AlterEcho: Loose Avatar-Streamer Coupling for Expressive VTubing

Man To Tang*
Purdue University

Victor Long Zhu
Purdue University

Voicu Popescu
Purdue University

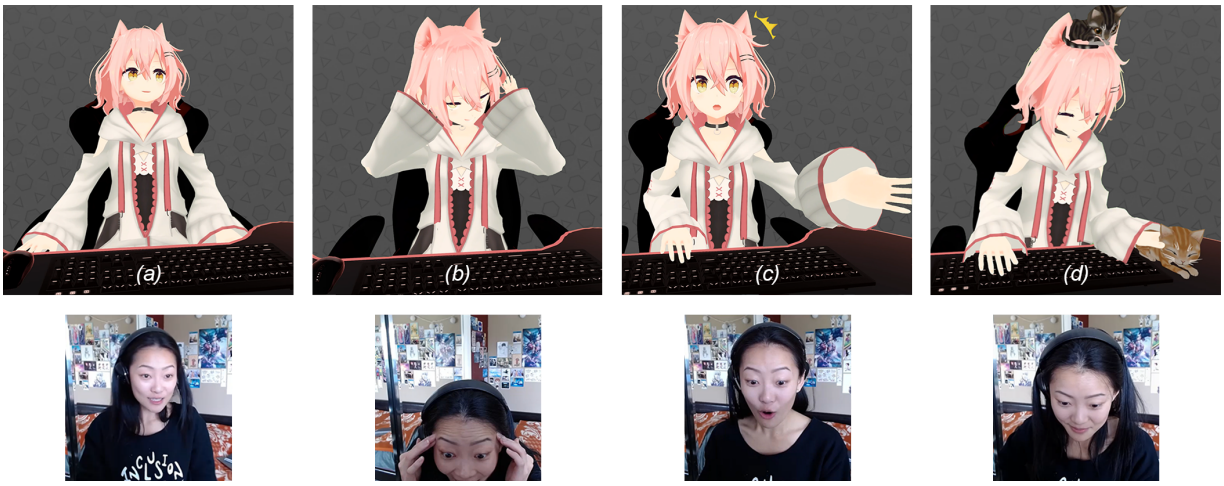

Figure 1: *AlterEcho* VTuber avatar animation (top) and corresponding streamer video frames (bottom), which are *not* shown to the viewer, and are shown here for illustration purposes. The avatar's coupling to the streamer is looser than in conventional motion capture, with the avatar making gestures that are identical (*a*), similar (*b* and *c*), or completely different (*d*) from those of the streamer.

Abstract

VTubers are live streamers who embody computer animation virtual avatars. VTubing is a rapidly rising form of online entertainment in East Asia, most notably in Japan and China, and it has been more recently introduced in the West. However, animating an expressive VTuber avatar remains a challenge due to budget and usability limitations of current solutions, i.e., high-fidelity motion capture is expensive, while keyboard-based VTubing interfaces impose a cognitive burden on the streamer. This paper proposes a novel approach for VTubing animation based on the key principle of loosening the coupling between the VTuber and their avatar, and it describes a first implementation of the approach in the *AlterEcho* VTubing animation system. *AlterEcho* generates expressive VTuber avatar animation automatically, without the streamer's explicit intervention; it breaks the strict tethering of the avatar from the streamer, allowing the avatar's nonverbal behavior to deviate from that of the streamer. Without the complete independence of a true *alter ego*, but also without the constraint of mirroring the streamer with the fidelity of an *echo*, *AlterEcho* produces avatar animations that have been rated significantly higher by VTubers and viewers ($N = 315$) compared to animations created using simple motion capture, or using *VMagicMirror*, a state-of-the-art keyboard-based VTubing system. Our work also opens the door to personalizing the avatar persona for individual viewers.

Keywords: VTuber, automatic avatar animation, live streaming.

Index Terms: Human-centered computing—Int. systems and tools

*Corresponding author: tigerhix@gmail.com

1 Introduction

Live streaming, or live video broadcasting through the Internet, has gained worldwide popularity in the past decade [25]. Platforms like YouTube and Twitch allow anyone with a webcam and an Internet connection to become a *streamer*, sharing their life experiences and creative content while engaging with an online audience in real-time. More recently, advances in motion capture and computer animation have empowered streamers to represent themselves with virtual avatars without revealing their real self. Virtual YouTubers, or *VTubers*, are streamers who embody a virtual avatar and role-play a specially designed persona [33]. Originating from East Asia where the subcultures of anime and manga are prevalent, the VTuber community has since rapidly grown and expanded, reaching a worldwide audience across cultural and language barriers. Kizuna Ai, a VTuber who is considered the first and most popular VTuber to date, has 4 million followers around the globe and was selected as a tourism ambassador for the Japanese National Tourism Organization in 2018 [28]; by July 2020, there were more than 13,000 active VTubers from Japan, and more than 6,000 from China [24, 43]. It is estimated that the VTuber industry in Japan alone will exceed 50 billion yen (471 million USD) by 2022 [2]. The year 2020 also saw VTubers reach popularity in the West [15]; the English-speaking VTuber Gawr Gura, for example, reached 1 million subscribers on YouTube in only 41 days [20].

A key contributing factor to a VTuber's streaming performance is their animated avatar; it must vividly portray their persona, in order to attract an audience that finds their personality and appearance appealing [33]. For example, VTuber Kizuna Ai's persona is a "recently-developed advanced artificial intelligence" whose words and actions are naive [33]. Thus, her avatar always has a cheerful and confident look and exhibits rich facial expressions [29].

However, animating an expressive VTuber avatar in real-time has remained a challenge. In a popular approach, streamers animate their avatars by relying on a series of keyboard shortcuts to trigger specific gestures or animations; but such a keyboard interface places

a significant cognitive burden on the streamer, which not only stifles their creativity and spontaneity, but also potentially results in an off-putting performance that feels contrived. Another approach is to wear a full-body motion capture suit that tracks one's body movements in real-time. The technology, however, is expensive, uncomfortable, and requires a large physical space. Furthermore–and probably most importantly–such real-time motion capture imposes an upper limit on the avatar animation, which can at best mirror, but never surpass, the streamer's movement. Removing this constraint of reality is important for VTubers to be able to express themselves in new ways that can more accurately fit their creative vision.

In this paper, we propose a novel approach to VTubing based on the central idea of loosening the coupling between the avatar and streamer, and we describe the first implementation of our approach in the *AlterEcho* VTubing animation system. In *AlterEcho*, the streamer is captured with a phone (video) and a headset microphone (audio). The phone provides basic motion capture data, including head position, orientation, and facial landmarks. Keyboard and mouse input events from the computer used for streaming are also captured. The video, audio, motion capture data, input events, and the desired *persona parameters* are fed into the *AlterEcho* system which then animates the avatar (Fig. 1). The avatar makes *tethered gestures* based on motion capture data (*a*); *inferred gestures* based on speech recognition, acoustic analysis, facial expression recognition, and avatar persona parameters (*b and c*); as well as *impromptu gestures* based on persona parameters (*d*).

