An Audio-Driven System For Real-Time Music Visualisation

Max Graf, Harold Chijioke Opara, Mathieu Barthet
Centre for Digital Music, Queen Mary University of London
Correspondence should be addressed to Max Graf (max.graf@qmul.ac.uk)

Disclaimer: this paper is an Author's Accepted Manuscript, the published version is available at https://www.aes.org/e-lib/browse.cfm?elib=21091
Main reference: *M. Graf, H. Chijioke Opara, and M. Barthet, "An Audio-Driven System for Real-Time Music Visualisation," Paper 10498, (2021 May.).*

***Abstract*—Computer-generated visualisations can accompany recorded or live music to create novel audiovisual experiences for audiences. We present a system to streamline the creation of audio-driven visualisations based on audio feature extraction and mapping interfaces. Its architecture is based on three modular software components: backend (audio plugin), frontend (3D game-like environment), and middleware (visual mapping interface). We conducted a user evaluation comprising two stages. Results from the first stage (34 participants) indicate that music visualisations generated with the system were significantly better at complementing the music than a baseline visualisation. Nine participants took part in the second stage involving interactive tasks. Overall, the system yielded a Creativity Support Index above average (68.1) and a System Usability Scale index (58.6) suggesting that ease of use can be improved. Thematic analysis revealed that participants enjoyed the system's synchronicity and expressive capabilities, but found technical problems and difficulties understanding the audio feature terminology.**

I. Introduction

Although the primary mode of consumption of music is auditory, efforts have been made to translate musical expression to other domains [1]. For people with normal or corrected-to-normal vision, visual representations can lead to deeper insights about musical expression.

This work is concerned with computer-generated visualisation of audio signals in a 3D visual environment. Current tools to produce music visualisations do not explicitly leverage advances made in the music information retrieval (MIR) field. We investigate methods for interactively mapping audio features (numerical data representing signal and perceptual attributes of sound/music obtained computationally) to visual objects to facilitate music visualisation. The main contributions of this work are as follows: First, we propose a method to enable users to create visualisations from audio signals routed to a digital audio workstation (DAW), and a modular software system called "Interstell.AR", which is versatile and flexible and can be used with many game engines. Embedded in a DAW, a source can be, for example, an instrument, a singer's voice, or a complete recording. Second, we provide design insights based on a user evaluation conducted with 34 participants that can inform the design of multimedia systems combining audio and visual modalities.

As a tool to facilitate the creation of new audiovisual media, the proposed system could be of interest to musical artists, graphic designers, and VJs (visual jockeys). Applications range from the production of live video-based music streaming, visuals for live music, augmented/virtual/mixed reality experiences, games, etc. Other applications include music education, e.g., to learn about musical attributes or a performer's musical expression using visual correlates.

Understanding the role of music visuals has been the object of studies in performance reception and experience and new interfaces for musical expression, some of which specifically focus on electronic music [2]. In live electronic music performance, often the gestures of performers do not provide explicit connections to the sound production mechanisms due to the virtualisation of musical instruments (use of synthesis and digital music interfaces); accompanying visuals may be used to compensate for the lack of visible efforts from performers [3], so as to enrich the audience experience. The wide variety of timbres in recorded electronic music makes it difficult, if not impossible, for listeners to relate to existing musical instruments; by establishing connections between the auditory and visual domains, music visuals experienced while listening to an (electronic) music recording may help listeners to develop their internal imagery and associations with sounds.

Compared to previous works, the proposed system uses a wide range of audio features and leverages audio engineering techniques to increase creative possibilities. Whereas related systems rely on relatively simple audio representations to create visualisations, our system supports a wide range of audio features (cf. section III-B1). Although this broad palette of audio features might be overwhelming for some novice users (as reflected in the user evaluation), the results discussed in the paper suggest that it can support creativity favourably for users, especially if they have some prior experience with audio processing. While feature-based music visualisation tools exist, they require explicit knowledge about coding (e.g. feature extraction and visualisation with Max/MSP and Jitter, Processing, etc.), whereas our system can be used by people with no coding experience. The insights from our study may thus be useful for future research in this domain.

