The Importance of Multimodal Emotion Conditioning and Affect Consistency for Embodied Conversational Agents

Che-Jui Chang
Rutgers University
New Jersey, USA
chejui.chang@rutgers.edu

Samuel S. Sohn
Rutgers University
New Jersey, USA
samuel.sohn@rutgers.edu

Sen Zhang
Rutgers University
New Jersey, USA
sen.z@rutgers.edu

Rajath Jayashankar
Rutgers University
New Jersey, USA
rajath.jay@rutgers.edu

Muhammad Usman
King Fahd University of Petroleum & Minerals
Dhahran, Saudi Arabia
muhammad.usman@kfupm.edu.sa

Mubbasir Kapadia
Rutgers University
New Jersey, USA
mubbasir.kapadia@rutgers.edu

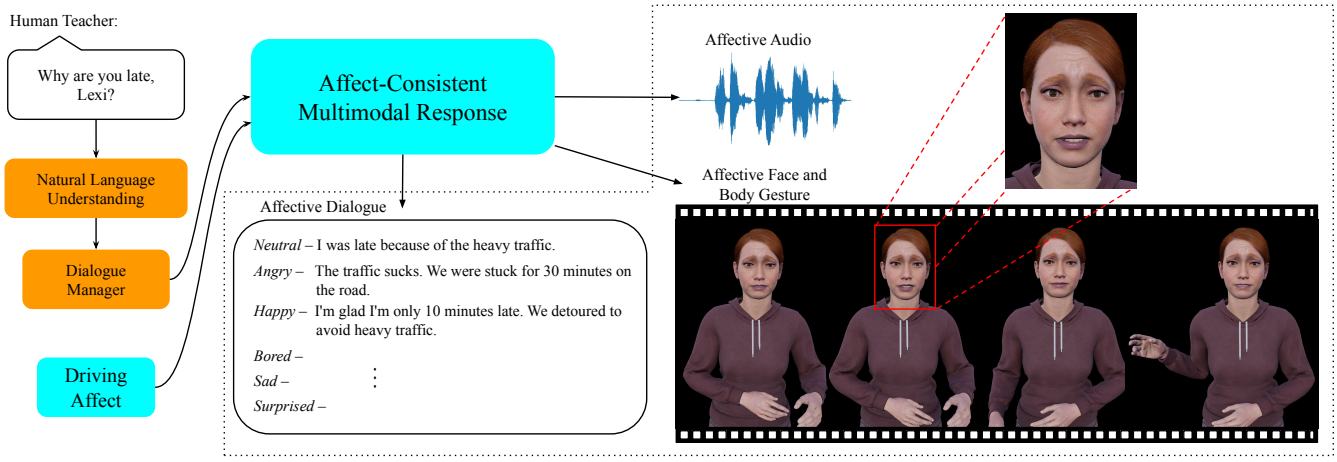

Figure 1: We propose, ACTOR, a conceptual framework with affect-consistent multimodal behaviors for autonomous embodied conversational agents that enhances human's perception of affects. Our framework consists of four affective modalities, dialogue (left, below center), voice (right, above center), face, and body gesture (lower right), generated from the same driving affect (lower left).

ABSTRACT

Previous studies regarding the perception of emotions for embodied virtual agents have shown the effectiveness of using virtual characters in conveying emotions through interactions with humans. However, creating an autonomous embodied conversational agent with expressive behaviors presents two major challenges. The first challenge is the difficulty of synthesizing the conversational behaviors for each modality that are as expressive as real human behaviors. The second challenge is that the affects are modeled independently, which makes it difficult to generate multimodal responses with consistent emotions across all modalities. In this work, we propose a conceptual framework, ACTOR (**A**ffect-**C**onsistent mul**T**imodal behavi**OR** generation), that aims to increase the perception of affects by generating multimodal behaviors conditioned on a consistent driving affect. We have conducted a user study with 199 participants to assess how the average person judges the affects perceived from multimodal behaviors that are consistent and inconsistent with respect to a driving affect. The result shows that among all model conditions, our affect-consistent framework receives the highest Likert scores for the perception of driving affects. Our statistical analysis suggests that making a modality affect-inconsistent significantly decreases the perception of driving affects. We also observe that multimodal behaviors conditioned on consistent affects are more expressive compared to behaviors with inconsistent affects. Therefore, we conclude that multimodal emotion conditioning and affect consistency are vital to enhancing the perception of affects for embodied conversational agents.

IUI '23, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0106-1/23/03…$15.00
https://doi.org/10.1145/3581641.3584045

CCS CONCEPTS

• **Computing methodologies** → **Intelligent agents**; *Virtual reality*; • **Human-centered computing** → **User studies**.

