Enhancing Participation Experience in VR Live Concerts by Improving Motions of Virtual Audience Avatars

Hiromu Yakura*
University of Tsukuba
National Institute of Advanced Industrial Science and Technology (AIST), Japan

Masataka Goto†
National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

While participating in live concerts is a promising application of virtual reality (VR), it falls short of our participation experience in the real world. In particular, to increase the engagement of participants, previous studies emphasized the importance of social experience among audience members, such as the sense of co-presence elicited by sharing physical reactions or body movements synchronized with music. In this respect, a common strategy in existing platforms is to present avatars of remote human participants in a VR venue and make every avatar imitate movements of the corresponding participant. However, this strategy implicitly assumes that a not small number of users connect simultaneously to watch the same content and thus is not applicable when only a few users gather or a user is watching alone. Therefore, with the aim of providing better experience to a user who participates in live concerts as one of the audience, we examine computational approaches to enhancing the sense of co-presence through virtual audience avatars. We propose four methods of presenting avatar movements: copying the user's own movements, copying other users' movements, repeating beat-synchronous movements, and synthesizing machine-learning-based movements. We compare their effectiveness in a user experiment and discuss application scenarios and design implications that open up new ways of active media consumption in VR environments.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Human-centered computing—Human computer interaction (HCI)—Interactive systems and tools; Applied computing—Arts and humanities—Performing arts;

1 INTRODUCTION

Participation in live performances is a promising application of virtual reality (VR). For example, Facebook launched a VR-based live broadcasting platform, *Oculus Venues*, in 2018 to let remote users participate in live concerts of famous artists, such as Post Malone and Billie Eilish. VR also enables a new kind of live entertainment where an artist embodying a 3D cartoon style avatar performs in a VR environment by using motion capture technology [51]. In the case of an event held in April 2018, more than 2,500 users gathered into a venue constructed in a VR environment to see a performance of an avatar-based vlogger (so-called "VTuber") [47]. The spread of inexpensive VR devices augments our experience of participating in live concerts [8] and consequently promotes the idea of Marshall McLuhan's *global village* [40], with the audience members anywhere in the world interacting with each other in VR venues regardless of their physical distance.

On the other hand, there are challenges to be overcome in providing immersive experience in remote participation in musical performances. Tarumi et al. [57] conducted a semi-structured interview and found out that remote participation dulls a sense of unity aroused by sharing reactions with other audience members. This result corresponds with the result of interviews with expert performers by Webb et al. [64], which suggested that, in computer-supported performances, a sense of co-presence plays an important role in engaging the remote audience and promoting liveness. It is also supported by the observation in the real world by Solberg and Jensenius [54], who stated that participation in musical performances is an embodied and social experience. They suggested that audience members create social bonds and enhance their entertainment through bodily responses showing interpersonal synchronization with music.

Consequently, Tarumi et al. [57] urged the development of technologies for sharing not only the performances of artists but also the reactions of remote audience members in order to provide better participation experience. In fact, the existing platforms, including Oculus Venues, implement a feature capturing users' body movements via sensors in head-mounted displays (HMDs) and their controllers and reflecting the movements in their avatars presented in the VR environment. However, this feature implicitly assumes that a not small number of users connect simultaneously through their HMDs to watch the same content. In other words, when only a few users gather or a user is watching alone, they cannot enjoy the social aspect of live participation.

This issue conversely suggests that VR-based live participation platforms would be able to offer users in such situations better experience if we could display the movements of other audience avatars computationally. Moreover, this approach can be extended to various applications other than live participation in terms of realizing immersive experience in VR environments. For example, VR video player applications that provide a cinema-like environment, such as SKYBOX Video Player[1] and CineVR[2], can deliver a new watching experience by presenting audience avatars. In particular, considering that many movie theaters have introduced *sing-alongs*, which allow the audience to shake glow sticks or dance along with the movie [1, 2, 60], offering sing-along-like experience in VR environments even while watching alone would have a huge potential. Nevertheless, to the best of our knowledge, the potential benefit of audience avatars has not received much attention in the literature.

Therefore, in this paper, we present our comprehensive study about the presentation of audience avatars during live participation in VR environments though the comparison of four methods we propose. The first two are simple methods: copying the user's own movements (hereafter called "self movements") and copying other users' movements, where the latter one requires the existence of users who have watched the same content beforehand. Given the synchronicity of human movements with music [54], the others exploit the information of the played song, the third repeating synchronous movements based on the beat information and the fourth synthesizing movements from the song's acoustic features by using machine learning. We then compare the effect of the proposed methods on the user's evaluation of participation experience through a

*e-mail: hiromu.yakura@aist.go.jp
†e-mail: m.goto@aist.go.jp

[1] https://skybox.xyz/
[2] https://cinevr.io/

978-1-7281-8508-8/20/$31.00 ©2020 IEEE
DOI 10.1109/ISMAR50242.2020.00083

user study using a custom-developed VR venue. Based on its results, we discuss design implications and application scenarios regarding participation in live performances using VR technologies.

1.1 Contribution

The following three points are the main contributions of this study.

- We introduce a new approach to enhancing media-consuming experience in VR environments featuring live participation by improving audience avatars, which had not previously been focused on.
- We propose four methods for presenting the movements of audience avatars leveraging the movement data of the real-world audience, one of which includes machine-learning-based synthesis.
- We present a design guideline that considers various application scenarios and is based on comparison of the proposed methods in a user study using a custom-developed VR venue.

We believe that our results and discussion open up new applications of active media consumption in VR environments that can fill a gap between presentational and participatory performances [61].

2 Background

In this section we situate our work by introducing prior work regarding the applications of VR in live performances and discussing the challenges of those applications. We also review the existing research on synthesizing human movements computationally that can be related to the presentation of the movements of audience avatars.

2.1 Virtual Reality for Live Performances

VR has been widely used as an entertainment medium for a long time [50]. In particular, its ability to induce a sense of presence, the user's sensation of "being there" during the immersive experience in VR environments [53, 63], has drawn much attention because this sense plays a key role in eliciting an emotional response from a user [48]. Thus, considering that computer-supported performances often lack presence [12], many researchers have proposed methods to leverage VR technologies to realize immersive participation in live performances [3, 16, 24, 28, 29].

Most of these methods are designed to capture the motions of artists using sensors and present them to the audience participating through HMDs in real-time, like Geigel [16] and Kaneko et al. [28] did. There are some proposals of exploiting VR environments to offer participation experience that is unfeasible in the real world, such as dynamically manipulating the spacial effect of audio [3] or displaying related posts in social networking services [29]. Furthermore, Horie et al. [24] augmented the participation experience by using an electroencephalogram recorder and a smartwatch and overlaying virtual effects that correspond to users' brainwaves and heartbeats.

Another use of VR technologies for live performances is cinematic VR [37], a panoramic presentation of video contents in VR environments. He et al. [22] compared the subjective responses and physiological signals of people watching a video of a theater performance in the cinematic setup with an HMD with those of people watching the same video projected on a screen. As a result, cinematic VR was shown to enhance not only a sense of presence and engagement but also the user's desire to watch the performance.

Consequently, as mentioned in Section 1, the commercial use of VR technologies for live performance has already taken place, thanks to the advent of inexpensive VR devices. For instance, the market is expected to reach revenue of 4.1 billion dollars in 2025 [5], and practical discussions on the legal treatment of intellectual property in VR live concerts are progressing [35]. In other words, participation in live performances is also considered in industry to be a promising application of VR.

