personal bias, exploring the foundation of multimodal interaction in audiovisual composition. This model does not attempt to account for all perceptual variables in modelling television advertisement perception. It considers the conundrum of composing musical-visual interactions by looking at the effect of physiological variables.

3. T.I.P AND BASIC EXPOSURE

The repercussions of audiovisual TIPs in television advertisements can theoretically be divided between the functional and the aesthetic. Modelling all the perceptual facets that construct an audioviewer's response would require a detailed integration of numerous fields of expertise including but not limited to advertising theory, consumer psychology, film theory and aesthetics. One commonality is present between all these elements in that they are all higher-level functions of the perceptual system during stimuli interpretation. An accredited model of interpretation for television advertisements or audiovisual media as a whole is not present in the literature. However, parallels can be drawn from the interpretation of other complex message carrying media, such as art as a generalised field.

Describing a 'Psycho-Historical Framework for the Science of Art Appreciation' Bullot and Reber [2] compartmentalise three steps in a hierarchical model detailing 'Modes of Appreciation' that are transferrable to audiovisual media interpretation.

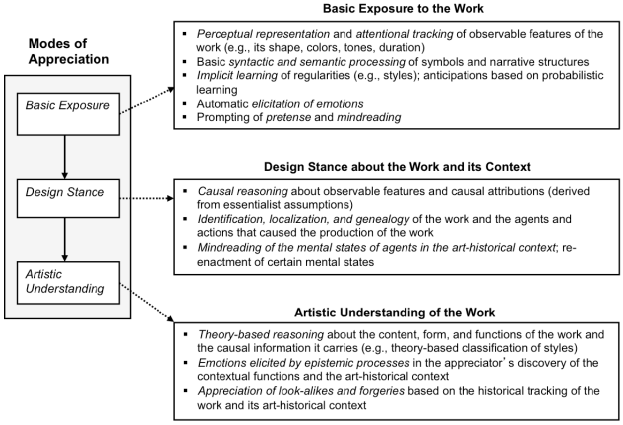

Figure 1. Bullot & Reber's three modes of art appreciation.

Basic Exposure is the first and therefore fundamental step in the successive 'Modes of Appreciation' process, described as 'the set of mental processes triggered by perceptual exploration of an artwork without knowledge about its causal history and art-historical context.' The solid arrows are representative of required conditions. Therefore the Design Stance cannot be adopted by the subject prior to progression from Basic Exposure, and Artistic Understanding is only achieved if the previous two Modes of Appreciation are fulfilled. The dotted lines link to the mental activities elicited by each mode. These are listed sequentially.

The foundation and requirement prior to all other functions in this hierarchical structure is the first bullet point, 'Perceptual representation and attentional tracking of observable features of the work...' In summary and contextualisation to audiovisual montage analysis, this defines the audioviewer's response to the component features of the presentation without yet inferring these qualities to any external associations or concurrent interpretations. This primary exposure analysis is therefore disregarding any semantic, hermeneutic or (to generalise) any higher-level cognitive functions that are hierarchically consequent in stimulus interpretation.

At Basic Exposure, transient features take precedence.

4. QUANTIFYING TRANSIENTS

The human perceptual system does not respond with the same linearity of a computational analysis. Therefore, in quantifying the human response to a stimulus perceptually relevant models need to be applied. Deviations in visual luminance and auditory spectral centroid (often referred to as 'brightness') are employed here as the perceptually relevant denotations of transients.

4.1 Visual Luminance

Luminance is the perceived intensity of a weighted sum of red, green and blue (RGB) tristimulus primary components proportional to physical power. It is perceived intensity, as the value for luminance is subject to the varying sensitivity of the human visual system across different colour bands (in the same manner that hearing sensitivity is dependent on frequency and defined by Fletcher-Munson curves, luminosity is defined in the CIE 1931 colour space[2]). Utilising luminance as opposed to a brightness analysis (which would average the largest and smallest RGB channel values without accounting for the perceptual system) provides a more ecologically valid quantification of transience for human interpretation.