KEYWORDS

embodied conversational agents, multimodal behavior generation, emotion conditioning, affect consistency

ACM Reference Format:
Che-Jui Chang, Samuel S. Sohn, Sen Zhang, Rajath Jayashankar, Muhammad Usman, and Mubbasir Kapadia. 2023. The Importance of Multimodal Emotion Conditioning and Affect Consistency for Embodied Conversational Agents. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581641.3584045

1 INTRODUCTION

Embodied virtual agents have garnered increasing attention in the field of graphics and virtual reality as they facilitate the building of realistic, immersive environments and low-risk virtual training platforms [4, 17]. Several studies [15, 22, 43] have shown the efficacy of emotion contagion with the use of expressive virtual agents during multimodal interactions with humans. However, it remains a challenge to generate affective multimodal responses for autonomous embodied conversational agents (ECAs) that are as expressive as humans. As it stands, creating an embodied conversational agent with expressive multimodal responses requires the effort of putting together all the modalities and synchronizing the multimodal behaviors, but the fact that the emotions are modeled separately could be another obstacle for building such an expressive framework. In this work, we propose a conceptual framework, ACTOR (**A**ffect-**C**onsistent mul**T**imodal behavi**OR** generation), with multimodal emotion conditioning and affect consistency that addresses the aforementioned issues and increases human's perception of affects.

The concept of emotion conditioning refers to the capability of taking an emotion as a conditional input and generating the behavior for a modality that matches the condition. For example, an emotion-conditioned face modality may use the audio and conditioned emotion as inputs to generate an affective facial animation. On the contrary, the non-conditioned modalities, which are included in most existing embodied conversational frameworks [18, 45, 56], only accept one input mode and output another mode without consideration of emotion. The notion of affect consistency refers to an integrated ECA framework being able to generate multimodal emotional behaviors with the same affect. Terminologically, we refer to the conditioned emotion for each modality as the driving affect in the ACTOR framework, as it is used to drive the affective behaviors.

Our affect-consistent framework consists of four modalities: dialogue, voice, face, and body. The conversational behaviors for each modality are generated given the same driving affect, as illustrated in Figure 1. Practically, the behaviors are generated using stylistic parameters that have been linked to each driving affect, as described in Section 3.1. We describe the preliminaries of our main user study, including the design of the experiments, creation of stimulus, and the comparison models, in Section 4. We show the confusion matrix of the perception scores for each comparison model to evaluate the efficacy of multimodal emotion conditioning and affect consistency on affect perception and conduct ANOVA tests to report the statistical significance in affect perception under 4 model conditions, 6 driving affects, and 6 perceived affects, in Section 5. We summarize the experimental results and discuss our key findings in Section 6.

This paper makes the following contributions. First, we propose a conceptual multimodal framework, ACTOR, that resolves the two aforementioned challenges for building autonomous ECAs, including the difficulty of generating the expressive behaviors for the conveyance of emotions and the issue of stitching together the modalities in which the affects are modeled separately. Such a framework can be applied to create intelligent virtual agents in video games or deployed in human-computer interaction interfaces to increase the immersive experiences for virtual training and assistance. Second, we discover from our user study that affect consistency maximizes the perception of the driving affect and that an inconsistent affect in just one modality can decrease the same affect perception. In fact, when a modality is not emotion conditioned, the behaviors tend to be less expressive, which then dilutes the perception of the driving affect. These findings evidence the importance of emotion conditioning and affect consistency for each modality. Finally, an additional statistical analysis reveals several more nuanced findings. (1) The voice and face modalities contribute to affect perception more than the body modality, as the removal of a consistent affect in the body modality does not necessarily disrupt the perception of the same affect. (2) The correlations in perception scores between the affects can be explained using their valence and arousal values. These findings are important for building an ECA framework with multimodal emotional responses in which the affects and modalities play a vital role. The video demonstration for the affect-consistent multimodal behaviors can be found in our supplementary materials.

2 RELATED WORK

Our proposed framework and user study for affect perception are related to the coordination of the agent's perception and responses, the modeling of synchronous modalities, and the evaluation of affect-driven or personality driven agents. In this section, we review the literature from the following perspectives: Multimodal Conversational Agents and Affect-Driven Avatars.

2.1 Multimodal Conversational Agents

2.1.1 Multimodal Communication. Multimodal communication is a natural form between human interlocutors where audio, facial expressions, eye gazes, head movements, hand gestures, and body gestures are used to provide vivid conversational behaviors. Likewise, the same communication strategy has been proven to be effective in increasing the realism of embodied virtual agents according to previous studies [26, 47]. Social skills including behavior matching [41], style matching [29], and emotion-awareness [56] have been investigated and developed in virtual conversational agents. In addition, research [8] has shown that synchronization of the modalities, specifically speech and gestures, for multimodal ECAs is the key to improving realism and human likeness. We leverage