The results from the user evaluation conducted for a specific song indicate that user-generated mappings yielded a significantly higher measure of complementarity between

audio and visuals judged by participants, in comparison to a control condition with no mapping. The system was found to be moderately easy to use with a mean system usability scale (SUS) [4] measure of 58.6, indicating that certain aspects of the system may be too complex. The system obtained a mean creativity support index (CSI) [5] of 68.1 showing a good aptitude for exploration and expressiveness whilst the sense of immersion in the creative activity still needs to be improved. Qualitative feedback analysed through thematic analysis [6] put forward that users liked the creative possibilities offered by the system and the synchronicity between audio and visuals in the resulting computer-generated graphics. However, issues around usability and the complexity to understand the meaning of audio features suggest that there are aspects to improve to better support users who are not familiar with MIR.

II. Related Work

Music visualisation. Torres & Boulanger presented an agent-based framework that produces animated imagery of three-dimensional, videogame-like characters in real-time [7]. Characters in this framework were designed to respond to several types of stimuli, including sound intensity. In a subsequent study [8], the authors used this framework to formulate a digital character's behaviour that responded to the emotionality of a singer's voice. Selfridge & Barthet investigated how music-responsive visuals can be experienced in augmented/mixed reality environments for live music [9]. However, their system does not enable audio feature or visual environment customisation, which is a feature of this work (see Section III).

Nanayakkara et al. presented an interactive system designed to rapidly prototype visualisations using Max/MSP and Flash [10]. Their visualisations were based solely on Musical Instrument Digital Interface (MIDI) signals. Although radically reducing the dimensionality of possible mappings by limiting the data types to trigger events and MIDI note numbers, this work is interesting for highlighting the use of MIDI data in the visualisation process.

Kubelka devised a hybrid system based on a combination of real-time and pre-generated visualisation techniques [11]. It operates directly on the audio stream and maps characteristics of music to parameters of visual objects. What is notable in this study is the idea of creating a separate component for the interactive design of visualisations (a scene editor), since the majority of existing works stipulate pre-configured mappings between audio and visual properties that are not to be changed at runtime.

Music visualisation was also applied in the context of music production and learning. McLeod and Wyvill created a software that accurately renders the pitch of an instrument or voice to a two-dimensional space, allowing a musician or singer to evaluate their performance [12]. This example illustrates the analytic possibilities of visualisation, which become particularly apparent when viewed in educational contexts.

Systems such as Magic Music Visuals [13] and Synesthesia [14] represent the most recent generation of commercial audio visualisation software. One particularly notable system is the ZGameEditor Visualizer, a toolkit integrated into the FL Studio DAW. It directly incorporates the visualisation process into the environment of the DAW. Those commercial systems are sophisticated tools with regard to their visualisation capabilities. However, they rely on a small number of audio features, which may limit the creative potential of the visualisations. Visual programming frameworks for music visualisation such as TouchDesigner[1], Max/MSP Jitter[2] or VSXu[3] do provide extensive audiovisual processing capabilities, but require significant (visual) programming skills or the use of pre-configured visualisations.

Our system differs from the previous works insofar that it can be directly integrated with a DAW. Audio engineering techniques are employed to process the incoming audio signal (cf. III-B1) and provide an interactive, fully configurable mapping routine for connecting sounds to visuals. This allows for sophisticated control over the source material per se, and also the way this source material affects the visuals.

Mapping. *Mappings* describe here virtual connections that define the behaviour of visualisations in response to audio features extracted from the music. Hunt et al. argue that the connection of input parameters to system parameters is one of the most important factors when designing digital music instruments (DMIs) [15]. This consideration is worth studying for audiovisual interaction based on procedural processes. In the system presented in section III, audio features act as input parameters reacting to sound changes. When updated, they trigger a change in the system (a change in visual properties in our case). As such, the design and quality of mappings is one of the system's core considerations. The Open Sound Control (OSC) specification [16] is one of the most popular ways of creating such mappings. Our system has been designed using Libmapper [17], an open-source, cross-platform software library based on OSC. The defining feature of libmapper is that mappings can be manipulated through a dedicated visual interface, while the system is actively running.