2.2 Challenges of Live Performances in Virtual Reality

During the expansion of commercial expectations, Tarumi et al. [57] inferred from the result of their semi-structured interview with 14 participants that remote participation in live music performances lacks a sense of unity. They described the sense of unity as a derivative of reacting (cheering or shouting) with other audience members and a precursor of the synergic effect among audience members on their excitement. As mentioned in Section 1, this result corresponds with the observations of audience in the real world that spotlight the embodied and social aspects of live participation [14, 54]. It also is consistent with the experiment by He et al. [22]; that is, among all subjective measures compared between cinematic VR and screen projection, cinematic VR diminished only the sense of co-presence with other audience members.

One good example that illustrates the importance of the social aspect of the audience members' existence is given by *Hatsune Miku* [27, 31, 33], the virtual character of a singing voice synthesis software. In "her" live concerts held across the world, an audience of thousands packed a venue and shook their hands with color-changing glow sticks synchronously in front of the holographic screen showing the dance of the character [31]. Leavitt et al. [33] pointed out that these concerts provide the audience with shared energy and the excitement of being involved, which enhances the liveness of their participation. This means that the existence of other audience members itself can be a major motivation to participate in live concerts.

Consequently, several papers suggested providing a function to share reaction information between remote participants in order to enhance the sense of unity or co-presence in live performances [57, 64]. This idea has actually been reflected in the design of the existing VR-based platforms by mirroring the body movements of the audience members in their avatars in real-time. However, as explained in Section 1, this approach implicitly assumes that a not small number of users are connecting simultaneously and thus has limited applicability. In addition, considering that a person moving off the beat is perceived negatively by people moving in synchrony with music [55], such a naive mirroring can be ineffective due to the delay caused by the latency of the network.

In summary, current VR-based live participation platforms have room for improvement regarding the deployment of audience avatars. A comprehensive study regarding the presentation of audience avatars in VR environments is therefore desirable if we are to discuss future directions of the use of VR technologies for media consumption including live participation.

2.3 Synthesizing Human Movements

In connection with our motivation of presenting the movements of audience avatars during live performances, there are many proposals to synthesize temporal human movements corresponding to music, such as choreography synthesis [15, 44, 56]. For example, Ofli et al. [44] proposed a method to synthesize dance figure sequences by selecting a dance figure for each musical measure in accordance with the acoustic feature of the measure. Tang et al. [56] used a long short-term memory (LSTM) network to synthesize dance choreographies from the input music in an end-to-end manner using a large-scale dance-music dataset. While these methods succeeded in synthesizing human movements synchronized with music, we note that their aim is to synthesize expressive movements of a single person. Thus they are not directly applicable to our situation, which is targeted at presenting the synchronized movements of numerous avatars to look like audience members in live performances.

Synthesizing the movements of multiple people is also discussed in the context of crowd simulation [58, 66]. For example, Lee et

al. [34] proposed a method for simulating crowd behaviors by imitating an aerial video of real human crowds and gave an example of the synthesized movements of an audience around a street dance performance. And Gu et al. [20] introduced a context-aware diversity control for crowd simulation replicating a wide range of situations: from pedestrians to the audience of a street performance or soldiers in a military march. These methods are not designed to accompany music, however, and are not associable with our motivation, which focuses on synchronicity with music in order to enhance participation experience in live performances. Yilmaz et al. [67] discussed the possibility of simulating the movements of music concert audience members in accordance with the acoustic features of the music but have not yet implemented it.

We conclude from the above that our motivation is different from that in existing literature on choreography synthesis or crowd simulation. Thus we believe that our proposal of the computational approaches to presenting the movements of audience avatars can make a fundamental contribution to the realization of a new way of active media consumption in VR environments.

3 Apparatus and Material

To examine the effect of our proposals for presenting audience avatars, we first developed a VR venue for Oculus Quest[3] using Unity[4]. In our application, a user is surrounded in a virtual concert area by more than 200 audience avatars, whose movements are controlled computationally by the proposed methods. To enable the user to see both the performance on the stage and other audience members, we used a 3D model of an amphitheater-style venue, and the location of the user was fixed at the audience seat in the fifth row in front of the stage. The movements of the user are captured at three points (the head and both hands) by using Oculus Quest and Oculus Touch and are reflected in the user's avatar through an inverse kinematics algorithm that calculates a realistic positioning of the avatar's skeleton from the positions and rotations of the three points.

To compare the effects between different presentation methods, the musical performance shown to the user should be replicable in combination with each method. Either of two strategies can be used to realize such a stimulus: playing a video that recorded the musical performance of a human artist in the cinematic VR setup, or replicating the performance of an artist by animating a 3D model along with the music. For the following reasons, we used the latter strategy; in particular, presenting the animated 3D model of Hatsune Miku.

First, the live concerts of Hatsune Miku can be one of the promising applications of our methods because, as we discussed in Section 2.2, the existence of other audience members plays an important role in affording the motivation to participate in the concerts. Moreover, Hatsune Miku is known for forming a collaborative creation community on the Internet [31, 33]. For instance, in addition to the fact that songs composed by musicians using the singing voice synthesis software are distributed in online video-sharing services, the motion data of her 3D model dancing to the songs are also uploaded by peer choreographers in the community voluntarily [21, 33]. Thus we can utilize such resources to demonstrate the musical performance to the user in the VR venue.

Furthermore, using these resources, some fans sometimes hold a hand-made live concert of Hatsune Miku by preparing a projector and screen [33]. They also often post the footage of the concert on the Internet, which can be a great help for our situation because it opens up the way for the statistical synthesis of the movements of audience avatars in a manner similar to that of the existing methods presented in Section 2.3. In particular, since the existence of the audience comprises a major motivation to participate in her concerts, as discussed in Section 2.2, not only the performance but also the movements of the audience members are filmed considerably more in such footage than they are in the footage of concerts of human artists. We would like to emphasize that, though the movement data of real audience members are essential for statistical synthesis, it is difficult to collect such data unless we use the footage or hold large-scale live concerts by ourselves.

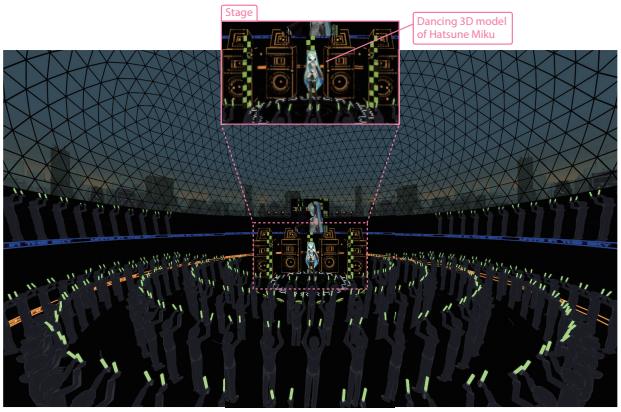

Figure 1: An example interface of our custom-developed VR venue. The 3D model of Hatsune Miku dances on the stage located in the center, and the surrounding audience avatars hold glow sticks in the same way that people at her concerts in the real world do.

At the same time, because of the equipment and resource limitation described above, we focus on the hand movements of audience avatars rather than synthesizing the entire body movements. In detail, the audience is not well lighted up in common live concerts to emphasize the stage lighting, and thus it is difficult to retrieve from the footage the body movements of the audience members other than the hand movements characterized by glow sticks. In addition, as previously mentioned, Oculus Quest captures the user's movement with the three points of the head and both hands, and then additional equipment is required to capture the entire body movements. We believe that the computational presentation of the synchronized hand movements of audience avatars still exhibits a wide range of applications, considering that the hand movements of audience members are often emphasized using glow sticks not only in live concerts in the real world but also in sing-along theaters.