We evaluated *AlterEcho* in a study ($N = 315$) with VTubers, non-VTuber streamers, and VTuber viewers through an online survey. The first part of the study compared three versions of the same segment from a VTubing stream, which had identical voice tracks, and differed only in the avatar animation. Participants rated the avatar animation generated by *AlterEcho* higher than the two control conditions, and a majority of participants rated the *AlterEcho* avatar as the most engaging, natural, and preferred avatar. The second part of the study evaluated *AlterEcho*'s ability to adapt the avatar's persona on the shy/introverted to confident/extroverted continuum. Participants were shown the same animation twice, with identical voice tracks, but with the avatar's nonverbal behavior "shier" in one animation, and "more outgoing" in the other. Participants were provided eight persona adjectives that are correlated with the two ends of the shy/unconfident to outgoing/confident continuum; on average 7.2 of the 8 adjectives were correctly attributed.

We also refer the reader to the accompanying video which shows the animations used in our user study along with the nonverbal animation repertoire of *AlterEcho*.

2 Related Work

While live-streaming culture and practices have been extensively studied [25, 34, 41] and live-streaming systems for various applications have been proposed [26, 55], VTubing is still a relatively new phenomenon; there is little prior research aimed specifically toward it. In a very recent paper, Lu et al. [33] have uncovered the nuances in the perceptions and attitudes of VTuber viewers compared to those of non-VTuber viewers; Otmazgin [40] discusses how VTubers, as an emerging new media industry in Japan, have integrated technology and fandom and have blurred the boundary between reality and virtual reality; Bredikhina [12] surveyed VTubers on identity construction, reasons for VTubing, and gender expression. Yet, there is a lack of technical discourse on VTubing.

2.1 Current VTubing Solutions

In the following, we discuss from a practical standpoint current VTubing solutions for both independent VTubers, who pursue VTubing as a hobby, and commercial VTubers, who are associated with a company and for whom VTubing is a job.

Full-body motion capture. Commercial VTuber agencies like Hololive or A-SOUL employ industry-grade motion capture systems for full-body tracking [6, 7]. The streamer wears a full-body motion capture suit, and their body movements and facial expressions are streamed and played on a rigged 3D avatar in real-time. Due to technical complexity, there are usually no interactions between the avatar and the virtual environment; even in their rare occurrences, they are often limited, e.g., grabbing and placing a static object. Basic cinematography may be achieved by either manually or automatically rotating between pre-configured virtual world positions, or by using a tracked real-world object that the streamer can freely move and rotate to adjust the virtual camera. Certainly, fully motion-captured avatars move realistically, but most VTubers cannot afford a full-body motion capture setup. The motion capture suit and/or gloves can also be unergonomic in streaming scenarios like gaming, which requires extensive keyboard and mouse use.

As a more affordable alternative, some VTubers may employ consumer VR headsets with additional trackers for full-body tracking [47]. However, they are less accurate and often constrain the streamer's actions, e.g., the streamer has to hold the VR controllers at all times, still rendering some tasks such as non-VR gaming impossible. Furthermore, facial expression data are not captured.

Head/Facial motion capture. Most independent VTubers do not consider body tracking at all; indeed, as modeling and animating 3D full-body avatars are expensive, many VTubers choose to instead create and embody a face-rigged 2D avatar, and purely rely on head and face motion capture using a webcam or phone camera [48] to bring their avatar to life. Popular software include *FaceRig* (RGB image-based tracking) [18] or *VMagicMirror* (ARKit-based tracking) [9]. There also exists VTubing software (e.g., *Luppet* [8]) that provides hand tracking at an additional cost through a LeapMotion sensor, but not without limitations; for instance, the sensor cannot capture the streamer crossing their arms.

Keyboard interfaces. Most consumer VTubing software systems allow the streamer to apply baked animations on the avatar by pressing keyboard shortcuts, giving the avatar character [33]. However, the streamer can only remember so many keyboard shortcuts, and triggering the correct animation at the correct time when the streamer needs to concentrate on another matter, e.g., fighting a video game boss, can become unmanageable.

Despite its many limitations, a head/facial motion capture setup that supports keyboard-triggered motions still remains largely popular for both its low cost and light burden on the streamer–they only need a webcam or phone in front of them. Thus, it is considered a more cost-effective solution than using a full-body motion capture setup and is more widely used. In fact, even commercial VTubers who can afford full-body motion capture would often switch to using head/facial motion capture only so that they can comfortably live-stream from home.

2.2 Conventional Streaming vs. VTubing

For both the audience and the streamer, VTuber streams are more than just live streams with a virtual avatar: whereas audiences of real celebrities are interested in the streamer's *actual* appearance and their complex, implicit personality [22], VTuber audiences are attracted to the *fictional*, often stereotypical, appearance and personality traits of a simple, "flat," and explicit VTuber persona, which often draws inspiration from the subcultures of anime and comics, i.e., the Otaku community; consequently, the VTuber community is also more accepting to actions taken by VTubers that may not be viewed upon as favorably by conventional communities [33]. For instance, the VTuber Inugami Korone has a quirk of asking the audience to offer their fingers [17], a reference to the Japanese mafia practice "*Yubitsume*" ("finger shortening") [11]. Her violent and dark catchphrase is unlikely to be as easily accepted by the conventional streaming community, as the virtual and more lighthearted

nature of VTubers makes their audience more tolerant of behaviors outside of the norm. The VTuber community's bigger focus on the "*moe* elements" [19] of the avatar rather than on the actions and characteristics of the streamer grants the VTuber more freedom in their creative expression.

Thus, these nuances naturally imply a difference in technical requirements between conventional streaming and VTubing. Indeed, existing VTubing systems have focused on reproducing the streamer's actions through the avatar as faithfully as possible. For example, when *Luppet* users raise their hand and wave to greet the audience, the LeapMotion sensor captures the hand image and generates hand landmarks to pose the avatar in the same fashion. However, even a system with perfect motion capture would fall short of ideal support for VTubing. Consider a VTuber persona who is talkative and makes many hand gestures, or a persona who is shy and always looks down during speech. While such nonverbal communication cues *can* be captured from the streamer's actions (assuming a full-body motion capture setup), it requires the streamer to have good acting skills to maintain the regular body movements associated with the persona; for hours-long streams, constantly maintaining the persona is both challenging and exhausting. As such, a VTubing system must not only capture reality, i.e., the streamer's actions, but also surpass reality, i.e., the avatar behaves differently to that of the streamer, to more effectively actualize the desired VTuber persona.

3 THE *AlterEcho* VTUBING ANIMATION SYSTEM

We first frame the challenges and requirements of VTuber avatar animation and introduce our approach, designed with formative feedback from four VTubers active on Twitch [54] and Bilibili [10], as well as the VTubing experience of one of the paper authors (Sect. 3.1). Then we give the *AlterEcho* system architecture (Sect. 3.2), describe the avatar animation repertoire (Sect. 3.3), and explain how *AlterEcho* achieves automatic animation (Sect. 3.4).

3.1 Design: Challenges, Requirements, and Approach

VTuber audiences have expressed appreciation for avatar motions and expressions that appear natural [33]; they desire subtle nonverbal communication cues, rich and natural facial expressions, as well as interactions between the avatar and the virtual environment, all of which cannot be easily provided with simple motion capture setups based on a webcam or a phone. However, relying on complex motion capture setups is also problematic due to budget and space limitations, along with the discomfort of wearing body suits or gloves for long streams.