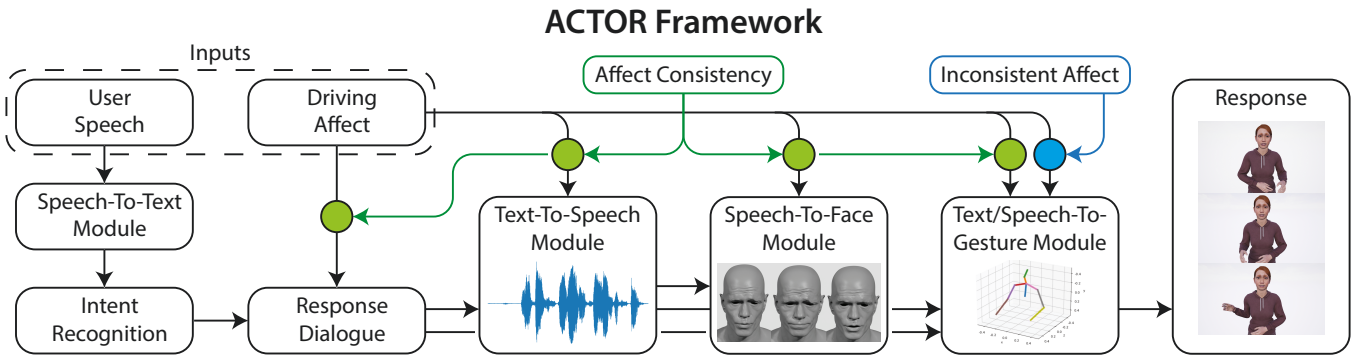

Figure 2: Our conceptual framework, ACTOR, with multimodal emotion conditioning and affect consistency. First the intent is recognized from the user's speech input and the corresponding sentence response is used to generate the voice for the virtual agent. The synthesized audio is then used to generate the face animation. Last, the body gestures are generated using the text and audio as inputs. The same driving affect is conditioned for all modalities to achieve affect consistency. To evaluate the influence of emotion conditioning and affect consistency on the perception of affects in the user study, we build affect inconsistent models for comparisons, where a single modality has an inconsistent affect.

the findings from these prior works by applying data-driven approaches for multimodal behaviors and synthesizing the behaviors with consistent affects.

2.1.2 Coordination of Agent's Perception and Response. Authoring an interactive ECA requires an integrated framework that coordinates the agent's perception and conversational responses. The interactive behavior tree was introduced in the literature [32, 34, 36, 54] for such a framework. PICA [23] models the active and reactive agent behaviors for a conversational agent, with the ability to initiate conversations, parse user intents and handles responses. Typically, a Natural Language Understanding (NLU) module is the core of the dialogue management that maps user's semantic inputs to the predefined intent. According to the intent and dialogue state, a pre-authored multimodal response is played. For instance, [37] generates dialogue responses by a rule-based NLU module and incorporates gestures as an additional modality. [57] creates and modifies the personality-driven multimodal behaviors offline and combines the modalities to render a response to an intent. The work from [56] takes multimodal perception inputs, processes the user emotion and dialogue intent, and generates the pre-authored animation response with emotion-awareness. Our work follows the same pipeline for intent parsing, but applies emotion-conditioned methods for multimodal response generation.

2.1.3 Synchronization of Modalities. As the modalities are generated through separate channels, synchronization becomes a key issue to human perception of realism, especially between speech and gestures [8]. [24] investigates the speech-gesture correlation and aims to understand the predictability of gesture from speech. Their follow-up work [25] leverages the speech-gesture relationship for gesture synchronization by finding the best gesture in the dataset through feature matching. Another branch of speech-gesture synchronization studies [2, 7, 12, 27, 28, 62] focuses on using a parameterized network to learn mappings between the two modalities. For example, Speech2Gesture [27] detects the hand gestures from video and builds a convolutional network for gesture generation. [2] uses a probabilistic flow-based model for likelihood maximization for co-speech gesture generation. On the other hand, several research groups [6, 7, 49, 62] have included semantic features to infer gestures. [62] encodes audio, text, and speaker identity for the generation of gestures. [6] includes an additional seed pose as input for autoregressive affective gesture generation. For more works regarding speech-driven gesture, we direct readers to these review papers [40, 60].

2.2 Affect-Driven Avatars

2.2.1 Affective Avatars. The accurate modeling of affects increases the believability of a conversational agent. Prior work [55] builds a computational model for affects with the psychologically-guided interplay of affective components including personality, motivation, emotion, and mood. [39] conducts a user study to create psychologically-plausible facial expressions. The increasing modeling capability of affects enables us to create ECAs capable of showing empathy and expressing emotional behaviors in terms of dialogues [9], faces [14, 39], and gestures [6]. Previous studies include emotionally-aware avatars with affect matching [56] and affect regulation [61], personality-based emotional characters [52], and an affective virtual student [17] to support teacher training.

Our framework does not incorporate affective perception but focuses on the synthesis of multimodal affect-consistent behaviors with our defined emotions.

2.2.2 Personality-Driven Avatars. Building an ECA with personality-driven behaviors has become a major focus in recent years, with evidence and findings from prior research [10, 19, 31, 55] to support the benefit of personality modeling. Typically, the OCEAN personality [10, 19, 42] is used in the literature for analysis and generation of character movements. [57] does personality-specific adjustments for all conversational modalities. The voices and facial expressions for OCEAN personalities are built on top of an affect layer. Their body motions are modified according to Laban

Table 1: The mapping of the affects and the stylistic parameters in the voice [30], face [46], and body [6] modality. For example, neutral audio would increase the speaking rate by 40%. For the face modality, the source shot represents the style of the facial expression. The smoothing parameter determines the smoothness of the face animation can be along the time axis. The strength parameter determines the degree of exaggeration. For the body modality, the seed pose ID refers to the pre-pose used for training in [6]. It determines the style of the predicted pose. We randomly select one of the listed IDs in our framework.