III. Design of the System

A. Design objectives

The production of visuals accompanying music is inherently a creative task which must take into account considerations from a number of stakeholders, e.g., artists, producers, labels. This work investigates how to design assistive tools for visual artists creating content aimed at accompanying music media. The end product is shaped by human factors (style, creative intent, a band's image and "universe", etc.), hence we do not target fully automatic music visualisation renderers here (e.g. the Winamp music visualisers[4]). With this in consideration, we pose two

[1] https://derivative.ca/product
[2] https://cycling74.com/products/max/
[3] https://www.vsxu.com/about/
[4] http://www.geisswerks.com/milkdrop/

design objectives (DO); DO_1: The system should allow music producers and visual artists to create responsive visualisations based on the sonic and musical attributes from individual musical instruments, groups of instruments, or the mix as a whole. DO_2: The system should be intuitive - the process of creating mappings should be mainly visual, without extensive need of coding.

B. Implementation

1) Backend - audio plugin: The backend is a DAW audio plugin implemented in C++ using the JUCE framework [18]. Its central task is audio feature extraction. Based feature extraction review by Moffat et al. [19], we used the *Essentia* software library [20], as it can run in real-time and offers a large number of audio features, giving users a broad palette of possible starting points for their desired visualisations with the system. By applying the plugin to one or more tracks, an arbitrary number of backend instances can be connected to the mapping space. This provides the flexibility required by DO_1. In the following, we give a detailed explanation of the different aspects of the backend component.

Feature extraction: Feature extraction is at the core of the plugin. Several audio features are computed globally, i.e. for whole, unfiltered chunks of incoming audio signals:

Loudness is an important aspect of musical expression. It provides an empirical estimate for the perceived intensity of a sound. Given that rhythmic elements can be contained in the source material, it can be employed to translate the pulse of a musical piece to periodic movements in the visualisations, for example.

Pitch is computed using the YIN algorithm [21]. It is one of the central features that allow listeners to distinguish between different components in music. In certain contexts it implicitly conveys aspects of emotionality [22]. This feature is most useful when the plugin is applied to isolated source material, such as an instrument or a vocal part.

The **spectral centroid**, a signal descriptor modeling the brightness perceptual attribute of a sound. It represents the frequency spectrum barycenter and has been shown to be an important correlate of timbral variations used to convey musical expression [23].

A simple **onset detection system**, based on the high-frequency content algorithm [24] is included. Its purpose is to recognise rapid changes in the music to identify the onset of notes or percussive instruments, which can be mapped to visual attributes.

Lastly, the **sensory dissonance** of the signal is included in the global calculations. It measures the tension resulting from instrumental tones at a given point in time, based on the spectral peaks of the audio signal.

Sub-bands: In order to support DO_1, the system provides an option to split the incoming audio signal into up to three frequency sub-bands. These are created using second-order low- and high-pass filters respectively. Each sub-band provides a number of *feature slots*, which allows for the injection of feature-computing algorithms into the given band. This is useful when the audio is a mix of several elements, e.g., a complete song. sub-bands allow the user to isolate certain parts of the signal, for example the bass on one end of the frequency spectrum and drum cymbals on the other end.

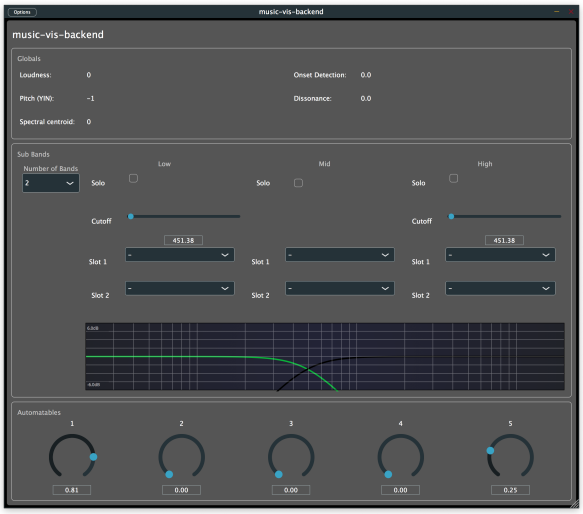

Fig. 1. Screenshot of the backend GUI

Automatables: To extend the range of possible mappings, we also incorporate user-controllable metadata into the data stream. They allow users to create mappings controlling visualisations independently of the audio feature data. By utilising a feature provided by libmapper, automatables can be combined with other mappings. This gives users fine-grained control over the way a given audio signal affects a visual property.