While current webcam-based and phone-based VTubing systems do allow user-defined hotkey-triggered avatar animations, remembering to manually trigger such animations is a mentally burdening task for the streamer: "*Manually switching my face expressions while playing a game can be very distracting, so I almost do not [switch facial expressions] at all, unless I am solely chatting with the audience*" (VTuber 3). Furthermore, when the streamer desires animations that correspond to unscripted, authentic reactions, accurately timing such animations is both essential and challenging. For example, consider a common scenario in which the streamer is startled by elements in a computer game. It takes approximately 0.5s for a person to recover from their startle reflex, and it can take up to 30s for them to fully recover their decision-making ability [52]. Thus, there is an inevitable delay between the startling event and the time the streamer is able to press a hotkey to trigger an avatar reaction, such as a facial expression of shock or a body gesture of covering the face. A delay in the avatar's reaction is a key facilitating factor for the uncanny valley phenomenon [53], which can lead to a less natural performance that disengages the viewers.

Even with a perfect motion capture system, a VTuber avatar cannot reach its full potential: the virtual world of the avatar can only approach, and never surpass, the real world. For example, the avatar should be able to interact with objects and characters that exist only in the virtual world and not in the real world. The avatar's abilities should not be restricted to those of the streamer: the avatar could be a better actor when role-playing a designed VTuber persona, and, for hours-long streams, the avatar's endless energy could help the streamer stay in character.

Consequently, a VTubing animation system should be designed to satisfy the following requirements:

- *Automatic.* The avatar animation should be generated without explicit intervention from the streamer.

- *Low cost.* The system should be implemented from inexpensive, unobtrusive, and compact components.

- *Support for real / virtual world differences.* The system should produce animation suitable for a virtual world that is controlled, but not limited, by the real world.

- *Adaptable.* The avatar persona should be parameterized such that the streamer can set it based on their preferences.

We propose *AlterEcho*, a VTuber avatar animation system designed to satisfy the requirements above based on the defining idea of loosening the coupling between the streamer and the avatar. The streamer motions are captured with an inexpensive system, i.e., a phone; instead of simply retargeting the captured motion to the avatar, the animation is generated by also taking into account the streamer's actions and speech, the virtual environment (which could be significantly different than the real world), and the specified avatar persona parameters. The resulting avatar animation not only includes elements from the captured motion, but also includes new elements blended in to leverage the richness of the virtual world, achieving and sustaining a desired avatar persona.

3.2 Architecture

Fig. 3.1 gives an overview of the *AlterEcho* system. The VTuber sits in front of a computer, e.g., a laptop or a desktop, wearing a headset with a microphone. The VTuber interacts with the computer using a mouse and keyboard interface, for example to play a game, or to show videos as part of their VTubing stream. The computer also runs the *AlterEcho* system. The VTuber head and torso are captured with a video camera and their speech is captured with the headset microphone. The video stream (1 in Fig. 3.1), the audio stream (2), the mouse and keyboard events (3), and the desired values of the avatar persona parameters (4) are input to the *AlterEcho* system.

The VTuber's head and facial features are tracked by any low-cost, real-time face tracking system that computes head position, head orientation and facial landmark data, e.g., OpenCV [39], MediaPipe [23] or ARKit [3]. A machine-learning facial expression recognition (FER) system then classifies the facial landmark data into facial expressions. Compared to the traditional approach of running the machine learning classification on video frames [31], using the facial landmark data has the advantage of faster training and increased classification robustness [46]. We cover in more detail our particular choice of components in Sect. 4.1. Speech is converted to text with punctuation and word-level timestamps [50], and the audio stream undergoes further acoustic analysis to measure the volume and pitch of each word. The head and face mocap data along with the recognized facial expressions (5), the results from speech analysis (6), and the keyboard and mouse events (3) are buffered to be used by the Gesture Triggering module to decide whether the preconditions of specific avatar gestures are met (7). The *AlterEcho* avatar gesture repertoire is described in Sect. 3.3, and the Gesture Triggering module is described in Sect. 3.4.1.

AlterEcho preloads graphics assets and animations (8). The graphics assets include the computer animation characters of the avatar and their companions (e.g., a cat), and the models of the keyboard, mouse, and desk. The graphics assets (9), the animation clips (a), and inverse kinematics (IK) algorithms [4] (b) are piped together with the triggered gesture data (c) into the Animation Integration

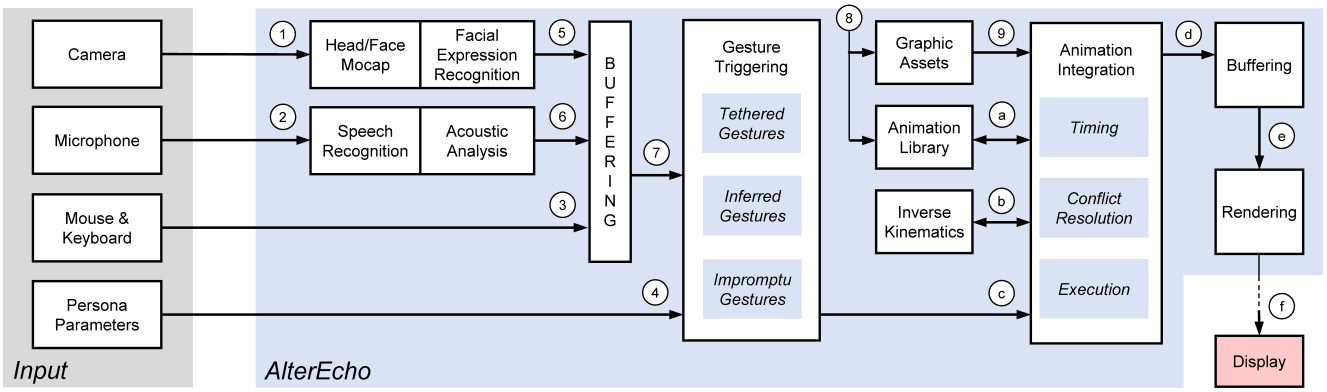

Figure 2: *AlterEcho* architecture.

module which implements avatar animation (Sect. 3.4.2).

The virtual environment, including the avatar, are queued up for rendering (d), and are rendered just in time (e) to maintain a constant time offset between real time and rendering time. This delay gives the Animation Integration module the ability to arbitrate between tethered gestures on one hand, and inferred and impromptu gestures on the other, to avoid lower priority gestures being truncated when interrupted by higher priority gestures. This delay (see Sect. 3.4.1) is fundamental to our looser avatar-streamer coupling, as it allows the avatar to temporarily diverge and then to gracefully sync back up with the streamer. Finally, the rendered frames are streamed to the viewer over the Internet (f).

3.3 Nonverbal Animation Repertoire

We categorize the *AlterEcho* avatar's nonverbal behavior into tethered, inferred and impromptu gestures, based on how loosely the avatar's actions are connected to those of the streamer. We use the term *gesture* to encompass all nonverbal behaviors, including hand, arm, head, face, body, and gaze movements. The accompanying video catalogues *AlterEcho*'s animation capabilities.

Tethered gestures. The avatar lipsyncs, makes facial expressions, moves their head, and uses the keyboard and mouse, mirroring the captured movements of the streamer. Examples of tethered gestures include the head position and the view direction in Fig. 1 (*a*).

Inferred gestures. The avatar also makes gestures which are not a direct replica but rather a result of the streamer's actions. *Beat gestures* are inferred from the streamer's speech, and, although they do not carry semantic information, they enhance the rhythm of the speech and serve as attention cues [27]. *Iconic gestures* are also inferred from the speech, and they evoke specific ideas based on nonverbal communication conventions [56]; for example, the avatar waves their hand to greet the viewer, shrugs their shoulders to express doubt or indifference, or grabs their head when startled . *Emotion gestures* are facial expressions that the avatar makes based on the inferred emotional state of the VTuber. Examples of inferred gestures in Fig. 1 are the frustration gesture (*b*) and the shocked expression with the beat gesture (*c*).