Modality	Affective Parameter	Neutral	Angry	Happy	Bored	Sad	Surprised
Voice	Pitch	-	20%	100%	-20%	20%	80%
	Pitch Range	-	100%	100%	-20	20%	40%
	Rate	40%	50%	10%	-30%	-10%	20%
	Breathiness	-	-	-100%	-	100%	-30%
	Glottal Tension	-	80%	60%	-	-20%	50%
Face	Source Shot	p3_neutral	g1a	g2b	g8b	g8c	g4b
	Upper, Lower Smoothing	0.011, 0.004	0.012, 0.000	0.012, 0.005	0.025, 0.004	0.056, 0.009	0.014, 0.002
	Upper, Lower Strength	0.600, 1.218	0.798, 1.608	1.874, 1.632	0.600, 1.400	1.594, 1.126	1.696, 1.330
Body	Seed Pose IDs	18, 30	5, 10	8, 27	23, 130	123, 182	142, 181

Movement Analysis. Though our work focuses primarily on affect-driven behaviors, we apply similar adjustments to the modalities for emotion conditioning.

3 CONCEPTUAL FRAMEWORK

In general, the process of generating multimodal responses involves the following sequential steps: (1) dialogue generation, (2) text-to-speech, (3) audio-driven (or text-driven) facial animation, and (4) audio-driven body animation. However, such an ECA framework that generates the behaviors without emotion conditioning could lead to *emotion dilution* –a multimodal response that is supposed to be emotional and expressive, but turns out to be perceived as more neutral. Specifically, the reason comes from the fact that each step of the sequential executions is trained separately on different domains and then concatenated in a framework. For example, a non-conditioning system, with a text-to-speech module followed by an audio-driven facial animation module, may have the two modules trained on different datasets. The quality of the face modality can be undermined when synthetic audio generated from a different domain is used as its input during inference. Additionally, the degree of emotional expressiveness can often be diluted in each sequential step due to the sequential execution of modules without emotion conditioning. For example, the same non-conditioning system can dilute the expressiveness in the face, even though an affective text input is given initially. This is because all of these data-driven modules have to decode the affect from the input and encode it into the output behavior. The learning of affects is not guaranteed, so the affect in the output modality can become diluted. Therefore, all modalities in our proposed framework are emotion-conditioned.

It is widely accepted that the perception of affects is the most heightened when all modalities share the same affect. Prior research [15, 22] regarding the effect of different modalities of a virtual character on the perception of emotions always assumed that combined modalities shared the same emotion in their study design. However, achieving affect consistency for ECAs is challenging because all modalities can model affects in different ways. The affect that is conditioned in one modality might not be used as the condition in another. Moreover, some modules do not have interpretable variables for emotion conditioning. For example, the work from [35] is able to generate expressive face animations from audio input, but the variables used to control the output expressions are not interpretable. Therefore, we propose the conceptual framework, ACTOR, with all modalities conditioned on a consistent affect, as illustrated in Figure 2. Like other ECA frameworks, our framework follows a similar sequential execution but uses the same driving affect as the condition input for generating multimodal behaviors. We detail the implementation of our conceptual framework in the following subsections. To show the importance of emotion conditioning and affect consistency in an ECA framework, we also conduct a user study (Sec. 4) to compare our affect consistent model condition with the affect inconsistent models with one modality having a different driving affect.

3.1 Implementation

We realize the conceptual framework based on existing data-driven methods and create a virtual student, Lexi, based on the Metahuman character [21]. We use pre-scripted affective text responses and apply text-to-speech synthesis, audio-to-face generation, and gesture generation from text and audio in our implementation. In order to create a shared affect space for multimodal emotion conditioning, we utilize a mapping from the emotions to the stylistic parameters for each modality. The parameters are either not annotated (e.g. latent codes learned in an unsupervised fashion) or interpretable but not directly related to emotions. A summary of the mappings is shown in Table 1. The modifications of the voice parameters are based on psychological studies [53] on features for emotional voices while the adjustments of the latent parameters for faces and gestures are validated in our preliminary user study (Sec. 3.2). Based on the framework, we design two scenarios: the late scenario and the homework submission scenario. The perception of the character in our implementation is done by a speech-to-text engine followed by an intent extraction. Depending on the intent

Table 2: An example of our designed dialogue with intent, sample question, and responses for all affects.

Intent	Cover-up 1
Question	"Then why are you the only one in class who's late today?"
Neutral	"I went to a doctor earlier and then came to school."
Angry	"Well, I had to see a doctor first. That was *not* my idea. My mom asked me to."
Happy	"Well, I came here after a doctor's appointment. My headache was relieved."
Bored	"I had to see a doctor first. Then come here."
Sad	"I had to go to see a doctor. Please don't blame me for that. I'm really sorry."
Surprised	"Oh! Um, my mom asked me to see a doctor first."

of the user, the scenarios guide users through different dialogue states and ultimately the dialogues are directed to different endings. We refer readers to our appendices for more details regarding the framework implementation and our scenario flowcharts.

The following 6 affects are used to condition the modalities: *neutrality, anger, happiness, boredom, sadness, and surprise*. We do not use all emotions from the universal model [20] because contemptuous, disgusted, and fearful voices are not as distinguishable as others for most widely used Text-To-Speech engines, such as IBM TTS [30].

3.1.1 Dialogue Modality. Research from prior works [16, 64] is intended for generating dialogue responses with the driving affect as the condition. However, we could not adopt these prior works for our affective dialogue generation because the generated dialogues do not fit in our two scenarios and guide users to specific branches and endings. To this end, we designed our own dialogues. Table 2 shows an example of the dialogue in the late scenario. For each dialogue state, we created 6 response sentences, which correspond to the 6 driving affects. These affective responses were then used as inputs for the generation of behaviors in other modalities.