Graphical user interface: A GUI was created to visualise audio feature numerical data computed by the plugin and provide control over the sub-bands and automatables. Figure 1 shows a screenshot of the backend GUI.

2) Frontend - visualisations: The Unity[5] platform was selected to create the frontend due to its accessibility, extensibility as well as its potential to function on virtual and mixed reality devices. The frontend is designed so that the visualisation engine is not bound to a specific platform. A software interface was created to integrate libmapper into the Unity platform. This template serves as the base class for all visual components that send or receive libmapper signals.

3) Middleware - libmapper: The third cornerstone of the system is the mapping framework. Mappings between audio features and visual objects can be pre-configured, but also changed at runtime. This facilitates the creation of new roles in the context of audiovisual performances, as imagined by Bain [25]. Libmapper offers a visual interface with detailed views of signals and their connections. To support DO_2, we employ this browser-based GUI to make designing dataflows from the frontend to the backend a fully visual experience that does not require coding skills. For users with basic coding skills, it offers the possibility to apply mathematical operations to signals. For

[5]https://unity.com/

the majority of audio features extracted by the backend engine, different characteristics of instruments, performers' musical expression and mix audio qualities will create differences in the numerical values output by the system. By applying operations to the data passing through the connections, signals can be tailored to an intended range or subjected to linear or non-linear transformations, for example.

IV. Evaluation

A. Procedure

We conducted an online study to evaluate the prototype. It was structured into two tasks. Task 1 involved the observation and commentary of music videos showing visualisations produced by the system. Participants were asked to rate how well the visuals complemented the music in the videos by rating their agreement level to a 10-point Likert item (from "not well at all" to "extremely well"). They were also asked to verbally describe positive and negative aspects of the visualisations. Task 1 had a suggested time limit of 15 minutes.

Task 2 of the study was an interactive user test. Participants were instructed to install the software on their own devices and completed a series of subtasks with the system, using a provided electronic music track. Subtasks included connecting certain audio features from the backend to certain visual properties of the frontend, as well as experimenting with the system on their own. For task 2, we suggested a time limit of 45 minutes. Please refer to section VIII-B for a full list of the survey questions.

B. Methods

To assess the visualisation capabilities of the system, we tested the following hypothesis: *"Visuals based on audio-driven mappings generated by users complement the music better than visuals that do not react to the music."*. We designed a three-dimensional scene for the frontend that served as the foundation of mappings for the user evaluation. Figure 2 shows a screenshot of this scene. A link to music videos generated with the system is included in section VIII-A. We recorded three music videos with visualisations produced by the system for a one-minute long song. The song consists of a hip-hop style drum beat, as well as several synthesizer elements for bass and lead sounds. Baseline automated graphic animations are incorporated into the visualisations for all three videos (e.g. slight movements of the stars). A music video only showing baseline animations with no audio-reactive mapping was used as control condition (M0). Two of the authors created mappings for the song independently. The mappings were added to the baseline animations provided in the control condition (M0) and yielded the two audio-driven mapping conditions, M1 and M2. The order of the three videos was randomised across participants.

We conducted a Friedman test [26] with the audiovisual complementarity as dependent variable and the mapping type as independent variable (three levels: M0, M1, M2). A post-hoc analysis accounting for multiple comparisons was necessary; we used the Wilcoxon signed-rank test [27] to test differences between mapping conditions (M0-M1, M0-M2, M1-M2) by using the Bonferroni correction (the p-value significance level was $0.017 = \alpha$ based on a Type I error $\alpha = 0.05$).

Fig. 2. A screenshot of a dynamic music visulisation generated with the frontend by combining several AV mappings

We also analysed participant feedback in task 1 by means of inductive thematic analysis based on the framework by Braun and Clarke [6]. The goal was to identify common themes regarding positive and negative factors considering the three videos created for task 1 of the user study.

We conducted a separate thematic analysis for the user study task 2, with the goal of finding common themes and idiosyncrasies in how users interacted with the system. We assessed the usability of the software components using the SUS [4]. It consists of a 10-item questionnaire, regarding topics such as complexity and ease of use. Finally, we integrated the CSI [5] into the study to assess the value of our system as a creativity support tool. It provides questions to measure aspects such as exploration, engagement and effort/reward trade-off.