Impromptu gestures. The naturalness and personality of the avatar is further enhanced with gestures that are independent of the streamer. For example, the avatar scratches their face or neck, turns their head to avoid eye contact, slouches or stands up straight, crosses their arms, or even pets a virtual cat that has no real world counterpart–see (*d*) in Fig. 1.

Persona parameters. Finally, *AlterEcho* allows the streamer to tune the avatar persona along two dimensions: energy level and degree of extroversion. These high-level controls can produce the typical "extroverted and energetic" and "introverted and reserved" avatars, while also allowing for a shy but energetic avatar who avoids eye contact while making numerous nervous gestures, or an outgoing but calm avatar who confidently delivers dry humor with an indifferent face. Fig. 3 shows the same frame with degree of extroversion set to minimum (left) and maximum (right).

Figure 3: Same frame with different persona parameter values: shy/introverted (left), confident/extroverted (right).

3.4 Automatic Animation

In this section we describe how *AlterEcho* animates the VTuber avatar automatically by first deciding which gestures to trigger (Sect. 3.4.1) and then by integrating the gestures (Sect. 3.4.2) into the final avatar animation.

3.4.1 Gesture Triggering

The triggering of gestures depends on the streamer's tracked actions and/or the input persona parameter values (Fig. 3.1). We have formulated the avatar persona with two high-level parameters: *energy level* K_e, and *degree of extroversion* K_x, with values in the $[-1,1]$ range. Both parameters control the default facial expression, and the set of allowable gestures, e.g., low energy and introverted avatars do not make beat gestures. In addition, K_e controls gesture frequency, and K_x controls body posture (i.e., slouching or sitting straight), amplitude of the beat gestures (i.e., narrow or broad), IK targets of the chest and elbows, magnitude of the tethered body movements, and eye contact, based on conventional nonverbal expressions of extroversion [13, 36]. For example, a shier avatar slouches more and keeps elbows closer to body to be less visible; a more confident avatar sits straight and keeps their elbows away from their body to maximize their perceived size.

A gesture is triggered when its preconditions are met. In the following, we discuss the preconditions for different categories of gestures. Note that a triggered gesture will not necessarily be played;

which triggered gestures to integrate and execute is decided in animation integration (Sect. 3.4.2).

Tethered gestures. These gestures are triggered based on motion capture data, as well as on the keyboard and mouse events, and do not depend on persona parameters. There is a one-to-one mapping between the keys of the real and virtual keyboard; a key press event generates a corresponding key press gesture for the left or right hand of the avatar. The captured mouse events include cursor movement and button events, and the avatar manipulates the mouse accordingly. A more loosely tethered gesture is triggered when the VTuber's head position becomes distant from the screen, indicating that the VTuber let go of the keyboard and mouse, which the avatar follows.

Inferred gestures. Inferred gestures are triggered based on speech recognition, acoustic analysis, facial expression recognition, and persona parameters. Iconic gestures are triggered by the utterance of specific keywords and phrases recognized from the audio input, e.g., the avatar shrugging their shoulders when saying "well"/"ok," or the avatar grabbing their head with their hands to show frustration when shouting out "wait" followed by a pause longer than a few seconds. Beat gestures are triggered with probability P_β by words uttered with a volume higher than a threshold; in our implementation, it is computed as one standard deviation above the average volume of word utterances in the latest minute of audio. Beat gestures are also triggered with probability P_β for the first word of a sentence if no beat gestures have been triggered in the past t_β seconds. P_β is positively related to K_e and K_x, and t_β is negatively related to K_e and K_x, so a more energetic and extroverted avatar is allowed to make more frequent beat gestures. Emotion gestures, such as joyful, sad, or angry facial expressions, are triggered by changes from a neutral expression, as detected and classified by our FER model.

To allow for more control and flexibility, each inferred gesture has its own K_e and K_x triggering thresholds, g_e and g_x. Both of the following conditions must be met for the gesture to be triggered: (1) $K_e \geq g_e$, and (2) $K_x \geq g_x$ ($g_x > 0$) or $K_x \leq g_x$ ($g_x < 0$).

Impromptu gestures. Like inferred gestures, each impromptu gesture has its own thresholds g_e and g_x for the persona parameters K_e and K_x. Impromptu gestures are triggered per second with probability P_ι, which is negatively related to K_e, so that more energetic avatars make more impromptu gestures.

3.4.2 Animation Integration

The Gesture Triggering module (Sect. 3.4.1) provides a list of gestures which need to be integrated into the final avatar animation (Fig. 3.1). This requires timing the gestures, arbitrating whether the gestures should be executed while resolving potential conflicts, and finally executing the gestures to produce a smooth animation.

The starting time of a gesture is based on when its preconditions are met; applying an offset can also be necessary, e.g., to better synchronize the beat gestures with the streamer's speech. Most gestures have a duration derived from the length of the animation clip, or from the time it takes to reach the target pose through IK, and always *restore* the avatar into their natural pose. However, there are *non-restoring* gestures that leave the avatar in a target state for an indefinite amount of time, thus not having a duration. For example, beat gestures are non-restoring; they hold until the avatar makes the next hand gesture. Thus, the timing submodule must keep track of non-restoring gestures and return the affected body parts to their natural poses when they persist beyond a specified time threshold, e.g., 3s for beat gestures.

The arbitration submodule decides upon whether a triggered gesture should indeed be executed, and resolves conflicts between candidate gestures, i.e., gestures may act simultaneously on the same skeletal joints. First, all inferred and impromptu gestures have an execution probability, which is connected to the persona parameters, to further control their frequency and add onto the avatar's naturalness; they also have a cooldown period during which they are not eligible for being repeated. Once candidate gestures are decided, conflicts are resolved based on the following priority order, from high to low: tethered, emotion, iconic, beat, and finally impromptu gestures. To guarantee that a low-priority gesture completes its execution without a high-priority gesture interrupting it, we use a sliding window that affords a t_d seconds glimpse into the "future;" the theoretical lower bound of t_d is thus the maximum duration of a gesture. As a consequence, the execution of all gestures are delayed by t_d. Note that this delay is **not** the delay in the avatar's reaction as perceived by the audience, which we discussed in Sect. 3.1; rather, it should be seen as an extension to the *end-to-end delay* associated with live-streaming. Implications are further discussed in Sect. 5.

Most gestures are implemented (1) through forward kinematics, e.g., head rotations, (2) through IK, e.g., grabbing the mouse or pressing a key, or (3) by playing animation clips in the animation library. The gaze of the avatar is implemented by aiming the gaze away from or towards the camera based on the degree of extroversion K_x. Head, spine, and pole IK targets are transformed based on K_x to adjust the default body pose. Facial blendshapes are interpolated based on motion capture data and offsetted based on K_e and K_x to modulate the facial expression [3, 16].

4 Results and Discussion

We compared avatar animations produced by *AlterEcho* to those produced by other approaches in a study with $N = 315$ participants from the VTuber community (Sect. 4.2).