As luminance is not equal across each of the tristimulus components of the visual spectrum, a perceptual weighting is required and calculated by finding RGB values linearly (after an inverse gamma function) and weighting:

r=0.2126
g=0.7152
b=0.0722

[2] 'Brightness is defined by the CIE as the attribute of a visual sensation according to which an area appears to emit more or less light. Because brightness perception is very complex, the CIE defined a more tractable quantity luminance which is radiant power weighted by a spectral sensitivity function that is characteristic of vision. The luminous efficiency of the Standard Observer is defined numerically, is everywhere positive, and peaks at about 555 nm. When an SPD is integrated using this curve as a weighting function, the result is CIE luminance.' http://www.poynton.com/notes/colour_and_gamma/ColorFAQ.html#RTFToC3

The is calculated in Adobe After Effects conforming to the ITU-R BT. 709 coefficients [3] that define the standard for high definition television formats.
The change in luminance is the definition of the transient and is calculated using 'Magnum', an Adobe After Effects Script created by Lloyd Alvarez is a commercially available software that detects edits in video footage[3]. The script utilises the alpha channel to carry the luminance of the RGB component of the frames content[4]. This is done using the 'shiftChannels' component.
The Adobe blending mode 'CLASSIC_DIFFERENCE' is applied as one frame is shifted for comparison relative to the other. In this application we can determine this as the luminosity change being calculated by taking the value of the current frame and subtracting the value of the previous frame.
The Magnum script returns a numerical value frame by frame for the difference of the luminance calculated in the difference blend that can then be used for statistical analysis.

4.2 Auditory Spectral Centroid

There are several psychophysical variables as candidates for quantifying transients within an auditory stream. Variations in the loudness of an auditory signal is a strong candidate for highlighting transients, and sudden deviations of loudness have been shown to be highly transient (reference). But due to the compression of audio and the multitude of listening conditions, reproduction fidelity and presentation volume in televisual advertising as well as a desire to evaluate the structural characteristics of music rather than just peaks in auditory signal (which are easily identifiable with conventional audio analysis tools) spectral centroid is preferred as the methodology[5].
The spectrum of the audio to be analysed is input and converted to a single numerical quantification of the spectral centroid at the chosen sampling instance within the audio stream. Utilising a MAX MSP patch designed by Tristan Jehan [4] which uses Fast Fourier Transforms for perceptual analysis structured upon research by Grey and Gordon [5], the Patch samples the audio at the frame rate of the video[6] thus attaining a frame accurate value.

[3]The script code can be found at `http://www.james-cheetham.com/Downloads/Tools/Magnum-The%20Edit%20Detector.jsx`

[4] The alpha channel is redundant in this application as we are gaining the luminance value via a summation of the RGB component.

[5] Jehan highlights that 'attack quality (temporal envelope) and spectral flux (evolution of the spectral distribution over time)' are potentially equally important at identifying the transient points within an auditory signal and this could be integrated into the model in future work.

[6] For example, for 25fps video the sampling rate would be every 40ms. This sampling rate is slow for auditory signals, but as the interest of analysis concerns frame level accuracy at this stage it is an appropriate figure (additionally, in most commercial audiovisual editing software packages frame level accuracy is the minimum of controllable resolution). If auditory and visual elements were to be analysed using the proposed transient detection methodologies provided here prior to coupling, a higher sampling rate may be appropriate to determine a more accurate a picture of transient structure for use in composition.

It is the difference in these values on a frame by frame basis that indicates the transient degree of the change i.e. a large jump from one value to the next marks a significant deviation in spectral centroid, which is perceived as transient to the perceptual system.