C. Participants

34 participants took part in the study in total. No prerequisite skills were required for task 1. For task 2, we recruited participants who fulfilled certain software requirements to be able to install the software and had basic DAW skills. Participants were recruited using departmental mailing lists at our institution as well as through the *Prolific* platform. We applied a custom prescreening while selecting participants, with regard to their interests (music) as well as their academic/professional backgrounds (computer science, computing, engineering or music). However, subjects were not prescreened in terms of their familiarity with music visualisation/multimedia systems specifically. Their mean age was 24.2 years (SD=5.1). 70% were male, 30% female, two did not disclose their gender. The majority of them were from Europe or the United States. 22 of them were students. Task 1 was completed by all participants and Task 2 by nine participants.

V. Results

A. Quantitative Evaluation

The results of the Friedman test showed a significant difference in the audiovisual complementarity ratings of

TABLE I
STATISTICS OF THE WILCOXON TEST FOR TASK 1

	M0 - M1	**M0 - M2**	**M1 - M2**
Z	-3.563^{a}	-2.974^{a}	-0.705^{b}
p (2-tailed)	$< \mathbf{0.001}$	$\mathbf{0.003}$	0.481

[a]Based on negative ranks.
[b]Based on positive ranks.

videos based on the underlying mappings, $\chi^2(2, N = 34) = 15.6$, $p < 0.001$.

The results of the Wilcoxon signed-rank tests are listed in Table I. The results indicated significant differences between M1 and M0, M2 and M0, but not between M1 and M2. Mean and interquartile range values are 4.5 $(2.0-7.0)$ for M0, 7.1 $(5.8-8.3)$ for M1, 6.8 $(6.0-8.3)$ for M2. This shows that visuals produced with audio-driven mappings were found to better complement the music than visuals with no audio-reactive mappings.

B. Qualitative Evaluation

1) Thematic Analysis: Thematic analysis was conducted by two coders and the results were integrated. For task 1, 82 codes were extracted in total. The prevalent themes of answers obtained in task 1 were centred around the aesthetic quality of the visualisations (38 codes) and the connections between music and visualisations (28 codes). An in-depth analysis of the themes is omitted here for space reasons. For details, please refer to section VIII-C. The thematic analysis for user study task 2 is concerned with participants' experiences while actively using the system. The codes gathered from the answers were compiled into the following themes (code occurence numbers are reported in brackets):

Synchronicity (6): Overall, participants were satisfied with the system's synchronicity, both in terms of latency and mappings. One participant felt that they experienced low latency when attempting one of the subtasks. Participants stated that *"It followed the beat of the song well"*, *"i could obtain a music visualization that fit the audio"* and *"the light object in the middle of the visualisation lights up according to the music"*. One participant experienced delays between sound and visualisations (*"Connection issues. its a bit laggy too"*).

Expressiveness (6): The majority could express their creative intent with the system (*"The loudness and onset detection were my most favourite features since i was dealing with rhythmic music to test the system"*). The option to apply transformations to signals was well received. One participant stated: *"Really liked the section to add equations to the effect response"*. Several participants described their use of mathematical operations to transform mappings (*"[...] the camera movement was quite rapid but looked wonderful after diluting doing a y = 0.5*x"*, *"[...] result is pretty extreme, changing the mapping function to y=0.01*x works better."*).

Technical Issues (5): Four out of nine participants reported technical difficulties at some point during the experiment. One participant mentioned an intial unresponsiveness of the system (*"The interstellar app crashed a few times upon opening but ran stably if it managed to open"*).

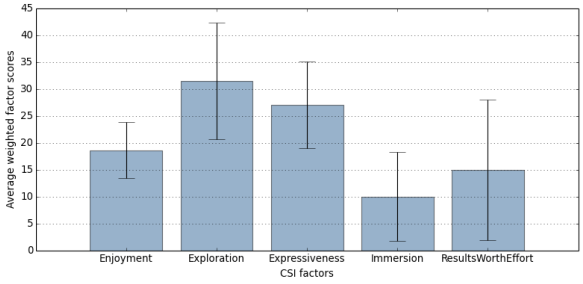

Fig. 3. Bar plot of the weighted average CSI factor scores

Ease of Use (4): Participants had a mixed experiences in terms of usability. Three participants lauded the intuitiveness of the system (*"It was very intuitive and fun to use"*, *"The connection was fairly simple to set up"*). Two participants reported that they experienced difficulty in using the system because they did not understand the terminology of certain audio features. They both highlighted that there was too much "tech jargon" and that the system should be more "user friendly". One participant claimed that they *"needed more guidance on what kind of mappings would work well"*.