In summary, we constructed a VR venue in which a user sees the musical performance of the dancing 3D model of Hatsune Miku while being surrounded by audience avatars as shown in Figure 1. Then, to present the dance, we collected five songs and corresponding motion data from the creation community on NicoNico[5], each of which lasts about 3–4 minutes. We also collected the footage of three fan-made concerts (one was held in the US, and the others were held in Japan), which contains 38 songs within 2 hours 54 minutes.

4 Proposed Methods

In this section we describe our proposal of four methods to present the hand movements of audience avatars in synchronization with music, which is implemented in our VR venue. The first two methods copy the user's own movements and other users' movements, respectively, whereas the others exploit the information of the played song.

[3] https://www.oculus.com/quest/
[4] https://unity.com/
[5] https://www.nicovideo.jp/

4.1 Copy of Self Movements

We first consider a simple method of copying the self movements of the user to the audience avatars, which inevitably forms the synchronized movements as long as the user moves in synchrony with music. The advantage of this method is that it does not require additional resources to display the movements of the audience avatars.

4.2 Copy of Other Users' Movements

We next consider copying the movements of other users who watched the same content. This method is inspired from a *Danmaku* interaction in online video-sharing services [18, 65]. In such services, text comments written by various users who have watched the same video clip are recorded in association with specific playback times that indicate when the users wrote them. Then the comments are overlaid on the video in synchronization with the playback, which gives an impression that people are watching together while typing comments. The Danmaku interaction is known to induce a sense of co-presence [10, 36] as well as an impression of liveness [26].

In our method, the users' movements are recorded along with the timeline of the musical performance and displayed using the audience avatars when the performance is replayed. The advantage of this method is that it is expected to present more realistic movements than copying the self movements does, considering that it results in an experience almost identical to watching the performance together with other users simultaneously. On the other hand, this method assumes that other users have watched the same content in advance and thus cannot be used in some situations, such as live streaming.

4.3 Repetition of Beat-Synchronous Movements

We also consider computational approaches to synthesizing the movements of the audience avatars rather than directly copying human movements. In our observation on the collected footage, we found that the hand movements of the audience members consist of a small number of characteristic sequences that are repeated over the musical measures so as to be synchronized with the beat of the music. This means that we can replicate the movements of the audience members by repeating such sequences according to the measure information extracted from the music by using beat detection algorithms [42].

To do that, we first applied a clustering algorithm to extract the representative sequences of the hand movements from the footage. To prepare the input features for the clustering algorithm, we captured the movements of the audience members through the glow sticks, as mentioned in Section 3, for each musical measure. In detail, we separated the footage by the musical measures based on the results of the beat detection provided by Songle [19]. Then we calculated the histogram of optical flow [32], which is one of the common features used in human action recognition, from the movements of the glow sticks in each musical measure. We consequently applied a time-series k-means algorithm [46] to the calculated features and determined the optimal number of clusters based on the elbow method [68]. As a result, $k = 4$ was suggested to be a reasonable choice, given that it was the first inflection point in Figure 2.

Figure 3 shows the extracted representative sequences, each of which is closest to the center of the corresponding cluster. They depict the transitions in the histogram of optical flow within a musical measure, and their dominant directions correspond to major hand movements, as illustrated in Figure 4. That is, the first cluster represents the movement of raising hands throughout the measure, the second cluster represents the movement of putting hands up twice in synchronization with the beat, the third cluster represents the faster movement of putting hands up at all four of the beats in the measure, and the fourth cluster represents the movement of shaking hands horizontally twice in the measure. Considering that these four clusters correspond to results reported in previous empirical taxonomy literature presenting four classes of the hand movements

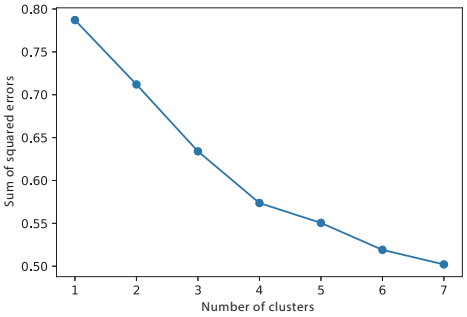

Figure 2: Change of the sum of squared errors for the different numbers of clusters. This result suggests that $k = 4$ is a reasonable choice for the number of clusters.

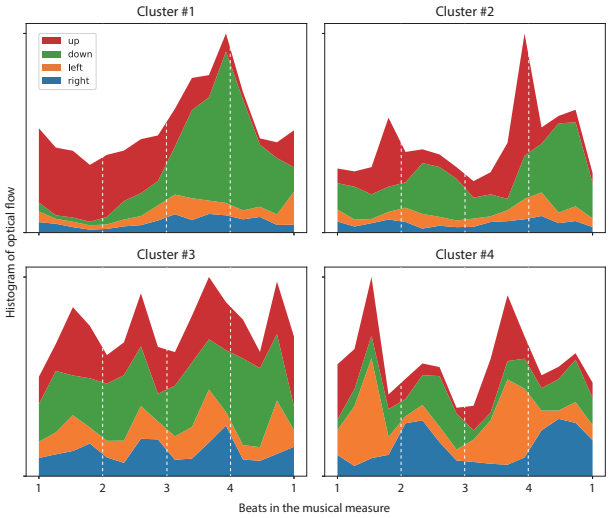

Figure 3: Histograms of optical flow within a musical measure: the depicted four clusters correspond to the extracted representative sequences of the hand movements. The white dotted lines represent the timing of beats, and the corresponding movements are illustrated in Figure 4. Note that, considering that the hand position reaches its peak at the beat, the optical flow indicating the relative movements of the hand would peak just before the beat.

of audience members in live concerts [41], we can conclude that these clusters have modeled the movements well.

Based on the above results, we propose a method presenting the movements of the audience avatars by repeating the representative sequences in time to the musical measures of playing songs. In detail, we first prepared motion sequences corresponding to the four clusters by recording the movement of one of the authors who imitated the extracted sequences while using Oculus Quest and Oculus Touch. When a song starts, one of the four motion sequences is randomly assigned to each audience avatar so that the derived distribution becomes similar to that of the footage from the concerts in the real world, which is shown in Table 1. Then each avatar iterates the assigned motion sequence for every musical measure, whose boundaries are recognized by the beat detection algorithm.

4.4 Synthesis Using Machine Learning

On further inspection of the footage, we found that the audience members seem to unconsciously choose a conforming movement from the typical movement patterns in accordance with the atmo-

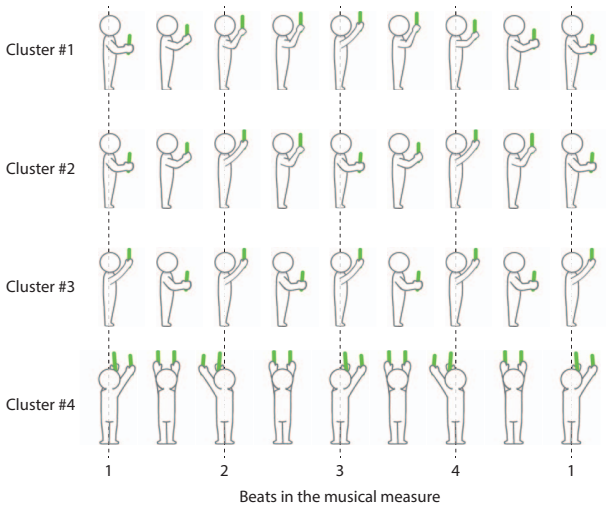

Figure 4: Illustrations of major hand movements corresponding to the four clusters of Figure 3. The dotted lines represent the timing of beats in the musical measure.