4.3 Comparing Audiovisual Transients

A frame by frame analysis outputs data which is graphed such as the example below (Amazon Kindle, Fig. 2) where the blue line represents loudness (auditory), the black line is spectral centroid (auditory) and the red follows the luminance differences (visual). These are represented on a frame by frame basis along the x-axis and subject to independent scales on the y-axis that are not normalised to either other. The relationship between the independents is difficult to ascertain visually[7]. Transients are visible in all analysis procedures (notably in the luminance analysis around a quarter of the way into the clip and auditory peaks and troughs towards the end of the clip), but the discrete figures and their directionality are confusing for interpretation[8].

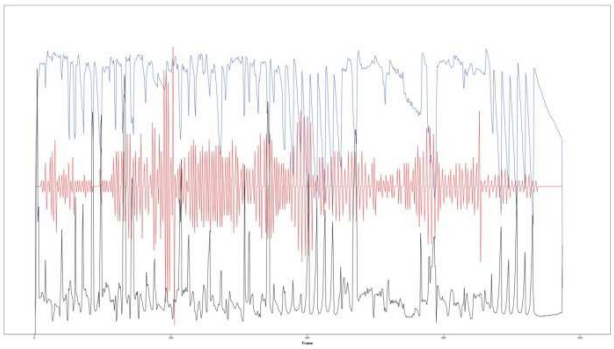

Figure 2. Amazon Kindle, Luminance (red), Brightness (black) and Loudness (blue) by Frame.

5. STATISTICAL TRANSIENT CATEGORISATION

The very definition of a transient is that it is a significant deviation from a neutral position. Statistically, we can quantify this by calculating z-scores which standardises the data set for the advert being analysed; expressing the scores in a standard distribution with a mean of zero and a standard deviation of one[9]. An option is to then catego-

[7] Animations including the audiovisual elements have been created to assist interpretation. This link (`http://youtu.be/ZSQnC38x4Pc`) demonstrates the Luminance/Loudness/Brightness transient analysis for Southern Comfort's 'Karate' advertisement. This link (`http://youtu.be/9u5dBsxVfiI`) runs the same analysis but at half speed for further ease in data interpretation.

[8] It should be noted that a sudden lack of sound is as transient as a sudden increase in sound. Likewise a luminance change in either direction can be equally transient – hence the positive and negative fluctuations are considered equally transient.

[9] This normalises the data values into a comparable range (as mentioned earlier, luminance and spectral centroid are defined on different, incompatible scales) – a further positive is that it neutralises the difference

rise these scores within defined statistical windows. This defines a level of significance for the transients based upon the sequence where the transients are relative to the advertisement sequence, rather than other advertisements[10].

In a normal statistical distribution we would expect to see 95% of cases with an absolute value below 1.96, with the remaining 5% above. In the Amazon Kindle example which is typical of all adverts the methodology has been tested on, 92.4% of cases fall within this range, leaving 7.6% outside, including 3.3% in the highest statistical bracket (Absolute z-score greater than 3.29). In this situation these extreme cases are exactly what we are looking for. This procedure marks transients within the sequence.

Figure 3 demonstrates the result of transient z-score classification. Frames are sequences along the x-axis while the height of the impulse on the y-axis marks the transient level as a statistical measure within 4 groups:

- Absolute Z-score = <1.96 coded as zero (therefore does not register on graph)
- Absolute Z-score = >1.96 = 1 (smallest impulse)
- Absolute Z-score = >2.58 = 2
- Absolute Z-score = >3.29 = 3 (highest impulse)

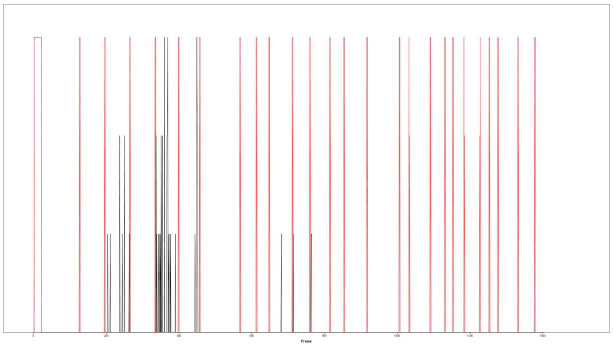

Figure 3. McDonalds advertisement, Z-Score Auditory Spectral Centroid (Black) and Luminance (Red). Values by Frame.