4.1 Implementation Overview

AlterEcho (Fig. 3.1) was implemented using Unity 2019.4.13f1 and Python 3. Head position, head orientation, and facial landmarks of the streamer are tracked with an iPhone 11 using face anchor data from Apple's ARKit [3], which is piped into Unity. We chose ARKit as our tracking solution as it is accessible and cost-effective: the VTubers we collaborated with all either own or have access to an iPhone/iPad that supports ARKit face tracking, which achieves real-time performance with high accuracy [51]. The ARKit face anchor data are then classified into one of four facial expressions, corresponding to emotions joy, sadness, anger, and neutral (no emotion). We used a support vector machine facial expression recognition (FER) model [21] trained with our own dataset, prepared by recording 10 minutes of the authors making different facial expressions, and labeling each frame of ARKit data with the corresponding facial expression. While deep learning FER approaches that directly work with video frames have been extensively studied [31], we chose to take as input the ARKit data for performance and robustness. Our FER model detects facial expressions for individual frames, which are filtered and used to trigger emotion gestures. Audio is continuously processed using Python. Speech recognition and transcription is performed using Azure's Speech to Text cloud service [35]; the audio is also processed locally with aubio [5] and PyAudio [42] to perform acoustic analysis, i.e., to measure the volume and pitch of each word. Keyboard and mouse events are captured using Python and piped into Unity.

The Gesture Triggering and Animation Integration modules are implemented in Unity as a C# script. In our implementation, the delay t_d is set to 15 seconds. For most gestures, their triggering thresholds g_e and g_x are set to 0 and -1 (effectively no thresholds), except for a few "outgoing gestures," e.g., shrugging, stretching the arms, etc., for which g_x is set to 0.5. The probability P_β of triggering a beat gesture is determined with a linear equation, where the range of the extroversion coefficient K_x is shifted from [-1, 1] to [0, 2] and weighted by w_β (Equation 1). The probability P_ι of triggering an impromptu gesture and the minimum time t_β between two beat gestures are similarly determined. The weights w_β, w_ι, and w_t are

Figure 4: Frames from the three-way comparison from our user study.

Table 1: Participant weekly VTubing watch time.

Watch time	VTubers	Other streamers	Viewers	
0	2	8	34	
$<$ 1 hr	9	14	42	
2-5 hrs	24	15	74	
5-10 hrs	11	2	26	
$>$ 10 hrs	10	10	34	
Total	**56**	**49**	**210**	**315**

Table 2: Participant weekly live-stream and video creation time.

Live-stream	VTubers	Other streamers	Video creation	VTubers	Other streamers
0 hrs	6	33	0 hrs	15	3
$<$ 5 hrs	10	9	$<$ 1 hr	23	38
5-20 hrs	15	6	1-2 hrs	7	4
$>$ 20 hrs	25	1	$>$ 2 hrs	11	4
Total	**56**	**49**	**Total**	**56**	**49**

determined to be 0.25, 0.06 and 15, respectively, through trial and error.

$$\begin{aligned} P_\beta &= w_\beta(K_x+1) \\ P_I &= w_I(K_e+1) \\ t_\beta &= w_t(1-(K_x+K_e+2)) \end{aligned} \tag{1}$$

After Unity handles buffering and rendering, *AlterEcho* is streaming-ready using existing streaming software such as *OBS Studio* [38]. Note that the audio output of the streaming software needs to be configured to add a delay of 15s to synchronize with the gestures (see Sect. 3.4.2).

4.2 User Study

We conducted a user study in May 2021 with $N = 315$ participants to compare animations produced with *AlterEcho* to animations produced with two other approaches, and to gauge *AlterEcho's* ability to modulate the avatar persona.

Participants. In order to reach a global audience, the study was implemented with an online survey using survey platforms Qualtrics [44] ($N_Q = 82$) and Wen Juan Xing [14] ($N_W = 233$). We recruited the participants on Reddit and National Geographic of Azeroth (NGA) through their dedicated VTubing forums [37, 45] to reach both VTubers and viewers. 84 of the participants identified as female and 200 identified as male. 56 were VTubers, 49 were other streamers or video content creators, and 210 were viewers not involved in content creation. Table 1 and Table 2 show that our participants have substantial VTubing experience. On average, each participant took 12.8 minutes to complete the survey.

4.2.1 Comparison to Other Approaches

The first part of the survey is a controlled within-subject user study to compare *AlterEcho* animations to two control conditions. The animations used in the comparison were generated as follows.

Conditions. Since chatting and gaming are two main activities of a VTubing stream [33], we contracted an experienced VTuber to perform a typical *Let's Play* session, in which the VTuber played a computer game and entertained the audience with related and unrelated commentary. The VTuber's speech was recorded with a headset microphone and then mixed into the computer's screen capture. The keyboard and mouse events were captured with a Python script. The streamer was also captured with an iPhone 11 that provided head and facial landmark motion capture data.

The audio/video capture, the events trace, and the mocap data were used to create animations for three conditions. For the first control condition (C1), the animation was created by directly applying the mocap data to the avatar. For the second control condition (C2), the avatar was animated using *VMagicMirror* [9], a popular VTubing system that takes mocap data as input and enhances it with animations triggered through key presses; we triggered animations manually and offline to make the best use of the *VMagicMirror* features and without putting pressure on the streamer's real-time performance. For the experimental condition (E), the avatar was animated using *AlterEcho*. All three animations have exactly the same audio track of the streamer's voice mixed in with the video game sounds. Each animation is 2min23s long. Note the animations do not show the streamer, just their avatar. During game play, the animation is juxtaposed to the computer screen capture. Please refer to Fig. 4 and to the accompanying video for an illustration of the animations used in the three conditions.

Experimental procedure. First, the participant watched the three animations one at a time, in random order. After watching each video, the participant was immediately asked to rate on a five-point Likert scale whether "the avatar was engaging" (*Engaging* Table 3), "the avatar was natural" (*Natural*), and if they "would be happy with this avatar in [their] streaming/VTubing" (*Preferable*). Then, the participant was shown all three videos simultaneously, with two videos at the top and one video at the bottom (see accompanying video), in the order they just watched. The participant was asked to select the video in which they thought the avatar was "most engaging," "most natural," and the animation that had their "preferred" avatar. The participant could provide optional free-form feedback.

Results and discussion. We converted the Likert scale answers into numerical values from 1 ("Strongly Disagree") to 5 ("Strongly

Table 3: Individual animation ratings, all participants ($N = 315$).

	Engaging		Natural		Preferable	
	Mean	S.Dev.	Mean	S.Dev.	Mean	S.Dev
Mocap only (C1)	3.25	1.02	2.80	1.09	3.04	1.10
VMagicMirror (C2)	3.55	0.92	3.11	1.13	3.22	1.11
AlterEcho (E)	**4.16**	0.89	**4.02**	1.04	**3.94**	1.08

Table 4: Individual animation ratings, VTuber participants ($N = 56$).

	Engaging		Natural		Preferable	
	Mean	S.Dev.	Mean	S.Dev.	Mean	S.Dev
Mocap only	3.43	0.99	2.77	1.10	3.23	1.13
VMagicMirror	3.77	0.83	3.20	1.13	3.52	1.18
AlterEcho	**4.25**	0.81	**3.95**	1.07	**4.18**	0.99

Agree"). The average individual ratings of the animations are given in Table 3. Table 4 gives the ratings for the VTuber participants. *AlterEcho* was rated higher than *VMagicMirror*, which was rated higher than Mocap only, for all three measures, by all participants, as well as by the VTuber subgroup. Average VTuber ratings were slightly higher than the average ratings over all participants.