3.1.2 Voice Modality. Given an input text and a driving affect, ACTOR synthesizes the speech for the voice modality. Although there exist emotion-controllable text-to-speech synthesis methods [38], these previous studies are not applicable because of either their poor quality, domain differences, or language disparity [11, 38]. Instead, we use the IBM Watson Text-To-Speech Engine [30] for our voice modality. The provided controllable features are *pitch, pitch range, rate, breathiness, and glottal tension*. *Pitch* refers to the frequency of the voice. *Pitch range* specifies the variation of pitch during speech. *Rate* refers to the talking speed. *Breathiness* determines the amount of air produced in the sound. *Glottal tension* decides how hard the voice is. A voice with higher breathiness and lower glottal tension sounds calm and soft. We carefully set the feature values for all affects according to the review paper [53]. The mapping is shown in Table 1. For detailed information about the features, we refer readers to these works [30, 57].

3.1.3 Facial Animation Modality. The face modality takes the synthesized audio and the driving affect as inputs, and then outputs a facial animation with synchronized lip movements and a matching affect. The previous work proposed by [35] builds an end-to-end model for the task using unsupervised latent codes weakly associated with affects. Recently, the work from [13] disentangles audio content and emotion and entangles the content with the driving affect for expressive emotion-conditioned animation synthesis. Although these works synthesize high-quality emotional facial animations, their facial motions cannot be retargeted to the Metahuman character's parametric blendshape [21]. As a result, we leverage Omniverse Audio2Face [46], where the parametric face model is driven by input audio and stylistic parameters, including the source shot, smoothing, and strength. The source shot controls the style of the facial expression. The smoothing parameter determines the smoothness of the face animation can be along the time axis. The strength parameter determines the degree of exaggeration. We create the mappings and link the affects with the parameters for emotion conditioning, as shown in Table 1. The generated animations are then retargeted to the Metahuman character.

3.1.4 Body Gesture Modality. As discussed in Section 2.1.3, body gestures can be synthesized from audio and text inputs by an end-to-end network. For instance, Text2Gestures [7] is designed to generate affective, natural-looking gestures from textual semantics, while Speech2AffectiveGestures [6] leverages multimodal inputs, including text, audio, speaker identity, and a seed pose, to synthesize the affective gestures. Notably, the affective constraint in the latter work enables the output gestures to share the same affect with the seed pose. The seed pose is the pose sequence in the first few frames of the training segment. During inference, it can be used to generate gestures for different talking styles. Therefore, we build our gesture modality on top of Speech2AffectiveGestures, where the dialogue and synthetic speech are passed to the pretrained model for body gesture generation. The Speech2AffectiveGestures model was trained on TED Gesture Dataset [63], where only the upper bodies (10 joints) were used. The prediction of 10 joint positions are converted to the rotational angles relative to their parent joints and then retargeted to their corresponding joints in our character. We select the seed pose ID as the affective parameter for the body modality. The selection of the affective parameter is shown in Table 1.

3.2 Preliminary User Study

We conducted a preliminary user study for the validation of our choice of affective parameters for both face and gesture modalities.

Table 3: Summary of the pairwise comparison for affects. Each row represents whether an affective behavior can be identified when paired with all other affective behaviors. The *average match* is defined as the percentage of *match* cases plus half *equal* cases.

	Affect	Match	Equal	Mismatch	Avg. Match
Face	Neutral	0.55	0.35	0.10	0.73
	Angry	0.64	0.29	0.07	**0.79**
	Happy	0.82	0.08	0.10	**0.86**
	Bored	0.63	0.27	0.10	**0.77**
	Sad	0.54	0.34	0.11	0.71
	Surprised	0.59	0.31	0.10	0.75
Body	Neutral	0.52	0.38	0.10	0.71
	Angry	0.55	0.43	0.02	**0.77**
	Happy	0.53	0.35	0.12	0.71
	Bored	0.73	0.18	0.08	**0.83**
	Sad	0.52	0.25	0.23	0.64
	Surprised	0.40	0.47	0.13	0.63

We adopted the pairwise comparison method in the study where a random pair of affective behaviors were presented at the same time. The participants were then asked which better matches the description. The two affects of the selected behaviors were queried. For example, a pair of angry and happy facial animations was presented and two questions, which animation is happy and which is angry, were asked. The users could select either left, right, or equal as their response. The survey includes two parts, one for facial animations and the other for body movements. For each affect, we rendered 3 stimuli with each lasting roughly 3 seconds long. The behaviors only contained one modality and the rendered videos were without audio. The selection of the pairs and the order of the queried affects were randomized. We recruited 18 participants from the university, and each participant completed 30 survey questions: 15 videos for the face modality and 15 for the body modality.

We report the percentages of match, mismatch, and equal for the pairwise comparison result for the affects. Each affect is compared with all other affects and the summary of the result is shown in Table 3. We can see from the table that happy faces have the highest match percentage among all affects. This means that when a happy facial animation is presented with other emotional facial animations, more than 80 percent of our raters can accurately recognize it. Sad and neutral faces receive a higher equal percentage because users can sometimes be uncertain when they are paired with other emotional faces. Nonetheless, all affective facial behaviors receive the average match (match + 0.5 * equal) larger than 71 percent. On average, our affective facial parameters lead to 77% average match, far beyond random guess (50% average match). Regarding body gestures, our selection of the affective parameters leads to a slightly lower average match than the face modality, with 71.5% average match for all affects. Boredom is the most distinguishable among all six affects, with 73% match and 83% average match. Sadness and surprise, however, receive 64% and 63% average match respectively.