6. TIP TRENDS IN ADVERTISEMENTS

The methodology outlined is useful in the analysis of all audiovisual TIPs. Trends can be explored by genre, director, target audience, product category or any number of audiovisual variables. In addition to analytical processes it can be used in the composition of audiovisual media by auditioning the TIP effect of multiple musical-visual combinations to determine an appropriate pairing for the purpose of the media.

For example, in a sample of contemporary advertisements analysed (n = 10, airing from 2012 to 2014, ranging from 20 to 60 seconds in length at 25 frames per second, all with purely musical soundtracks i.e. no sound effects, dialogue or other layered audio[11]) we can elaborate on typical TIPs by advertisement concept.

Partitioning the advertisements into two categories, conceptually dominant (for example, technology products and lifestyle products such as perfume and alcohol choose to demonstrate a brand ethos as opposed to the literal representation or description of the product) as opposed to product dominant (for example children's toys or cleaning products, these show the usage and functionality of the sale item) [6] a clear difference in the compositional ideology is distinguished via TIPs.

Conceptually dominant advertisements rely on content, and are therefore less reliant on audiovisual montage.

Product dominant advertisements rely on demonstration, and thus have higher cut rates in the montage which is typically reflected in the choice of musical accompaniment via upbeat, higher tempo selections.

Therefore it is unsurprising that we see much larger temporal gaps between transient peaks in conceptual advertising as opposed to product-based advertising. The TIP is therefore more dispersed for product advertising than conceptual advertising.

Analysis	**Rate (per minute)**
Luminance Transient Rate (Conceptual)	33.5
Luminance Transient Rate (Product)	40.3
AV Transient Rate (Conceptual)	61.4
AV Transient Rate (Product)	88.2

A commonality between all advertisements is the tendency to cluster auditory transients within a defined region of the advertisement[12]. This would appear to be an effective strategy for orienting attention to key moments within the advertisement. Figure 4 shows an extreme example of this, where all auditory transient information is delivered at the end of the clip. This is an audiovisual hitpoint in relation to the delivery of the brand logo, but fails to register a visual transient at this instance. This TIP composition therefore chooses to separate the auditory and visual transients, which may have perceptual efficacy benefits due to modality switching latencies in multimodal integration [7].

between positive and negative fluctuations so transients in both directions are coded positive.

[10] Transient comparison across multiple advertisements is difficult due to the variations in visual and auditory composition and compression/broadcast methods. This research is concerned with improving advertising structure within itself, rather than directly improving advertisement efficacy relative to other advertisements. However, this would hopefully be a reciprocal outcome.

[11] Examples were from major international advertising campaigns including Amzon Kindle, Southern Comfort and Microsoft Windows.

[12] The percept of a transient depends on the plateau level of the stimulus as the transient threshold is highly adaptable in short time frames. This should therefore be integrated into further revisions of TIP models.

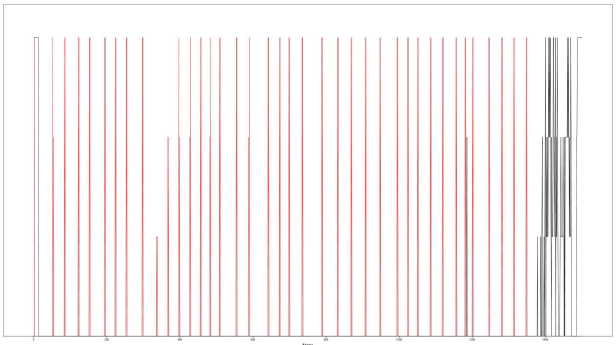

Figure 4. BSkyB Advertisement shows a separation of audiovisual transients in its TIP with steady distribution of visual transients and strong clustering of auditory transients at the end of the clip.