2) Usability and creativity support: The sample size for the SUS and CSI analyses was 9. Our system obtained a mean SUS score of 58.6 (SD=15.7). This is below the average score of 68[6], but nevertheless shows that our system exhibits a reasonable level of usability.

The system obtained a mean CSI score of 68.1 (SD=13.2), which indicates that its facility as a creativity support tool is above average, but not excellent. Figure 3 shows the individual weighted scores of CSI factors. The CSI results revealed that the aspects most positively received were "Exploration", "Expressiveness" and "Enjoyment" (in decreasing order). The factors "Immersion" and "ResultsWorthEffort" were rated rather poorly in comparison.

3) Duration of use in task 2: As mentioned above, we imposed a (non-strict) time limit of 45 minutes for task 2. On average, participants worked with the system for 58.8 minutes (SD=32.6). Table II lists the durations of use and resulting SUS and CSI scores for each participant in task 2. We subjected these data to analysis to see whether there existed an interaction between the time that participants spent working on the system and their indicated usability scores. The working durations and SUS scores were not significantly correlated, Pearson's $r(7)$=.27, p=.47 (p¿.05, the type I error). The same was true for the working durations and CSI scores, Pearson's $r(7)$=.61, p=.07 (p¿.05, the type I error). This indicates that there was no linear

[6]https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

TABLE II
WORKING DURATIONS IN MINUTES (M), SUS SCORES AND CSI SCORES FOR EACH PARTICIPANT IN TASK 2

Duration (m)	SUS score	CSI score	PID
31	30.0	52.3	9
34	50.0	80.0	5
40	60	46.0	1
42	67.5	76.0	8
45	72.5	65.6	3
50	62.5	67.0	2
60	50.0	58.3	4
105	87.5	81.0	6
123	47.5	87.0	7

correlation between the time that users spent with the system and their given usability ratings.

C. Discussion and limitations

The results presented for task 1 of the study are limited by the fact that only one song was used. Although the results evidence a significant improvement compared to a baseline, future work should be conducted to assess whether the results generalise to other songs and other genres of music. It would also be worth comparing the proposed system against existing commercial and open-source software that can be used for computational music visualisation generation.

In order to improve usability, information and tutorials on audio features could be provided since most users would not be familiar with MIR, psychoacoustics and music perception (e.g., links to online support material could be provided in the plugin menu). Ways to simplify the audio-visual mapping process should be investigated e.g., by using abstractions hiding the complexity in the naming and potentially large number of audio and visual attributes, and/or interactive machine learning.

VI. CONCLUSION

We have presented a novel framework for interactively visualising sound and music. To our knowledge, this is the first system operating on the demonstrated level of interactivity. By enabling users to map every implemented audio feature to every exposed visual property, a broad range of possible visualisations is supported. We explained the system's design and implementation and discussed its application in various settings. The results of the quantitative and thematic analyses showed that music videos produced with audio-driven mappings were perceived to have a higher audio-visual complementarity than videos showing non audio-reactive visualisations. Future work could address the limitation of one song for task 1 of the user study. Analysing the audio-visual complementarity using songs from different musical genres may increase the explanatory power of the evaluation. The results of the usability and creativity support analyses showed that there is still room for improvement to integrate help on technical audio features and to make the application less computationally expensive to reduce audiovisual lags during production.

VII. ACKNOWLEDGEMENTS

This work has been partly funded by the UKRI EPSRC Centre for Doctoral Training in Artificial Intelligence and Music (AIM), under grant EP/S022694/1.

REFERENCES

[1] Olowe, I., Barthet, M., Grierson, M., and Bryan-Kinns, N., "FEATUR.UX: An approach to leveraging multitrack information for artistic music visualization," in *Proc. Int. Conf. on Technologies for Music Notation and Representation*, 2016.

[2] Correia, N. N., Castro, D., and Tanaka, A., "The role of live visuals in audience understanding of electronic music performances," in *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, pp. 1–8, 2017.