4 MAIN USER STUDY

We conducted the main user study for our framework with all 4 modalities included: dialogue, voice, face, and body. We presented our stimulus with all modalities in the user study but changed the driving affect in each modality to see how the configuration of affects influences the affect perception. We compared our affect-consistent model (AC) where all modalities share the same driving affect with the three model conditions that have an inconsistent voice (IV), inconsistent face (IF), or inconsistent body (IB) modality. Table 4 lists the names of all our comparison models as well as their affect settings for all modalities. We denote the driving affect as Affect X and the inconsistent driving affect as Affect Y. The affect of the dialogue modality remains unchanged because the sentences are pre-scripted and all other modalities are dependent on the dialogue's content.

Table 4: The affect setting for the four model conditions in our user study. AC refers to our affect consistent model. IV, IF, and IB refer to the models with inconsistent voice, face, and body modality. Affect Y is any different affect from the driving affect, Affect X.

Model Condition	Dialogue	Voice	Face	Body
AC	Affect X	Affect X	Affect X	Affect X
IV	Affect X	Affect Y	Affect X	Affect X
IF	Affect X	Affect X	Affect Y	Affect X
IB	Affect X	Affect X	Affect X	Affect Y

We rendered the multimodal responses of each model as the stimulus in our user studies. We chose 5 dialogues from the late scenario, each containing 6 text responses associated with the 6 driving affects (Affect X in Table 4). For every text response, we then generated the multimodal behaviors for the 4 comparison models. The affect-consistent model was used to generate only one sample, but the 3 affect-inconsistent models (IV, IF, and IB) were used to generate 3 samples, with different Affect Y (explained in Table 4). One was neutrality and the other two were randomly selected from the remaining affects. In total, 300 video samples (5 dialogues $\times$ 6 sentences $\times$ 10 affect settings) were generated for the main user study.

We distributed our survey and collected responses from the participants through Qualtrics [48]. In the survey, each participant was first presented with a recorded video with a question displayed. We then asked the participants to answer to what extent the character's response aligned with the 6 defined affects using the 7-point Likert Scale. An answer of 1 on the scale indicates strong disagreement, 7 indicates strong agreement, and 4 is the threshold between agreement and disagreement. The same survey question was repeated 25 times for each participant, with videos randomly selected from the rendered responses. In the experiment, a total of 199 participants were recruited, most of which were university undergraduates with little or no knowledge of ECAs. On average, each video was rated by 49 different participants. The survey was taken anonymously and strictly followed the university IRB rules.

Table 5: The average and standard deviations of the Likert scores for the four comparison models. The columns indicate the driving affects and the rows represent the perceived affects. Bold face means the highest average score among all perceived affects. The symbol * denotes the significant decrease in the correct perception score after the removal of a consistent affect under $\alpha < 0.05$, while † means the significant decrease under $\alpha < 0.001$. Entries with gray background mean the highest perception score does not occur when the perceived affect is the same as the driving affect. CPIN refers to the correct perception of the driving affect when the inconsistent affect is neutrality.