Certainly the most significant trend is the lack of transient synchrony being utilised in television advertisements would suggest that the technique is being avoided. Yet even those clips where visual elements have been constructed to reflect the musical soundtracks narrative[13] via the synchrony of musical-visual elements fail to register these on significant perceptual instances of the TIP on the majority of occasions. This may be due to the desire to avoid 'Mickey Mousing'[14] or other musical-visual clichés to maintain the integrity of the advertisement. However, synchrony does not have to create clichéd effects and disregarding the perceptual benefits of transient elements is to the detriment of the advertisements efficacy. One such component of the advertisements rationale, is the recall of information held within.

7. PSYCHOPHSICAL EVALUATION WITH PERCEPTUAL IMPLICATIONS

To understand the implications of transient interactions and particularly the synchrony of audiovisual transients an experiment was designed to incorporate one of the key targets of advertising, implicit memory of visual information. It was hypothesised that by displacing the synchrony between musical and visual transients, measurements of implicit memory recall would be effected. We hypothesised that placing auditory transients prior to visual transients would increase recall due to attention priming [8].

Ten participants (average age 23, 3 female) participated in a word fragment completion task to assess implicit memory recall of information delivered within a video sequence. Thirty-six words (twelve for each of the three following conditions – auditory and visual transients in sychrony, audio displaced prior to visual transient by four frames and audio displaced after visual transient by four frames) from the Word Fragment Completion Set created by Anderson [9] were presented on a screen with music from the Southern Comfort 'Beach' advertisement ('Gotta be 'Me' by Odetta). Subjects believed they were participating in a beat recognition task and were instructed to tap along on a keyboard to their interpretation of the rhythm of the music as a distraction task. After viewing the clips the subjects were asked to complete the word fragments.

As hypothesised, an increase in implicit memory recall was noted following the early onset of auditory transients prior to the visual transients where the words were presented. Early onset returned a 36% memory recall compared to transients in synchrony at 24% and transient delay at 28%.

Conclusively TIPs have an effect on higher-level perceptual features as demonstrated here through implicit memory recall. By analysing the relative timing of audiovisual transient delivery, TIPs can be designed to increase the likelihood of implicit memory recall. Here a simple relationship is determined, showing that the delivery of an auditory transient, prior to the visual transient containing the information for implicit memory increases recall. Synchronous transients relatively inhibit recall by a large degree, and auditory post transient structures to a lesser degree. Further experimentation is needed to determine the effects beyond the range of eight frames (200ms) tested here.

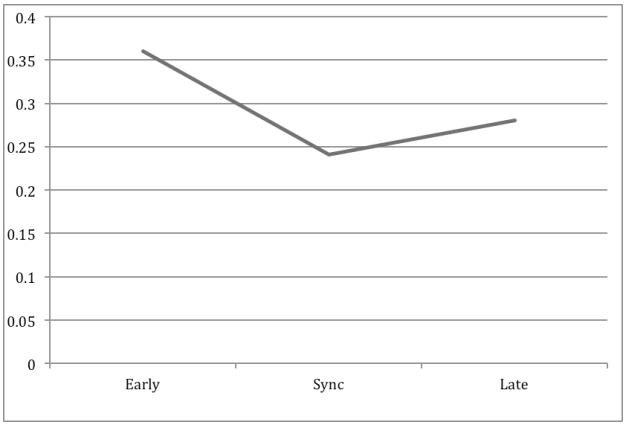

Figure 5. Level of visual implicit memory recall against the displacement of auditory transients relative to visual information onset.

8. CONCLUSIONS AND IMPLICATIONS

Demonstrating transient structures in a perceptually valid manner is the first step in understanding effective TIP composition.