[3] Schloss, W. A., "Using contemporary technology in live performance: The dilemma of the performer," *Journal of New Music Research*, 32(3), pp. 239–242, 2003.

[4] Brooke, J., *"SUS-A quick and dirty usability scale." Usability evaluation in industry*, CRC Press, 1996, iSBN: 9780748404605.

[5] Carroll, E. A. and Latulipe, C., "The Creativity Support Index," in *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '09, p. 4009–4014, Association for Computing Machinery, New York, NY, USA, 2009, ISBN 9781605582474, doi:10.1145/1520340.1520609.

[6] Braun, V. and Clarke, V., "Using thematic analysis in psychology," *Qualitative Research in Psychology*, 3(2), pp. 77–101, 2006, doi:10.1191/1478088706qp063oa.

[7] Torres, D. and Boulanger, P., "The ANIMUS Project: A Framework for the Creation of Interactive Creatures in Immersed Environments," in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, VRST '03, p. 91–99, Association for Computing Machinery, New York, NY, USA, 2003, ISBN 1581135696, doi:10.1145/1008653.1008671.

[8] Taylor, R., Boulanger, P., and Torres, D., "Real-Time Music Visualization Using Responsive Imagery," 2006.

[9] Selfridge, R. and Barthet, M., "Augmented Live Music Performance using Mixed Reality and Emotion Feedback," in *14th Int. Symposium on Computer Music Multidisciplinary Research*, p. 210, 2019.

[10] Suranga Chandima Nanayakkara, Taylor, E., Lonce Wyse, and Ong, S. H., "Towards building an experiential music visualizer," in *2007 6th Int. Conf. on Information, Communications Signal Processing*, pp. 1–5, 2007, doi:10.1109/ICICS.2007.4449609.

[11] Kubelka, O., "Interactive music visualization," *Central European Seminar on Computer Graphics*, 2000.

[12] Mcleod, P. and Wyvill, G., "Visualization of musical pitch," in *In Proc. of the Computer Graphics Int.*, pp. 300 – 303, 2003.

[13] Color & Music LLC, "Magic Music Visuals," 2012, https://magicmusicvisuals.com/.

[14] Gravity Current, "Synesthesia," 2018, http://synesthesia.live/.

[15] Hunt, A., Dd, Y., Wanderley, M., and Paradis, M., "The Importance of Parameter Mapping in Electronic Instrument Design," *J. ofNew Music Research*, 32, 2002, doi:10.1076/jnmr.32.4.429.18853.

[16] Wright, M., "The Open Sound Control 1.0 Specification," Specification, OSC, 2002.

[17] Malloch, J., Sinclair, S., and Wanderley, M. M., "Libmapper: (A Library for Connecting Things)," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, p. 3087–3090, ACM, New York, NY, USA, 2013, ISBN 9781450319522, doi:10.1145/2468356.2479617.

[18] Storer, J., "JUCE: Jules' Utility Class Extensions," 2020, https://github.com/juce-framework/JUCE.

[19] Moffat, D., Ronan, D., and Reiss, J., "An Evaluation of Audio Feature Extraction Toolboxes," in *DAFX*, 2015, doi:10.13140/RG.2.1.1471.4640.

[20] Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X., "ESSENTIA: An Open-Source Library for Sound and Music Analysis," in *Proceedings of the 21st ACM Int. Conf. on Multimedia*, MM '13, p. 855–858, Association for Computing Machinery, New York, NY, USA, 2013, ISBN 9781450324045, doi:10.1145/2502081.2502229.

[21] Cheveigné, A. and Kawahara, H., "YIN, A fundamental frequency estimator for speech and music," *The J. ofthe Acoustical Society of America*, 111, pp. 1917–30, 2002, doi:10.1121/1.1458024.

[22] Jaquet, L., Danuser, B., and Gomez, P., "Music and felt emotions: How systematic pitch level variations affect the experience of pleasantness and arousal," *Psychology of Music*, 42, pp. 51–70, 2014, doi:10.1177/0305735612456583.

[23] Barthet, M., Depalle, P., Kronland-Martinet, R., and Ystad, S., "Acoustical correlates of timbre and expressiveness in clarinet performance," *Music Perception*, 28(2), pp. 135–154, 2010.