Table 1: Distribution of the obtained clusters in the collected footage of the concerts. The movement of putting hands up at every beat was the most dominant.

Cluster	#1	#2	#3	#4
Occurrence	276	682	1,132	207
Proportion	12.0%	29.7%	49.3%	9.0%

sphere of the music. This suggested that more realistic movements of the audience avatars could be synthesized if we modeled their selection process computationally. For this purpose, we preliminarily examined whether the relationship between the song and the movement sequences can be predicted using machine learning algorithms.

For the input features representing the atmosphere of the music, we prepared four feature sets and compared their effectiveness. The first three feature sets are based on the previous work for music genre classification by Tzanetakis et al. [62]: timbral texture features, rhythmic content features, and pitch content features. For the timbral texture features, we calculated the mean and variance of spectral centroid, spectral rolloff, spectral flux, zero-crossing rate, and Mel-frequency cepstral coefficients (MFCCs) for each musical measure using librosa [39]. For the rhythmic content features, we first extracted the transition of onset strength within each musical measure. This transition represents how quickly the power spectrum of the audio signal is increasing, and its peaks could be regarded as onset events. After normalizing their heights by the height of the highest peak, we calculated the normalized heights of the second to ninth highest peaks as well as the local tempo in beats-per-minute (bpm). For the pitch content features, we first extracted the chromagram – which represents the distribution of energy over the 12 distinct pitch classes ($C, D\flat, D, E\flat, E, F, G\flat, G, A\flat, A, B\flat, B$) – over each musical measure. Then, we calculated the relative amplitudes of the first to fifth peak within this 12-dimensional vector, which will be higher for a measure that does not have many harmonic changes, and used them in addition to the original vector.

We also added global structure features that refer to the structural information of the entire song, for example, the existence of repeating sections or the position of prominent thematic sections (often referred to as chorus sections). This is because the movement selection by audience members can be affected by them, such as an energetic movement being likely to be chosen in chorus sections. In detail, we added two features (four variables): flags representing whether the current measure is in repeating sections and chorus sections, and the numbers of preceding repeating sections and chorus sections. Note that the calculation of these features is based on the result of RefraiD [17], which is also provided by Songle [19].

Table 2: Accuracy of the prediction of the audience members' movements (four classes) for each combination of the input feature sets. The combination of all four feature sets showed the highest accuracy.

Feature sets				Accuracy
Timbral texture	Rhythmic content	Pitch content	Global structure	
✓				69.35%
	✓			62.87%
		✓		62.17%
			✓	64.21%
✓	✓			70.70%
✓		✓		70.70%
✓			✓	74.18%
	✓	✓		66.57%
	✓		✓	70.27%
		✓	✓	69.31%
✓	✓	✓		72.18%
✓	✓		✓	74.27%
✓		✓	✓	73.97%
	✓	✓	✓	71.01%
✓	✓	✓	✓	**76.67%**
Expected value for the random selection				35.36%

Using the paired data of the above feature sets and the audience movements in the collected footage, we compared the prediction accuracy of three-fold cross validation while changing the combination of the feature sets to use. For the machine learning algorithm we used LightGBM [30] because it is known for its accuracy and computational efficiency. Here the output of the prediction model is a 4-class label indicating which of the four movements of Figure 3 was observed in the given musical measure.

The results are presented in Table 2, which shows that our prediction scheme achieved much better accuracy than the random selection we used in Section 4.3, especially in the case of combining all four of the feature sets. By comparing the combinations, we confirmed that all of the feature sets contributed to the improvement of the accuracy. This is also supported by the fact that the top three features regarded as important by LightGBM were (in order) the first MFCC, the tempo, and the flag indicating whether or not the current measure is in a chorus section. We recognize that the accuracy could be improved by devising better features and parameters, but as mentioned in Section 1, here we are presenting a comprehensive study regarding the presentation of audience avatars rather than the prediction of movement sequences.

Thus, we propose another method to synthesize the movements of the audience avatars by using the above prediction model in order to present a natural movement that matches the music. This method calculates the likelihood of the four movements of Figure 3 for each musical measure by using the prediction model and assigns one of the movements to each audience avatar with a probability of the corresponding likelihood. Then, during the musical measure each avatar replays the same motion sequence of the assigned movement as that described in Section 4.3.

5 Experiment

To evaluate the effects of the proposed methods on the participation experience of live performances in VR environments, we conducted a user experiment using our custom-developed VR venue. In this

section we describe its setup and results in detail.

5.1 Measure

As we explained in Section 1, our motivation is to enhance a sense of unity or co-presence by presenting the movements of audience avatars computationally so as to provide better participation experience. In other words, we anticipate that our methods can foster the sense of unity or co-presence among users.

Unfortunately, however, there is no established scale for measuring what Tarumi et al. [57] called a sense of unity. In addition, conventional scales used to measure the sense of co-presence in VR environments [6, 43, 52] assume human-to-human interaction considering the assessment of the quality of telepresence systems, such as "I think the other individual often felt alone" or "I was interested in talking to my interaction partner." Though some research has used these scales to evaluate human-to-agent interaction of dyadic conversations [4, 45], they would not be directly applicable to our situation, which just presents the movements of other avatars and does not provide explicit interaction with them.

We therefore used a modified version of these scales proposed by Hwang et al. [25], who investigated the effect of online communication while watching a live sports event on TV on the viewers' sense of co-presence. Since their scale of the sense of co-presence was designed under the assumption that viewers were participating in a specific event, its items such as "I feel like I was rooting for my country's team along with many others" conform to our situation. In addition, though the study conducted by Hwang et al. was not intended to evaluate interaction techniques, recent studies [7, 9] used their scale to measure the effect of the *Danmaku* interaction, and thus it seems to be a suitable scale. We note that the scale consists of four items, and each of them was slightly modified to fit with our scenario as follows:

- I feel like I was participating in the live concert with others.
- I feel like many people were participating at the same time.
- I feel like I was physically communicating with others.
- I feel like I was rooting for the artist along with many others.

We also measured a sense of presence to evaluate the quality of the whole experience of live participation in our VR venue, measuring it in a manner similar to that in which it was measured in previous studies [22, 48]. In the work reported here we used the Slater-Usoh-Steed (SUS) scale [63] consisting of six items, which is one of the most popular presence scales applicable in VR environments [49].

5.2 Participants

Our experiment involved 20 participants, ranging in age from 20 to 46, five of whom were female. They were recruited mainly from a local university student community using word-of-mouth communication with the condition that they had an experience of using glow sticks, which potentially reflected the homogeneity of the Japanese fandom culture [59] in the demography of the participants. They participated voluntarily and spent about 15 minutes on our experimental procedure. They were allowed to stop the procedure at any time, but all of them completed it.

5.3 Procedure

Each of the participants experienced two sessions, each with a different song and a different presentation of the audience avatars. Here, in addition to presenting audience avatars with movements generated by each of the four methods proposed in Section 4, we prepared a baseline situation presenting audience avatars without movement. We note that, for the case of copying other users' movements, we used the movements of five participants who had watched the same song with other presentation methods previously.