Music is a common feature of audiovisual material, often used as a structural component of the montage, with visual rhythms deliberately constructed to reflect the musical tactus. Such compositional techniques are commonplace in all audiovisual mediums, and especially prevalent in shorter formats such as television advertisements which are the focus here.

Figure 6 demonstrates how musical and visual elements interact dependent on their alignment, and Figure 7 demonstrates how a simple displacement of the auditory track rearranges the musical and visual elements, thus

[13] Examples of such adverts include the McDonald's 'Symphony' (http://www.youtube.com/watch?v=F2ig42bfVdw) and United Airline's 'The Meeting' (http://www.youtube.com/watch?v=nU5DasW5nUY).

[14] 'Mickey Mousing' is a term used to describe mimicry in the musical soundtrack to visual elements. Often used in a negative manner due to its cliché status consequent of overuse in the 1930's and 40's.

creating a new audiovisual composition, with the same initial auditory and visual content.

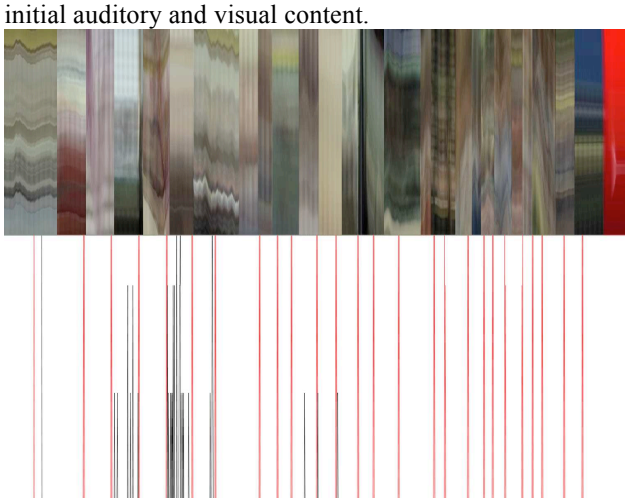

Figure 6. A movie-barcode (strip of pixels extracted from each frame of video and arranged linearly) with the TIP analysis (Spectral Centroid (Black) and Luminance (Red)). Visually we can see clear contrasts between the montage elements in the movie-barcode, and similarly definable sections created by contrasts in the TIP.

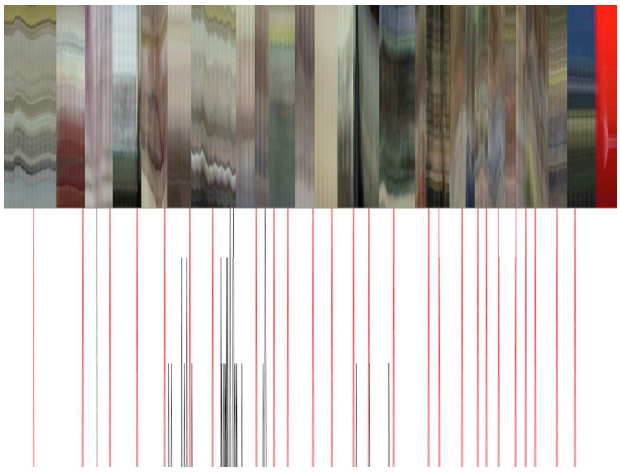

Figure 7. Here the audio has been displaced relative to the visual. Thus a new TIP compared to that in Figure 1 is created while the content of both clips has remained as consistent as possible.

As a creative in the audiovisual composition process, the effects of such displacement should not be taken lightly as although the displaced clips are exceedingly similar in passive reception, there are strong implications for audioviewer interpretation.