Since our data are not normally distributed, we compared the three conditions for each of the three dependent variables in Table 3 with a Friedman test followed by a post hoc test. There was a statistically significant difference between the three conditions for the dependent variables *Engaging* ($\chi^2(2) = 169.33$, $p < 1e{-}5$), *Natural* ($\chi^2(2) = 177.662$, $p < 1e{-}5$), and *Preferable* ($\chi^2(2) = 136.064$, $p < 1e{-}5$). Post hoc analysis was conducted to investigate differences between pairs of conditions. We used the Wilcoxon signed-rank test with Bonferroni correction, resulting in a significance level set at $p < .017$. Table 5 shows that *AlterEcho* (E) ratings are significantly higher than those of of Mocap only (C1), and of *VMagicMirror* (C2).

Table 6 shows that a majority of participants selected the *AlterEcho* animation as the most engaging, most natural, and as their preferred avatar. A similar pattern was also recorded for the VTuber participant subgroup.

Indeed, most participants praised the *AlterEcho* animation as "*vivid and natural*" (P254), "*more engaging because of the body movements and the fact [the avatar] can move its hands too*" (P261), and "*the one [they] ended up re-watching the most between the three [videos]... [their] eyes kept going back to [AlterEcho animation]*" (P265). Some participants thought the *AlterEcho* animation was produced with full-body motion capture: "*the tracking ... is very well-managed ... [but] it will come with a cost of spending more time and resources on equipment and will also demand more control of [the streamer's] own body*" (P279). As a participant noted, "*while the increased movement does lead to the characteristic unnatural/uncanny moments that are unavoidable for highly tracked models, [the AlterEcho avatar] still plays far better into creating a character and telling the audience about that character*" (P300).

However, some participants felt that the *AlterEcho* avatar was distracting: "*[It] feels like the model swayed/trembled too much left and right*" (P256), "*[It] has too many motions... The viewer may want to focus on the game but the substantial amount of motions may result in visual fatigue, because the viewer has to pay attention to both the game and the avatar*" (P120). Some participants also felt

Table 5: Analysis of differences between individual animation ratings, all participants ($N = 315$). Asterisk denotes significance.

	Engaging		Natural		Preferable	
	Z	p	Z	p	Z	p
C2−C1	4.43	1e−5*	3.51	.000457*	2.42	.016*
E−C1	10.47	<1e−5*	11.03	<1e−5*	9.65	<1e−5*
E−C2	8.91	<1e−5*	9.32	<1e−5*	8.63	<1e−5*

Table 6: Three-way animation ratings, all participants ($N = 315$) and VTuber participants only ($N = 56$).

	Most engaging		Most natural		Most preferred	
	All	VTubers	All	VTubers	All	VTubers
Mocap only	12%	13%	13%	9%	17%	18%
VMagicMirror	16%	18%	17%	16%	18%	20%
AlterEcho	**72%**	**70%**	**70%**	**75%**	**65%**	**63%**

that human-like body gestures would reduce the unique charm of an anime persona that tends to be "flat" and ideal: "*Too many subtle body movements would ruin the role-playing aspect of the VTuber. It made people think the world behind the screen is not so ideal*" (P85); "*[The AlterEcho animation] and [the VMagicMirror animation] feel like watching a real human, instead of an anime character, playing a game*" (P112); "*Although [the AlterEcho avatar] looks the most natural, it seems too similar to a real person when combined with the [streamer's] voice. [The VMagicMirror avatar] is just right: there are some cartoonish gestures, but they don't feel split with the voice*" (P203). Although one can change the frequency and variety of gestures in *AlterEcho* easily by adjusting the persona parameters, the "sweet spot" of the amount and types of avatar gestures for a typical VTuber audience remains to be explored by future studies.

Interestingly, some participants suggested going in the opposite direction of a flat avatar and asked for more frequent and pronounced facial expression changes: "*The facial expression in all videos should switch more aggressively*" (P3); "*The avatar expressions are too few... there should be more quirky expressions*" (P21); "*When a VTuber is playing a game such as Syobon Action [the game played by the VTuber], I expect the VTuber to have a plethora of reactions and facial expressions, so I think the facial expressions should be more enhanced and dramatic*" (P34). Some also preferred the facial expressions in the *VMagicMirror* animation over the ones in *AlterEcho*: "*I feel as if the facial expressions of [the VMagicMirror avatar] were more engaging, however [the AlterEcho avatar] has more dynamic movement*" (P270); "*While I enjoyed [the AlterEcho animation] the most overall, it seems like [the VMagicMirror animation] is more natural when it comes to the facial expressions*" (P268). The diversity in participants' preferences, some of which are contradictory, validate our persona modulation approach, which should be extended to allow precisely defining a desired persona.

Some participants noticed model clipping errors in the *AlterEcho* animation due to the lack of collision detection, finding it "*very distracting*" (P51), and that it even "*broke the immersion of it all*" (P255). Clipping is not a fundamental limitation of *AlterEcho*, and future implementations should remove it through algorithmic solutions for adapting baked animations to any new and possibly dynamic virtual environment geometry.

Some participants noticed the reuse of gestures: "*it is obvious [that] a lot of the animations [in the AlterEcho animation] are prescripted ones and the arm movement is off-putting*" (P286), "*though [the AlterEcho animation] was what I found the most natural and engaging, I did notice that the animations appeared to be canned after seeing the same one repeated so often*" (P298). One solution is to simply include more gestures in the animation library to make the gesture reuse less noticeable; another is to incorporate machine-learning techniques to generate gestures on-the-fly, such as speech-driven gesture synthesis [1].

4.2.2 Avatar Persona Modulation

The second part of the study evaluates the effectiveness of *AlterEcho*'s persona parameterization.

Conditions. We extracted a 40s segment from the source materials used in the first part of the survey, and generated two animations with *AlterEcho* using different extroversion parameters, i.e., $K_x = -1$, i.e., shiest, and $K_x = 1$, i.e., most outgoing. Note that we did not evaluate the energy parameter K_e, as it mainly affects

Table 7: Correct attribution % of personality adjectives, all participants ($N = 315$).

	Shy/introverted adjectives			
	shy	*introverted*	*reserved*	*insecure*
VTubers	96%	96%	77%	77%
Others	88%	95%	94%	69%
Total	90%	95%	91%	71%
	Outgoing/extroverted adjectives			
	confident	*extroverted*	*outgoing*	*positive*
VTubers	91%	95%	93%	93%
Others	91%	97%	96%	90%
Total	91%	96%	95%	90%

frequency of gestures, which is difficult to observe in a video of the length that an online survey permits.

Experimental procedure. The participant first watched a video that showed both animations side-by-side (similar to Fig. 3) in a random order. Then, the participant was provided the following twelve adjectives in a random order: "shy", "introverted", "reserved", "insecure", "confident", "extroverted", "outgoing", "positive", "likeable", "intimidating", "friendly", and "arrogant". They were asked to choose the avatar that best fits each adjective. The choices were "Neither", "Left", "Right", and "Both". The first four adjectives are correlated with a "shy/introverted" persona, and the next four with "outgoing/extroverted." The last four adjectives have a lesser correlation with the shy-outgoing avatar persona continuum, and they were included to make the attribution of the other eight adjectives more challenging. Finally, VTuber participants were asked if they would like to have high-level control over the personality of their avatars, and to provide optional free-form feedback.