The participants did not go through all combinations of the songs and presentation methods because that would have taken too long and make the participant recruitment more demanding. Instead, they were randomly and uniformly assigned using a balanced Latin square of the songs and the presentation methods so as to avoid the influence of the learning effect and the users' preference for the songs. More specifically, there are 20 possible ordered combinations of choosing two from the five conditions, and thus, each of them was assigned to exactly one participant so as to maintain the balanced block design. Furthermore, in order to corroborate the evaluation using the measures mentioned in Section 5.1, we analyzed their pairwise comparison between the two presentation methods each participant experienced by using the Plackett-Luce model [38].

Figure 5: Example setup of the experiment. The participants donned Oculus Quest and watched the performances in our VR venue.

Before the participants started the sessions, they were given a brief explanation about the experiment and donned Oculus Quest in a meeting room, as shown in Figure 5. After they finished each session (i.e., after watching the performance of the single song in our VR venue), they removed the Oculus Quest and completed a 10-item questionnaire for measuring senses of co-presence and presence on a 7-point Likert scale. In addition, after their second session, we asked them which of the two sessions they felt was better as well as for their comments about the participation experiences.

5.4 Results

Figure 6 shows the scores indicating the participants' sense of co-presence for each presentation method. Using Durbin's test [11, 13], we found in the scores a significant effect of the presentation methods ($D(4) = 12.84, p = 0.01$). From its post-hoc comparisons, we found that the scores for showing no movements and for copying self movements differed significantly from those for any of the other methods ($p < 0.05$).

Figure 7 shows the scores indicating the participants' sense of presence for each presentation method. According to Durbin's test, there was also a significant effect of the presentation methods in the scores ($D(4) = 10.22, p = 0.04$). From the post-hoc comparisons, we found that both copying other users' movements and synthesizing machine-learning-based movements showed significant differences against both showing no movements and copying self movements ($p < 0.05$). In addition, we also found that the scores of the presence were significantly correlated with those of the co-presence (Spearman's $\rho = 0.52, p = 0.001$).

As mentioned in Section 5.3, we also analyzed their response for the pairwise comparison between two methods assigned to each participant by using the Plackett-Luce model. The obtained score of each method is shown in Table 3, which indicates that the probability of being judged as better between all combinations of any two methods would correspond to the ratio of their scores. These scores showed similar trends with the scores of the senses of co-presence and presence, with particularly high scores for copying other users'

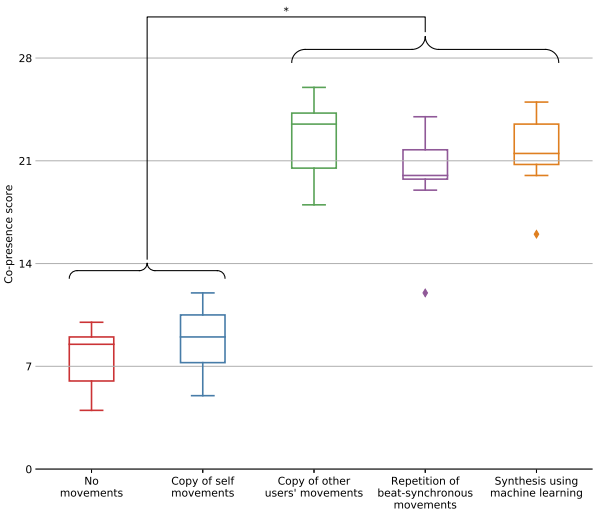

Figure 6: Scores indicating the participants' sense of co-presence for the proposed methods. Both showing no movements and copying self movements showed significant differences from all three of the other methods in our post-hoc analysis ($p < 0.05$).

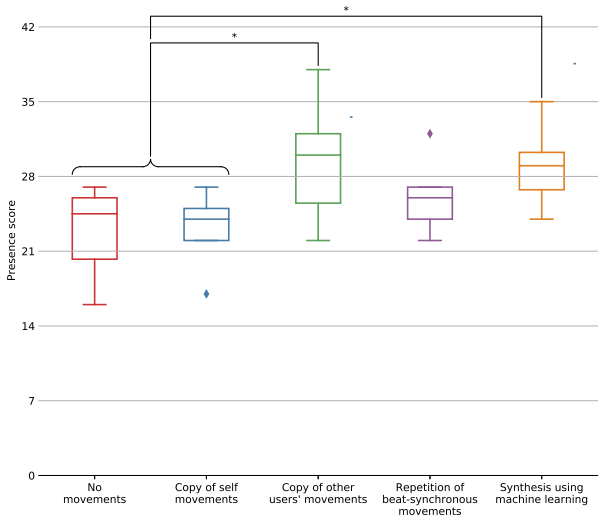

Figure 7: Scores indicating the participants' sense of presence for the proposed methods. Both copying other users' movements and synthesizing machine-learning-based movements showed significant differences against both showing no movements and copying self movements in our post-hoc analysis ($p < 0.05$).

Table 3: Scores calculated from the participants' pairwise comparisons of their participation experiences using the Plackett-Luce model.

Presentation method	Score
No movements	0.085
Copy of self movements	0.020
Copy of other users' movements	0.387
Repetition of beat-synchronous movements	0.140
Synthesis using machine learning	0.368

movements and synthesizing machine-learning-based movements. This result also suggested that, in contradiction with our expectation, copying self movement did not attract the participants, which demands further analysis.

5.5 Analysis

As presented in Section 5.4, the proposed methods of presenting the movements of audience avatars contributed to the enhancement of the users' sense of co-presence, except for the case of copying the movements of the users themselves. This interesting result can be understood from the comments of the participants who experienced the situation, such as "I immediately noticed that my movements were copied and felt a little creepy because the avatars seemed like clone humans." To put it the other way around, the reason why copying other users' movements was positively perceived even though the audience avatars showed similar movements can be attributed to the human-like variation in their movements, such as the angle of the hands or the peak positions. We also observed that, once the participants noticed that their movements seemed to be copied, they often tried to figure out how the audience avatars were manipulated rather than pay attention to the performance, like moving their hands in various directions as if they had seen a mirror for the first time. These results suggest that copying self movements would appear strange to users rather than enhance their sense of co-presence.

At the same time, the other three methods provided better participation experience; in particular, copying other users' movements and synthesizing machine-learning-based movements significantly enhanced the senses of co-presence and presence. The effectiveness of copying other users' movements is consistent with our statement in Section 4.2 that it results in an experience almost identical to watching the performance together with other users simultaneously. On the other hand, the effectiveness of the machine-learning-based synthesis can be attributed to the naturalness of the movements against the music brought by the machine learning algorithm considering that the case of repeating beat-synchronous movements used the same motion sequences. This is also supported by the comments of the participants who experienced both situations, such as "In the second time *[note: a session with the machine-learning-based synthesis]*, the audience members were more realistic than they were in the first time *[note: a session with the repetition of beat-synchronous movements]* because they changed their movements in accordance with the music" and "I felt that the audience members in the first time *[note: a session with the machine-learning-based synthesis]* had personalities they didn't have in the second time *[note: a session with the repetition of beat-synchronous movements]*."

In addition, the scores of the senses of co-presence and presence, and also the score from the Plackett-Luce model showed similar trends among the presentation methods. This result is consistent with the previous literature that pointed out the importance of the sense of unity [57] or co-presence [64] in providing better participation experience. Our experimental results thus complement their arguments based on qualitative interviews.

While the above results suggest the advantage of presenting audience avatar movements, there were comments about a side effect of the presentation such as "I felt like being encouraged to move in the same manner as the surrounding avatars." However, none of the respondents regarded this feeling as unfavorable, and thus we consider that it would not negatively affect their responses to the questionnaire.

6 Discussion

In this section we discuss design implications regarding the presentation of audience avatars in VR environments during live performances. We also discuss the limitations and future directions of our proposal of the computational approaches.