Model Condition	Perceived Affect	Driving Affect					
		Neutral	Angry	Happy	Bored	Sad	Surprised
AC	Neutral	**4.76 ± 1.69**	3.12 ± 1.29	3.59 ± 1.53	4.10 ± 1.97	3.39 ± 1.44	3.78 ± 1.57
	Angry	3.76 ± 1.73	**4.87 ± 1.63**	2.97 ± 1.29	2.86 ± 1.13	2.97 ± 1.19	2.98 ± 1.38
	Happy	2.70 ± 0.79	3.52 ± 1.43	**4.39 ± 1.72**	3.38 ± 1.16	2.81 ± 0.95	3.41 ± 1.51
	Bored	3.55 ± 1.53	3.10 ± 1.13	3.07 ± 1.41	**4.40 ± 1.69**	3.05 ± 1.38	3.05 ± 1.28
	Sad	2.82 ± 1.21	3.03 ± 1.16	3.42 ± 1.41	3.79 ± 1.99	**5.37 ± 1.66**	3.63 ± 1.65
	Surprised	2.76 ± 0.78	3.27 ± 1.73	3.73 ± 1.63	3.45 ± 1.21	3.22 ± 1.30	**4.11 ± 1.90**
IV	Neutral	†3.83 ± 1.76	3.67 ± 1.92	**3.70 ± 1.91**	**4.30 ± 1.84**	3.41 ± 1.55	**3.78 ± 1.83**
	Angry	3.62 ± 1.65	†**3.67 ± 1.67**	3.05 ± 1.44	3.21 ± 1.57	2.91 ± 1.52	3.48 ± 1.54
	Happy	3.17 ± 1.21	3.16 ± 1.45	†3.48 ± 1.54	3.19 ± 1.29	2.86 ± 1.05	3.48 ± 1.45
	Bored	**3.95 ± 1.81**	3.62 ± 1.71	3.51 ± 1.67	†3.53 ± 1.52	3.19 ± 1.54	3.40 ± 1.66
	Sad	3.54 ± 1.76	3.33 ± 1.60	3.34 ± 1.59	3.79 ± 1.91	***4.75 ± 1.78**	3.56 ± 1.60
	Surprised	3.44 ± 1.37	3.16 ± 1.63	3.07 ± 1.69	3.18 ± 1.40	3.10 ± 1.41	3.72 ± 1.63
	CPIN	N/A	†3.72 ± 1.70	†3.65 ± 1.63	†3.59 ± 1.48	*4.59 ± 1.87	3.70 ± 1.71
IF	Neutral	†**3.81 ± 1.66**	3.58 ± 1.61	**3.66 ± 1.89**	**4.39 ± 2.07**	3.52 ± 1.91	**3.71 ± 1.82**
	Angry	3.13 ± 1.58	†**3.80 ± 1.91**	2.79 ± 1.47	3.02 ± 1.30	2.93 ± 1.35	3.36 ± 1.51
	Happy	3.14 ± 1.27	3.14 ± 1.40	†3.49 ± 1.54	3.29 ± 1.16	2.89 ± 1.01	3.46 ± 1.37
	Bored	3.55 ± 1.77	3.33 ± 1.71	3.33 ± 1.48	†3.57 ± 1.69	3.00 ± 1.55	3.43 ± 1.59
	Sad	3.47 ± 1.73	3.31 ± 1.44	3.27 ± 1.45	3.81 ± 2.06	†**4.05 ± 1.92**	3.69 ± 2.01
	Surprised	3.21 ± 1.33	3.08 ± 1.63	3.14 ± 1.56	3.12 ± 1.26	3.01 ± 1.34	*3.62 ± 1.69
	CPIN	N/A	†3.83 ± 1.86	†3.41 ± 1.46	†3.63 ± 1.36	†3.99 ± 1.89	*3.63 ± 1.76
IB	Neutral	***4.18 ± 1.69**	3.42 ± 1.57	3.00 ± 1.60	**3.99 ± 2.12**	3.27 ± 1.60	3.58 ± 1.76
	Angry	3.39 ± 1.69	***4.39 ± 1.97**	3.16 ± 1.54	3.04 ± 1.37	2.93 ± 1.27	3.40 ± 1.53
	Happy	3.09 ± 1.25	3.23 ± 1.37	†3.45 ± 1.58	3.05 ± 1.15	2.99 ± 1.01	3.22 ± 1.39
	Bored	3.61 ± 1.93	3.72 ± 1.66	**3.49 ± 1.43**	†3.55 ± 1.78	3.19 ± 1.38	3.24 ± 1.61
	Sad	3.55 ± 1.74	3.20 ± 1.40	2.96 ± 1.51	3.65 ± 1.97	***4.69 ± 1.91**	3.34 ± 1.87
	Surprised	2.99 ± 1.27	3.43 ± 1.58	3.31 ± 1.87	3.14 ± 1.19	3.26 ± 1.40	**3.69 ± 1.63**
	CPIN	N/A	4.46 ± 1.94	†3.56 ± 1.44	†3.78 ± 1.65	4.87 ± 1.96	3.86 ± 1.71

5 RESULTS

5.1 Confusion Matrices for All Model Conditions

For each of the 4 model conditions, Table 5 reports the average and standard deviation of the Likert scores for each affect's perceived alignment with a driving affect, where the affect being compared to the driving affect is referred to as the *perceived affect*. Each model's results are a 6×6 confusion matrix, for which the diagonal entries match the perceived and driving affects. We refer to it as the perception of the driving affect or the correct perception in our results.

5.1.1 ***When all modalities are affect-consistent, participants recognize the driving affect.*** The AC model condition receives the highest Likert scores along the diagonal entries, with all average scores greater than 4. This indicates that with consistent affects across all modalities, participants correctly recognize that there is the most alignment when the perceived and driving affects are the same Among all the diagonal entries, participants found sadness to be the most recognizable driving affect and surprise to be the least, which indicates that the sad and surprised behaviors have the highest and lowest degree of expressiveness respectively. Moreover, we can make observations about the non-diagonal perceived affects that may be confused with the driving affect for the AC model. This confusion of alignment between the driving affect and perceived affect is most notable where the average Likert score of a misperceived affect is close to or greater than 4, meaning that the average participant somewhat agreed with the misalignment. When the driving affect is happiness, the Likert score of surprise is close to 4. When the bored behaviors are presented to the participants, the perceived neutrality and sadness are also strong. When surprise is used as the driving affect, the perceived neutrality score is also high. Some confusions can be explained by the valence-arousal circumplex. For example, happiness and surprise are close to each

other in terms of their valence arousal positions, and boredom and sadness have close proximity. However, other confusions, including the boredom-neutrality and surprise-neutrality pairs, can be the cause of low degree of expressiveness, so the participants tend to rate the emotional behaviors with higher neutrality scores.

5.1.2 ***One inconsistent modality can disrupt the recognition of almost all driving affect.*** We can see how the removal of a consistent affect from one modality influences the perception of the driving affect, which is otherwise perceived correctly as evidenced by the AC model in Table 5. For example, the confusion matrix of the IV model shows that the inconsistent affect in the voice modality decreases the Likert scores at the diagonal entries. Most of the Likert scores drop below 4, which suggests that the participants do not agree the perceived affects are the same as the driving affects. We conduct the one-tailed t-test [59] between the AC model and three other model conditions on their correct perception scores. The result is provided at the diagonal cells in Table 5 (denoted as * and †). There are significant decreases in Likert scores after the removal of consistent affect in voice and face modality for the correct perception of all affects. However, there is no significant decrease after the removal of a consistent affect in body modality for surprise perception. The result also implies that an inconsistent affect in the voice and face modalities leads to more statistically significant decreases than the body modality in the correct perception scores.