Results and discussion. The percentage of correct attribution of the four shy/introverted and of the four outgoing/reserved adjectives is given in Table 7. An adjective is considered to be attributed correctly if it is attributed to the appropriate avatar, and not to the other avatar. Seven of the eight adjectives were attributed correctly 90% of the time or more, and a majority (59%) of participants attributed all eight adjectives correctly. On average, 7.2 of the 8 adjectives were attributed correctly, which indicates that the two personas were easily distinguishable. The highest error rate was measured for the attribute "insecure", whose placement on the introverted/extroverted scale might be ambiguous, as insecurity could also drive someone to being overly outgoing, not limited to being shy.

As we hypothesized, a majority (87.5%) of VTuber participants would like to be able to modulate the personality of their avatar, based on both better actualization of their designed VTuber persona and ergonomic reasons: "*I think both the VTuber and viewers would prefer to have an avatar that is more closely aligned with the VTuber persona and streamer's voice performance. As a VTuber, I would like to have a way to easily adjust my avatar to fit the setting*" (P34); "*Being able to adjust the avatar expressions would be interesting. The VTuber can use different parameters for different streaming scenarios, or even to achieve a comic effect like making the avatar unnecessarily outgoing when playing a horror game*" (P46); "*I think I would like to have this option in the case I have a hard time conveying a behaviour. For example, I might want to look more outgoing even though I'm stressed because it's only my third stream. Just as you would use a voice changer*" (P265).

Finally, a recurring feedback was the inconsistency between the timid gestures that the shy avatar makes and the streamer's energetic voice: "*The soul of the avatar needs to match the avatar; even if I personally prefer [the shy avatar], [the confident avatar] seems more fitting with the voice*" (P101); "*Because only the pose of the avatar was changed and not the tone of voice, [the shy avatar] felt significantly more incongruous, which is why I think [watching the confident avatar] is a significantly better viewing experience*" (P89). Future work should investigate the influence of the streamer's voice on how the VTuber persona is perceived by the audience.

5 Conclusion, Limitations, and Future Work

We have presented *AlterEcho*, a VTubing avatar animation system designed to be automatic, to give high-level control over the avatar persona, to have modest equipment constraints, and not only to tolerate, but also to take full advantage of differences between the real and virtual world. *AlterEcho* was evaluated in a large controlled study with participants with substantial VTuber and VTubing viewing experience; the *AlterEcho* animations surpassed animations produced automatically from similar motion capture input, and even animations where motion capture was enhanced manually, offline, with a state-of-the-art keyboard-based VTubing system.

We have begun discussing the limitations of *AlterEcho* in Sect. 4.2 where we summarized the shortcomings reported by our participants. Furthermore, our approach of loosening the coupling between avatar and streamer requires finalizing the animation with a buffering delay (see Sect. 3.4.2) that allows the system to trigger and integrate a complex set of possibly conflicting candidate gestures. The delay has to be at least as long as the gesture with the longest duration; we used a conservative 15s delay. This delay implies that the streamer does not see their avatar react in real-time, and it also precludes low-latency live interactions between the streamer and the audience, which is a fundamental limitation of our approach. Future work could investigate the feasibility of a smaller delay to decrease streaming latency. While a few seconds of additional latency is likely acceptable as existing live streaming exhibits a latency of several tens of seconds [49], this hypothesis should be tested once *AlterEcho* is deployed to be used by VTubers; such a deployment could also further validate *AlterEcho* through viewers comparing VTubing streams animated using *AlterEcho* to those animated using other systems.

In addition, our avatar persona modulation is currently limited to nonverbal communication; future work could investigate modifying verbal content and delivery of the avatar. Future work should also examine the granularity of avatar persona adaptation, which so far has only been tested at the two endpoints of the shy/introverted to outgoing/extroverted continuum, and the avatar persona could also be modulated along additional dimensions. Finally, our implementation of *AlterEcho* assumes the VTuber has a 3D avatar; however, many VTubers, especially independent VTubers, still rely on animating a 2D avatar using systems like *Live2D* [32, 57] to avoid the higher cost of modeling and rigging a 3D avatar. Future work could look into extending *AlterEcho* with support for 2D avatars.

Our work opens the door to personalizing the avatar persona for individual viewers, either automatically, based on inferences driven by viewer actions, or under the viewer's control. For pedagogical agents, personalization to individual learners is one of the biggest potential advantages over conventional in-class education [30], whereas in VTubing, the role of avatar personalization is less clear: VTubers might be reticent to hand over avatar control to viewers, and viewers might not want to break the illusion that the fictional persona is "real" by manipulating the avatar. Ultimately, future work could explore increasing the avatar's autonomy from the streamer, in the quest of rich artistic expression, enabled by complex, evolving, contradictory, rebellious, or even completely independent avatars.

Acknowledgments

We thank VTubers Ying and Probe, illustrator Chiongyi, and all survey participants for their contribution to our study. We also thank developers Hinzka for their work in applying blendshapes to VTubing avatars and Megumi Baxter for their work on VMagicMirror; both were great inspirations to *AlterEcho*.

REFERENCES

[1] S. Alexanderson, G. E. Henter, T. Kucherenko, and J. Beskow. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum*, 2020. doi: 10.1111/cgf.13946

[2] AMBI. VTuberって何？4年後には市場規模500億円超の試算も [What are VTubers? Some estimate that the market will exceed 50 billion yen in 4 years]. https://www.huffingtonpost.jp/ambi/vtuber-shijyo_a_23472458/, Nov 2020.

[3] Apple Inc. ARFaceAnchor. https://developer.apple.com/documentation/arkit/arfaceanchor.

[4] A. Aristidou, J. Lasenby, Y. Chrysanthou, and A. Shamir. Inverse Kinematics Techniques in Computer Graphics: A Survey. *Computer Graphics Forum*, 37(6):35–58, Nov 2017. doi: 10.1111/cgf.13310

[5] T. aubio Team. aubio, a library for audio labelling. https://aubio.org/.

[6] 东西文娱. 乐华推首个虚拟女团A-SOUL：偶像经纪公司布局虚拟偶像的时机到了吗？ [Yuehua Entertainment pushes its first virtual idol group A-SOUL: Is it time for idol agencies to deploy virtual idols?]. https://www.36kr.com/p/994429983111300, Dec 2020.

[7] カバー株式会社. 「ホロライブ」3Dリアルタイム配信アプリのモーショントラッキングをアップデート！ [Hololive's 3D real-time streaming app updated with motion tracking!]. https://prtimes.jp/main/html/rd/p/000000052.000030268.html, Apr 2019.

[8] 合同会社ラペットテクノロジーズ. Luppet. https://luppet.appspot.com/.

[9] 獏星. VMagicMirror. https://malaybaku.github.io/VMagicMirror.

[10] Bilibili. Bilibili. https://www.bilibili.com/.

[11] A. N. Bosmia, C. J. Griessenauer, and R. S. Tubbs. Yubitsume: ritualistic self-amputation of proximal digits among the Yakuza. *Journal of injury and violence research*, 6(2):54, 2014.

[12] L. Bredikhina. Designing identity in VTuber Era. *ConVRgence (VRIC) Virtual Reality International Conference Proceedings*, Apr. 2020. doi: 10.20870/IJVR.2020…3316

[13] A. Cafaro, H. H. Vilhjálmsson, and T. Bickmore. First Impressions in Human–Agent Virtual Encounters. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(4):1–40, 2016.

[14] Changsha Ranxing Information Technology Co., Ltd. 问卷星 [Wen Juan Xing]. https://www.wjx.cn/.