Table 4: Design Implications for deploying computational approaches to presenting the movements of audience avatars in VR applications. They are derived from the results of our user experiment.

	Live streaming via the Internet	**Playback of private contents**	**Playback of shared contents**
Copy of self movements	Not recommended	Not recommended	Not recommended
Copy of other users' movements	Not recommended – unless low-latency access to the Internet is guaranteed	Not applicable	Highly recommended if the movements are available – but it is also recommended to combine with the machine-learning-based synthesis to avoid the cold-start problem
Repetition of beat-synchronous movements	Recommended	Applicable	Applicable
Synthesis using machine learning	Potentially applicable – if the prediction model is trained without the global structure information	Recommended	Recommended

6.1 Design Implications

Our results revealed that presenting the movements of audience avatars is a reasonable method to provide better participation experience in live performances unless they copy the self movements of a user. The remaining three methods have different resource requirements for presenting the movements. Considering this point, we discuss implications for designing VR applications based on the above results considering various scenarios.

As mentioned in Section 1, our approaches can be used not only for participation in live performances but also in cinematic VR to watch pre-recorded videos with audience avatars, like a sing-along theater. Here these applications can be divided into three scenarios according to the available resources: live streaming via the Internet, playback of private contents, and playback of shared contents.

In live streaming, copying other users' movement is not recommended despite its effectiveness because the delay in the movements caused by the network latency potentially degrades the user's participation experience [55], unless low-latency access to the Internet, such as 5G networks, becomes widespread. In addition, it is also difficult to retrieve the information about the song to be played next, which is required by the machine-learning-based synthesis to calculate the global structure features. Therefore, although the machine-learning-based synthesis showed the better scores, we conclude that the repetition of beat-synchronous movements is the best choice for live streaming via the Internet, given that the beat information can be extracted from the audio signal in real-time [42].

However, as shown in Table 2, the omission of the global structure features did not drastically reduce the prediction accuracy of the movement selection. By omitting these features, the synthesis using machine learning can be potentially applied to live streaming with a short buffering for calculating the musical-measure-wise features.

In playback of private contents, in contrast, we can retrieve the acoustic information of the entire song beforehand whereas other users' movements for such private contents are not available. Thus, according to our results, the machine-learning-based synthesis is a reasonable choice.

In playback of shared contents, the machine-learning-based synthesis is recommended for the same reason. In addition, we can imagine that, in the near future, users watching the shared contents would upload their movement data to be shared with other users, in an analogy to the playback-synchronized user comments in some online video-sharing services as described in Section 4.2. In such an age, by downloading the movement data beforehand and replaying in synchronization with the playback, we can duplicate the situation of copying other users' movement, which gives an impression of watching together with other users simultaneously. From these points, it is worth considering the employment of copying other users' movement as well as a way to record and collect the movement data from many users.

At the same time, we note that it raises a concern for a *cold-start* problem, e.g., the second user watching a content sees a strange situation that all of the audience avatars copy the same movements of the first user. Thus it is also recommended to combine copying other users' movements with the machine-learning-based synthesis until a certain number of users have watched the content.

We summarized the above discussion in Table 4. We hope that these design implications will help researchers and practitioners provide better participation experience for users.

6.2 Limitations and Future Work

Though our results have paved the way for new approaches to improving VR-based live participation platforms, there are still some limitations. For example, additional investigations involving a greater number and diversity of participants with better gender-balancing are desirable to ensure the generalizability of the results. The measures we used to examine the effects of the proposed methods (i.e., the senses of co-presence and presence) shed light on an important but limited part of participation experience. Since other aspects, such as participants' liking of songs or the amount of their movements, would affect the experience, we plan to explore their link to the proposed methods.

The proposed methods also have room for improvement. For instance, as mentioned in Section 4.4, the prediction model used in the machine-learning-based synthesis can be improved by devising features, such as adding standard deviation, kurtosis, or skew, which might offer ways to surpass the case of copying other users' movements.

Other than these points, as mentioned in Section 3, we have so far focused on presenting the hand movements of audience avatars rather than the entire body movements because of the limited data available. If we can obtain the entire body movements of individual audience members during live concerts, the end-to-end synthesis from music can be realized using previous methods for choreography synthesis, which we mentioned in Section 2.3. We suspect that the presentation of the entire body movements would contribute to the improvement in the user's sense of co-presence, but that confirmation is left as a future work because of the difficulty and the cost of collecting such extensive data on body movements.

In addition, there is a limitation due to the computational performance of the device we used, Oculus Quest. In detail, it is difficult to increase the number of the audience avatars from the current setting without causing dropped frames. Given that live performances in the real world often involve an audience of thousands, we are also interested in how the number of audience avatars affects the user's

impression, in particular, in the case of presenting thousands of audience avatars. However, in order to do that we need to re-implement the VR venue in a more sophisticated way using a more powerful device. We therefore also leave this point for future work.

Toward the deployment in real applications, there is wide room for research expansion. For instance, in the same manner as Zajonc observed for various subjects [69, 70], long-term exposure to the audience avatars would affect the user's perception of them. In addition, although we did not consider the presence of avatars controlled by other human audience members in our experiments, it is possible to use the proposed methods when other users also gather in the same VR environment, as we mentioned in Section 1. In such a situation, the next goal is to integrate human-to-human interactions with the proposed methods so that a user can enjoy the social aspect of live participation with the help of both human and virtual audience avatars.

Lastly, exploration of different application scenarios would be a fruitful endeavor. As discussed in Section 6.1, our methods can be applied to other VR applications, such as cinematic VR for replicating sing-alongs. As collective VR experiences [23] showed that VR-based storytelling applications can also be enriched by the existence of audience members, it is worth extending the proposed methods to such narrative contents. Furthermore, our approach of synthesizing the movements of audience avatars can offer the sensation of becoming a famous artist to a user through emulating the experience of being on a stage surrounded by many audience members. We believe that the development of such a new way of active media consumption can be based on our results.

7 Conclusion

In this paper we examined computational approaches to presenting the movements of audience avatars during the participation in live performances via VR environments for the purpose of providing better participation experience by improving the user's sense of co-presence. We proposed four different methods: copying self movements, copying other users' movements, repeating beat-synchronous movements, and synthesizing machine-learning-based movements. We then compared their effectiveness in a user experiment using a custom-developed VR venue. Our experiment revealed that, while copying movements of other users who watched the same content before is quite effective for enhancing participation experience, the machine-learning-based synthesis demonstrated comparable performance and would be a reasonable choice considering its availability in many situations. We hope that our results and discussion not only help VR-based live participation platforms provide immersive experience but also open up new VR applications for active media consumption.

Acknowledgments

This work was supported in part by JST ACCEL Grant Number JPMJAC1602, Japan.

References

[1] B. Alexander. Sing-along 'Frozen' coming to theaters. USA TODAY, https://www.usatoday.com/story/life/movies/2014/01/22/frozen-sing-along/4783327/, January 2014. Accessed: October 15, 2019.

[2] R. C. Allen. Reimagining the history of the experience of cinema in a post-movie-going age. *Media International Australia*, 139(1):80–87, May 2011. doi: 10.1177/1329878X1113900111

[3] R. Atarashi, T. Sone, Y. Komohara, M. Tsukada, T. Kasuya, H. Okumura, M. Ikeda, and H. Esaki. The software defined media ontology for music events. In *Proceedings of the 1st International Workshop on Semantic Applications for Audio and Music*, pp. 15–23. ACM, New York, NY, 2018. doi: 10.1145/3243907.3243915

[4] J. Bailenson, R. Guadagno, E. Aharoni, A. Dimov, A. Beall, and J. Blascovich. Comparing behavioral and self-report measures of embodied agents' social presence in immersive virtual environments. In *Proceedings of the 7th Annual International Workshop on Presence*, pp. 1–8. ISPR, Philadelphia, PA, 2004.