[24] Masri, P. and Bateman, A., "Improved Modelling of Attack Transients in Music Analysis-Resynthesis," in *ICMC*, 1996.

[25] Bain, M. N., *Real Time Music Visualization: A Study in the Visual Extension of Music*, Ph.D. thesis, The Ohio State University, 2008.

[26] Friedman, M., "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *J. ofthe American Statistical Association*, 32(200), pp. 675–701, 1937, doi:10.1080/01621459.1937.10503522.

[27] Wilcoxon, F., "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, 1(6), pp. 80–83, 1945, ISSN 00994987.

TABLE III
QUESTIONS FOR TASK 1 OF THE SURVEY

Visualiser 1, 2 and 3 (separately for each video)
What did you enjoy about the viewing experience?
What did you dislike about the viewing experience?
How well do you think the visualisation complemented the audio?
What are the reasons for your answer to the question above?
General Questions
Would you use a system such as Interstell.AR to produce music visualisations and if so, in which context(s)?
We want to improve Interstell.AR! Please report any ideas or recommendations you may have on how to improve the experience and/or what you would be interested in doing with the system.
Which headphones/earphones did you use?
Demographics and Personal Questions
Gender?
Age?
Please indicate your occupation:
Please indicate your nationality:
Do you have any hearing impairment? If so, please specify which, if you wish.
Do you have any visual impairment? If so, please specify which, if you wish.
How would you describe your experience as a musician?
How would you describe your experience as an audiovisual artist?

VIII. APPENDIX

A. Links to the videos created with the system

Videos of the three conditions created for task 1 of the user study were recorded and are available at the following link: https://gofile.io/d/C1kxhI.

B. Survey Questions

Tables III and IV list the survey questions for tasks 1 and 2, respectively, excluding the questions of the CSI and SUS analyses.

C. Thematic analysis for task 1 of the user study

For task one, the most striking positive theme regarding the two mapping conditions was the connection between music and visual elements. Approximately half of the overall answers mentioned this connection ("Consistency between audio and visuals", "I really enjoy the way the the visuals represent the sounds during certain sequences", "I enjoyed the virtual symbiosis between the effects and music [...]"). The particle systems in the visualisation seemed to catch participants' interest the most. Six out of 34 answers stated their positive effect on the viewing experience ("The timings of the particles with the beat were perfect", "The floating particles in the background, i love how they react to the music").

The main negative themes of the two mapping conditions was the lack of connections between audio and visualisations. More specifically, it appears that participants expected every visual object in the scene to be mapped to an audio feature ("Some elements weren't obviously mapping to a single feature of the audio", "I don't each 3d object was

TABLE IV
QUESTIONS FOR TASK 2 OF THE SURVEY

Subtasks
Subtask 1: Connect the overall loudness of the signal to the size of the particles in the centre of the scene. Please comment on the audiovisual mapping process conducted with the system and the music visualisation you obtained. Subtask 2: In this task, you will be asked to control the camera view used in the Interstell.AR visualiser based on audio onset features. Please comment on the audiovisual mapping process conducted with the system and the music visualisation you obtained. Subtask 3: Use the system to create your own mapping by experimenting with different audio features and visual elements. Please comment on the audiovisual mapping process conducted with the system and the music visualisation you obtained. Briefly describe what audio features you chose and how they manifested in the visual domain.
General Questions
Please describe your experience with the audio features available in the system. What additional audio features would you like to see included in the application? Do you have any other suggestions on how to further improve the system? Did you encounter any bugs or issues while working with the system?

mapped to a single musical object [sic]", "[...] I didn't understand how the foggy texture related with thee [sic] music").

Thematic analysis of the control condition led to interesting insights about the system in its default state, without any mappings applied. The most prevalent positive theme was the aesthetics of the visual scene itself ("Looks very clean", "The mid of the picture was nice to look at", "Just how HD it looks", "The imagery is calming"). These answers suggest that participants perceived the visual elements differently from the scenarios where they reacted to music, focusing more on the general appearances of objects.

The prevailing negative theme was the disjuncture between sound and visual objects. 21 participants commented on the stasis of the scene ("Nothing was happening in it, just repetitive", "Didn't feel like it reacted to the music at all", "It seemed way too static. No movement exept for the middle of the picture").