At Basic Exposure transient factors of the audiovisual montage take precedence, creating attentional distributions that construct the foundation of an audioviewer's consequent interpretation of the production. This effect falls to the functional aspect of the aforementioned aesthetic-functional divide. To qualify as a functional response it must return a measurable outcome over a population that can be replicated by adhering to a specified design principle. This approach could be labelled as Neuroaesthetics[15], an attempt to interpret the brains predisposition for certain forms and constructs that succeed over others in their purpose for a measured reason. Neuroaesthetics considers the human appraisal as consistent, a blanket response disengaged from individualities influence. A philosophical appraisal would highlight this omission of individuality as a confounding variable in modelling perception, the argument being that not merely the innate, but also the personalised and enculturated behaviours of interpretation and conscious cognition of stimuli are the key to our percept of the material. And indeed they are, which is why great effort and expense can surround advertising campaigns. But, from the standpoint of a music editor, who is uninvolved with the creation of the semiotic contents in either visual or musical modalities, yet charged with the duty of integrating the two in the most effectual manner, TIP are the key concern.

Digital Audio Workstations allow for highly synchronous musical-visual interaction, tailoring musical transients to key visual 'sync-points' via manipulation of tempo tracks which can be adjusted – pulled like elastic pinned by point of synchrony. The concern of the audiovisual composer is how to effectively compose these transients to the benefit of their production. Crucial narrative aspects of the visual and musical content can be identified as requiring heightened attention. For example, a key delivery point within an advertisement would be the presentation of the brand's name and logo. A key musical aspect could be the transition from one repeating musical motif to a new motif. To emphasize these key points – a music composer may choose to increase the salience of the downbeat of the transition – potentially through the addition of musical accent, or the addition of a crash cymbal as a simple example. Ultimately, the transient level of a new section is determined by its distinction from the previous. The same is true visually; the orientation of attention is modulated by the signification of change via contrast in the montage sequence i.e. a transient difference from one frame to the next. Mapping such contrasts enables an evaluation of the physiological factors with statistical prominence via their rarity tied to Basic Exposure. Beyond this numerous questions arise for further research. What are the effects of transient synchrony on cognition? Is this effect genre dependent? What are the boundaries of asynchrony, and how does this alter the TIP effect?

Further research will investigate improving the efficacy of audiovisual media interpretation via control of the transient structure in musical and visual content interaction.

9. REFERENCES

[1] L. G. Craton and G. P. Lantos, "A model of consumer response to advertising music," *J. Consum. Mark.*, vol. 29, no. 1, pp. 22–42, 2012.

[15] The term 'neuroaesthetics' was coined by Zeki and Lamb [10] and it denotes the biological factors in aesthetic interpretation

[2] N. J. Bullot and R. Reber, “The artful mind meets art history: Toward a psycho-historical framework for the science of art appreciation,” *Behav. Brain Sci.*, vol. 36, no. 02, pp. 123–137, 2013.

[3] P. ITU-T RECOMMENDATION, “Subjective video quality assessment methods for multimedia applications,” 1999.

[4] T. Jehan, “Creating music by listening.” Massachusetts Institute of Technology, 2005.

[5] J. M. Grey and J. W. Gordon, “Perceptual effects of spectral modifications on musical timbres,” *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1493–1500, 1978.

[6] C. Bullerjahn, “The effectiveness of music in television commercials,” in *Music and manipulation: On the social uses and social control of music*, 2006, pp. 207–235.

[7] T. Koelewijn, A. Bronkhorst, and J. Theeuwes, “Attention and the multiple stages of multisensory integration: A review of audiovisual studies,” *Acta Psychol. (Amst).*, vol. 134, no. 3, pp. 372–384, 2010.

[8] N. Cason and D. Schön, “Rhythmic priming enhances the phonological processing of speech,” *Neuropsychologia*, vol. 50, no. 11, pp. 2652–2658, 2012.

[9] C. A. Anderson, N. L. Carnagey, and J. Eubanks, “Exposure to violent media: the effects of songs with violent lyrics on aggressive thoughts and feelings.,” *J. Pers. Soc. Psychol.*, vol. 84, no. 5, p. 960, 2003.

[10] S. Zeki and M. Lamb, “The neurology of kinetic art,” *Brain*, vol. 117, no. 3, pp. 607–636, 1994.