[5] H. Bellini, W. Chen, M. Sugiyama, M. Shin, S. Alam, and D. Takayama. Virtual & augmented reality: Understanding the race for the next computing platform. Goldman Sachs Global Investment Research, https://www.goldmansachs.com/insights/pages/technology-driving-innovation-folder/virtual-and-augmented-reality/report.pdf, January 2016. Accessed: October 16, 2019.

[6] F. Biocca, C. Harms, and J. Gregg. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In *Proceedings of the 4th Annual International Workshop on Presence*, pp. 1–9. ISPR, Philadelphia, PA, 2001.

[7] J. A. Cameron. *The on-screen water cooler: Effects of televised user-generated comments on cognitive processing, social presence, and viewing experience*. PhD thesis, University of Tennessee, Knoxville, TN, August 2016.

[8] J.-P. Charron. Music audiences 3.0: Concert-goers' psychological motivations at the dawn of virtual reality. *Frontiers in Psychology*, 8:800, May 2017. doi: 10.3389/fpsyg.2017.00800

[9] L. Chen. How danmaku influences emotional responses: Exploring the effects of co-viewing and copresence. Master's thesis, Nanyang Technological University, Nanyang Avenue, Singapore, February 2018.

[10] Y. Chen, Q. Gao, and P. P. Rau. Watching a movie alone yet together: Understanding reasons for watching Danmaku videos. *International Journal on Human-Computer Interaction*, 33(9):731–743, February 2017. doi: 10.1080/10447318.2017.1282187

[11] W. J. Conover. *Practical Nonparametric Statistics*. John Wiley and Sons, New York, NY, 3 ed., 1999.

[12] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Mühlberger. The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in Psychology*, 6:26, January 2015. doi: 10.3389/fpsyg.2015.00026

[13] J. Durbin. Incomplete blocks in ranking experiments. *British Journal of Statistical Psychology*, 4(2):85–90, 1951. doi: 10.1111/j.2044-8317.1951.tb00310.x

[14] M. Ellamil, J. Berson, J. Wong, L. Buckley, and D. S. Margulies. One in the dance: Musical correlates of group synchrony in a real-world club environment. *PLoS ONE*, 11(10):1–15, October 2016. doi: 10.1371/journal.pone.0164783

[15] S. Fukayama and M. Goto. Music content driven automated choreography with beat-wise motion connectivity constraints. In *Proceedings of the 12th International Conference in Sound and Music Computing*, pp. 177–183. Zenodo, Genève, Switzerland, 2015. doi: 10.5281/zenodo.851119

[16] J. Geigel. Creating a theatrical experience on a virtual stage. In *Proceedings of the 14th International Conference on Advances in Computer Entertainment Technology*, pp. 713–725. Springer International Publishing, Cham, Switzerland, 2017. doi: 10.1007/978-3-319-76270-8_49

[17] M. Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1783–1794, September 2006. doi: 10.1109/TSA.2005.863204

[18] M. Goto. Music listening in the future: Augmented music-understanding interfaces and crowd music listening. In *Proceedings of the 42nd AES International Conference Semantic Audio*, pp. 21–30. AES, New York, NY, 2011.

[19] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano. Songle: A web service for active music listening improved by user contributions. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pp. 311–316. ISMIR, Montreal, Canada, 2011.

[20] Q. Gu and Z. Deng. Context-aware motion diversification for crowd simulation. *IEEE Computer Graphics and Applications*, 31(5):54–65, September 2011. doi: 10.1109/MCG.2010.38

[21] M. Hamasaki and M. Goto. Songrium: a music browsing assistance service based on visualization of massive open collaboration within music content creation community. In *Proceedings of the 9th International Symposium on Open Collaboration*, pp. 4:1–4:10. ACM, New

York, NY, 2013. doi: 10.1145/2491055.2491059

[22] L. He, H. Li, T. Xue, D. Sun, S. Zhu, and G. Ding. Am I in the theater?: usability study of live performance based virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, pp. 28:1–28:11. ACM, New York, NY, 2018. doi: 10.1145/3281505.3281508

[23] S. Herscher, C. DeFanti, N. G. Vitovitch, C. Brenner, H. Xia, K. Layng, and K. Perlin. CAVRN: an exploration and evaluation of a collective audience virtual reality nexus experience. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pp. 1137–1150. ACM, New York, NY, 2019. doi: 10.1145/3332165.3347929

[24] R. Horie, M. Wada, and E. Watanabe. Participation in a virtual reality concert via brainwave and heartbeat. In *Proceedings of the 8th International Conference on Applied Human Factors and Ergonomics*, pp. 276–284. Springer International Publishing, Cham, Switzerland, 2018. doi: 10.1007/978-3-319-60495-4_30

[25] Y. Hwang and J. S. Lim. The impact of engagement motives for social TV on social presence and sports channel commitment. *Telematics and Informatics*, 32(4):755–765, November 2015. doi: 10.1016/j.tele.2015.03.006

[26] D. Johnson. Polyphonic/pseudo-synchronic: Animated writing in the comment feed of Nicovideo. *Japanese Studies*, 33(3):297–313, December 2013. doi: 10.1080/10371397.2013.859982

[27] A. C. Jones, R. J. Bennett, and S. Cross. Keepin' it real? life, death, and holograms on the live music stage. In A. C. Jones and R. J. Bennett, eds., *The Digital Evolution of Live Music*, pp. 123–138. Chandos Publishing, Oxford, UK, 2015. doi: 10.1016/B978-0-08-100067-0.00010-5

[28] T. Kaneko, H. Tarumi, K. Kataoka, Y. Kubochi, D. Yamashita, T. Nakai, and R. Yamaguchi. Supporting the sense of unity between remote audiences in VR-based remote live music support system KSA2. In *Proceedings of the 1st IEEE International Conference on Artificial Intelligence and Virtual Reality*, pp. 124–127. IEEE, New York, NY, 2018. doi: 10.1109/AIVR.2018.00025

[29] T. Kasuya, M. Tsukada, Y. Komohara, S. Takasaka, T. Mizuno, Y. Nomura, Y. Ueda, and H. Esaki. LiVRation: Remote VR live platform with interactive 3D audio-visual service. In *Proceedings of the 2019 IEEE Games Entertainment & Media Conference*, pp. 1–7. IEEE, New York, NY, 2019. doi: 10.1109/GEM.2019.8811549

[30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, pp. 3146–3154. NeurIPS Foundation, San Diego, CA, 2017.

[31] K. Y. Lam. The Hatsune Miku phenomenon: More than a virtual J-pop diva. *The Journal of Popular Culture*, 49(5):1107–1124, October 2016. doi: 10.1111/jpcu.12455

[32] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, New York, NY, 2008. doi: 10.1109/CVPR.2008.4587756

[33] A. Leavitt, T. Knight, and A. Yoshiba. Producing Hatsune Miku: Concerts, commercialization, and the politics of peer production. In P. W. Galbraith and J. G. Karlin, eds., *Media Convergence in Japan*, pp. 200–229. Kinema Club, New Haven, CT, 2016.