[15] J. Chen. The Vtuber takeover of 2020. https://www.polygon.com/2020/11/30/21726800/hololive-vtuber-projekt-melody-kizuna-ai-calliope-mori-vshojo-youtube-earnings, Nov 2020.

[16] E. Chuang and C. Bregler. Performance driven facial animation using blendshape interpolation. *Computer Science Technical Report, Stanford University*, 2(2):3, 2002.

[17] C. Corp. Inugami Korone. https://en.hololive.tv/portfolio/items/inugami-korone/.

[18] FaceRig. FaceRig. https://facerig.com/.

[19] P. W. Galbraith. Moe: Exploring virtual potential in post-millennial Japan. *Electronic Journal of Contemporary Japanese Studies*, 9(3), 2009.

[20] @gawrgura. “Thank you. I am an overwhelmed, but very happy shark. Thank you to each of my hard working senpai, and all of you. From my big shark heart, thank you.”. https://twitter.com/gawrgura/status/1319328921305964544, Oct 2020.

[21] S. Ghosh, A. Dasgupta, and A. Swetapadma. A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification. In *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 24–28, 2019. doi: 10.1109/ISS1.2019.8908018

[22] D. C. Giles. *The Popularity and Appeal of YouTubers: ‘Authenticity’ and ‘Ordinariness’*, p. 131–153. Emerald Publishing Limited, 2018. doi: 10.1108/978-1-78743-708-120181011

[23] Google. Mediapipe Face Mesh. https://google.github.io/mediapipe/solutions/face_mesh.

[24] T. Higa. バーチャルYouTuberの人数が1万3000人突破、人気1位はファン数286万人のキズナアイ [The number of virtual YouTubers has surpassed 13,000, and the most popular is Kizuna Ai with 2.86 million fans]. https://jp.techcrunch.com/2020/11/09/userlocal-virtual-youtuber/, Nov 2020.

[25] M. Hu, M. Zhang, and Y. Wang. Why do audiences choose to keep watching on live video streaming platforms? An explanation of dual identification framework. *Computers in Human Behavior*, 75:594–606, 2017. doi: 10.1016/j.chb.2017.06.006

[26] Y. Hu, S. Xie, Y. Xu, and J. Sun. Dynamic VR live streaming over MMT. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–4, 2017. doi: 10.1109/BMSB.2017.7986127

[27] A. L. Hubbard, S. M. Wilson, D. E. Callan, and M. Dapretto. Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, 30(3):1028–1037, 2009. doi: 10.1002/hbm.20565

[28] Japan National Tourism Organization. JNTO to Launch ‘Come to Japan’ Campaign With Kizuna AI, the World’s First Virtual YouTuber. https://www.prnewswire.com/news-releases/jnto-to-launch-come-to-japan-campaign-with-kizuna-ai-the-worlds-first-virtual-youtuber-300608037.html, Mar 2018.

[29] Kizuna AI Inc. A.I. Channel. https://www.youtube.com/channel/UC4YaOt1yT-ZeyB0OmxHgolA.

[30] N. C. Krämer and G. Bente. Personalizing e-learning. The social effects of pedagogical agents. *Educational Psychology Review*, 22(1):71–87, 2010.

[31] S. Li and W. Deng. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, pp. 1–1, 2020. doi: 10.1109/TAFFC.2020.2981446

[32] Live2D Inc. Live2D Cubism. https://www.live2d.com/.

[33] Z. Lu, C. Shen, J. Li, H. Shen, and D. Wigdor. More Kawaii than a Real-Person Live Streamer: Understanding How the Otaku Community Engages with and Perceives Virtual YouTubers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445660

[34] Z. Lu, H. Xia, S. Heo, and D. Wigdor. You Watch, You Give, and You Engage: A Study of Live Streaming Practices in China. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174040

[35] Microsoft. Speech to Text. https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/.

[36] M. Neff, Y. Wang, R. Abbott, and M. Walker. Evaluating the effect of gesture and language on personality perception in conversational agents. In *International Conference on Intelligent Virtual Agents*, pp. 222–235. Springer, 2010.

[37] NGA. Vtuber综合讨论区 [Vtuber General Discussion Forum]. https://bbs.nga.cn/thread.php?fid=-60204499.

[38] OBS Project. Open Broadcaster Software. https://www.obsproject.com/.

[39] OpenCV. Face Recognition with OpenCV. https://docs.opencv.org/3.4/da/d60/tutorial_face_main.html.

[40] N. Otmazgin. Creative Masses: Amateur Participation and Corporate Success in Japan’s Media Industries. *Creative Context Creative Economy*, p. 65–82, Apr 2020. doi: 10.1007/978-981-15-3056-2_5

[41] A. J. Pellicone and J. Ahn. The Game of Performing Play: Understanding Streaming as Cultural Production. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI ’17, p. 4863–4874. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3025453.3025854

[42] H. Pham. PyAudio: PortAudio v19 Python Bindings. https://people.csail.mit.edu/hubert/pyaudio/.

[43] H. Qi. Virtual reality concert draws large crowds in Shanghai. http://www.chinadaily.com.cn/a/201907/22/WS5d357fdea310d830564005bc_2.html, July 2019.

[44] Qualtrics. Qualtrics XM. https://www.qualtrics.com/.

[45] Reddit Inc. r/VirtualYouTubers. https://www.reddit.com/r/VirtualYoutubers/.

[46] R. K. P. Rohith Raj S, Pratiba D. Facial Expression Recognition using Facial Landmarks: A novel approach. *Advances in Science, Technology and Engineering Systems Journal*, 5(5):24–28, 2020. doi: 10.25046/aj050504

[47] sh_akira. Virtual Motion Capture. https://vmc.info/.

[48] A. Shirai. REALITY: Broadcast your virtual beings from everywhere. In *ACM SIGGRAPH 2019 Appy Hour*, pp. 1–2. 2019.

[49] Y. Shuai and T. Herfet. Towards reduced latency in adaptive live streaming. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pp. 1–4. IEEE, 2018.

[50] A. P. Singh, R. Nath, and S. Kumar. A Survey: Speech Recognition Approaches and Techniques. In *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pp. 1–4, 2018. doi: 10.1109/UPCON.2018.8596954

[51] C. Strassberger and R. Sikkema. Democratising Mocap: Real-Time Full-Performance Motion Capture with an IPhone X, Xsens, and Maya. In *ACM SIGGRAPH 2018 Real-Time Live!*, SIGGRAPH '18. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3229227.3229236

[52] R. I. Thackray and R. M. Touchstone. Recovery of motor performance following startle. *PubMed*, Feb. 1970. doi: 10.2466/pms.1970.30.1.279

[53] A. Tinwella and R. J. S. Sloan. Children's perception of uncanny human-like virtual characters. *Computers in Human Behavior*, 36:286–296, July 2014. doi: 10.1016/j.chb.2014.03.073

[54] Twitch. Twitch. https://www.twitch.tv/.

[55] B. Wang, X. Zhang, G. Wang, H. Zheng, and B. Y. Zhao. Anatomy of a personalized livestreaming system. In *Proceedings of the 2016 Internet Measurement Conference*, pp. 485–498, 2016.

[56] Y. C. Wu and S. Coulson. How iconic gestures enhance communication: An ERP study. *Brain and Language*, 101(3):234–245, 2007. Gesture, Brain, and Language. doi: 10.1016/j.bandl.2006.12.003

[57] G. Çakır. How to become a VTuber. https://dotesports.com/streaming/news/how-to-become-a-vtuber, Jan 2021.