[34] K. H. Lee, M. G. Choi, Q. Hong, and J. Lee. Group behavior from video: a data-driven approach to crowd simulation. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 109–118. ACM, New York, NY, 2007. doi: 10.2312/SCA/SCA07/109-118

[35] C. H. Low. Assessing the future IP landscape of music's cash cow: What happens when the live concert goes virtual. *New York University Law Review*, 91(2):425–457, May 2016.

[36] X. Ma and N. Cao. Video-based evanescent, anonymous, asynchronous social interaction: Motivation and adaption to medium. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 770–782. ACM, New York, NY, 2017. doi: 10.1145/2998181.2998256

[37] A. MacQuarrie and A. Steed. Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. In *Proceedings of the 2017 IEEE Conference on Virtual Reality*, pp. 45–54. IEEE, New York, NY, 2017. doi: 10.1109/VR.2017.7892230

[38] L. Maystre and M. Grossglauser. Fast and accurate inference of Plackett-Luce models. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*, pp. 172–180. NeurIPS Foundation, San Diego, CA, 2015.

[39] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pp. 18–24. SciPy, Austin, TX, 2015. doi: 10.25080/Majora-7b98e3ed-003

[40] M. McLuhan. *Understanding Media: The Extensions of Man*. McGraw Hill, New York, NY, 1964.

[41] M. Mori, T. Moriya, and T. Takahashi. e-Ovation: A live karaoke system by using virtual reality tools. In *Proceedings of the 2017 International Workshop on Advanced Image Technology*, pp. 3A3:1–3A3:4. Multimedia University Press, Selangor, Malaysia, 2017.

[42] M. Müller, D. P. W. Ellis, A. Klapuri, and G. Richard. Signal processing for music analysis. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1088–1110, October 2011. doi: 10.1109/JSTSP.2011.2112333

[43] K. L. Nowak and F. Biocca. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 12(5):481–494, October 2003. doi: 10.1162/105474603322761289

[44] F. Ofli, E. Erzin, Y. Yemez, and A. M. Tekalp. Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis. *IEEE Transactions on Multimedia*, 14(3-2):747–759, June 2012. doi: 10.1109/TMM.2011.2181492

[45] A. T. Pereira, R. Prada, and A. Paiva. Improving social presence in human-agent interaction. In *Proceedings of the 2014 CHI Conference on Human Factors in Computing Systems*, pp. 1449–1458. ACM, New York, NY, 2014. doi: 10.1145/2556288.2557180

[46] F. Petitjean, A. Ketterlin, and P. Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693, March 2011. doi: 10.1016/j.patcog.2010.09.013

[47] N. Rio. The 29-year-old man from Kyoto University developing a virtual event market with VTuber (in Japanese). Business Insider Japan, `https://www.businessinsider.jp/post-171784`, July 2018. Accessed: October 14, 2019.

[48] G. Riva, F. Mantovani, C. S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcañiz. Affective interactions using virtual reality: The link between presence and emotions. *CyberPsychology & Behavior*, 10(1):45–56, February 2007. doi: 10.1089/cpb.2006.9993

[49] V. Schwind, P. Knierim, N. Haas, and N. Henze. Using presence questionnaires in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 360:1–360:12. ACM, New York, NY, 2019. doi: 10.1145/3290605.3300590

[50] W. R. Sherman and A. B. Craig. *Understanding Virtual Reality: Interface, Application, and Design*. Morgan Kaufmann, Burlington, MA, 2002.

[51] A. Shirai. REALITY: Broadcast your virtual beings from everywhere. In *Proceedings of the 46th ACM SIGGRAPH Conference – Appy Hour*, pp. 5:1–5:2. ACM, New York, NY, 2019. doi: 10.1145/3305365.3329727

[52] M. Slater, A. Sadagic, M. Usoh, and R. Schroeder. Small-group behavior in a virtual and real environment: A comparative study. *Presence: Teleoperators and Virtual Environments*, 9(1):37–51, February 2000. doi: 10.1162/105474600566600

[53] M. Slater and S. Wilbur. A framework for immersive virtual environments five: Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 6(6):603–616, December 1997. doi: 10.1162/pres.1997.6.6.603

[54] R. T. Solberg and A. R. Jensenius. Group behaviour and interpersonal synchronization to electronic dance music. *Musicae Scientiae*, 23(1):111–134, March 2019. doi: 10.1177/1029864917712345

[55] J. Stupacher, P.-J. Maes, M. Witte, and G. Wood. Music strengthens prosocial effects of interpersonal synchronization – if you move in time

with the beat. *Journal of Experimental Social Psychology*, 72:39–44, September 2017. doi: 10.1016/j.jesp.2017.04.007

[56] T. Tang, J. Jia, and H. Mao. Dance with melody: An LSTM-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1598–1606. ACM, New York, NY, 2018. doi: 10.1145/3240508.3240526

[57] H. Tarumi, T. Nakai, K. Miyazaki, D. Yamashita, and Y. Takasaki. What do remote music performances lack? In *Proceedings of the 9th International Conference on Collaboration Technologies and Social Computing*, pp. 14–21. Springer International Publishing, Cham, Switzerland, 2017. doi: 10.1007/978-3-319-63088-5_2

[58] D. Thalmann and S. R. Musse. *Crowd Simulation*. Springer-Verlag London, London, UK, 2007.

[59] K. Y. To. The voice of the future: Seeking freedom of expression through VOCALOID fandom. Master's thesis, The University of Texas at Austin, Austin, TX, May 2014.

[60] K. Tomoko. New movie showings give audiences chance to cheer, sing along. The Mainichi, `https://mainichi.jp/english/articles/20161023/p2a/00m/0et/006000c`, October 2016. Accessed: October 15, 2019.

[61] T. Turino. *Music as Social Life: The Politics of Participation*. University of Chicago Press, Chicago, IL, 2008.

[62] G. Tzanetakis and P. R. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002. doi: 10.1109/TSA.2002.800560

[63] M. Usoh, E. Catena, S. Arman, and M. Slater. Using presence questionnaires in reality. *Presence: Teleoperators and Virtual Environments*, 9(5):497–503, October 2000. doi: 10.1162/105474600566989

[64] A. M. Webb, C. Wang, A. Kerne, and P. César. Distributed liveness: Understanding how new technologies transform performance experiences. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pp. 431–436. ACM, New York, NY, 2016. doi: 10.1145/2818048.2819974

[65] Q. Wu, Y. Sang, and Y. Huang. Danmaku: A new paradigm of social interaction via online videos. *ACM Transactions on Social Computing*, 2(2):7:1–7:24, June 2019. doi: 10.1145/3329485

[66] M. Xu, H. Jiang, X. Jin, and Z. Deng. Crowd simulation and its applications: Recent advances. *Journal of Computer Science and Technology*, 29(5):799–811, September 2014. doi: 10.1007/s11390-014-1469-y

[67] E. Yilmaz, Y. Y. Çetin, Ç. E. Erdem, T. Erdem, and M. Özkan. Music driven real-time 3D concert simulation. In *Proceedings of the 2006 International Workshop on Multimedia Content Representation, Classification and Security*, pp. 379–386. Springer-Verlag, Berlin, Heidelberg, 2006. doi: 10.1007/11848035_51

[68] C. Yuan and H. Yang. Research on k-value selection method of k-means clustering algorithm. *J – Multidisciplinary Scientific Journal*, 2(2):226–235, June 2019. doi: 10.3390/j2020016

[69] R. B. Zajonc. Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2):1–27, June 1968. doi: 10.1037/h0025848

[70] R. B. Zajonc. Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, 10(6):224–228, December 2001. doi: 10.1111/1467-8721.00154