5.1.3 ***Some affects are more resilient than others to the inconsistency of one modality.*** When a consistent affect is removed from any of the three modalities, we see that the two driving affects, anger and sadness, can still be recognized by participants. Neutrality and surprise are less resilient because the perception of the same driving affect is the highest after the removal of each of the two affects in the body modality. When a happy or bored consistent affect is removed from any of the three modalities, the correct perception of the two affects is largely influenced. In fact, we can see a link between the decrease in the correct perception of driving affect and the increase in the perception of neutrality. The entries highlighted in gray at Table 5 indicate that the perceived neutrality score also increases for those irresilient driving affects. The three driving affects, happiness, boredom, and surprise are even perceived as more neutral for IV and IF models. We attribute the decrease in the correct perception of driving affect as well as the increase in perceived neutrality after the removal of a consistent affect to the cause of emotion dilution, when the expressiveness of the multimodal behavior is discounted by affect inconsistency. Overall, we observe that the removal of a consistent affect in voice (IV) and face (IF) modalities increases the perception of neutrality more than the removal in the body (IB) modality. This also implies that the emotion dilution issue is more obvious when either the voice or the face modality has an inconsistent affect.

5.1.4 ***A modality without emotion conditioning can decrease the perception of the driving affect.*** We have mentioned in Section 3 that emotion conditioning can be helpful as it mitigates the emotion dilution issue during the sequential executions for multimodal behaviors. We regard the stimulus with the inconsistent neutrality affect in the IV, IF, and IB model conditions as the behaviors generated without emotion conditioning in the voice, face, and body modalities respectively. Each 'CPIN' row in Table 5 reports the average and standard deviation of the Likert scores for the correct perception of all driving affects. Without emotion conditioning in one modality, the perception of the driving affect decreases, when compared with the AC model with all emotion-conditioned modalities. We also notice that emotion dilution is more obvious when the correct perception score is much more influenced by the unconditioned modality. This suggests that when the emotion conditioning is removed from one modality, the generated behaviors in that modality are perceived as less expressive and more neutral, which then decreases the correct perception of affects when other affective modalities are combined.

5.2 Interaction and Main Effects on Perception

We conduct the 3-way ANOVA test [33] to analyze the effect of the three independent variables, *model*, *driving affect*, and *perceived affect*, on the perception Likert scores and we apply Tukey HSD [1] for the post hoc tests. There are 4 conditions in the model variable and 6 conditions each for the two remaining variables. The result shows that there is a statistically significant interaction between the effects of *model*, *driving affect*, and *perceived affect* on the perception Likert score, with $F(75, 12444) = 2.105$ and $p < 0.001$.

Specifically, we observe a main effect on *perceived affect*, with $F(5, 12444) = 28.040$ and $p < 0.001$. The post hoc test indicates that across all conditions in the *model* and *driving affect* variables, the perception scores of neutrality and sadness are significantly higher than the scores of anger, happiness, boredom, and surprise, and the boredom perception score is significantly higher than happiness. This implies that the participants tend to rate the stimulus with higher Likert scores in neutrality and sadness. The reason behind the higher sadness rating is that it receives the highest when the driving affect is sadness. For neutrality, we can observe from Table 5 that neutrality is often misperceived as the highest for all affect inconsistent models. Across all driving affects, there is a simple main effect of the *perceived affect* variable on the Likert score for every model condition. The perception of the neutrality affect under AC model is statistically more significant than the perception of boredom, surprise, anger, and happiness, while the neutrality perception is much more significant than the perception of all other affects for all three affect inconsistent models (IV, IF, and IB). This shows that the emotion dilution is more obvious when the affects are inconsistent.

We also observe a main interaction effect between the *driving affect* and *perceived affect*, with $F(25, 12444) = 24.328$ and $p < 0.001$. Ideally, when the driving affect matches the perceived affect, the Likert score should be significantly higher than other driving affects. For all 6 conditions in the *perceived affect* variable, all the conditions in the *driving affect* variable with the same affect condition show statistically significant differences from other driving affects with dissimilar conditions, except one pair, neutrality, and boredom. The reason can be seen from Table 5 that when the perceived affect is neutrality, the boredom driving affect is rated relatively high for the three affect inconsistent models. We observe a similar simple effect of the *driving affect* variable on the Likert score when the

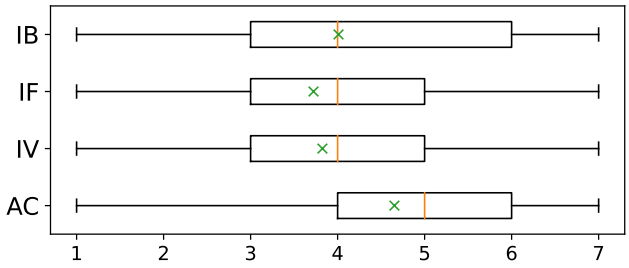

Figure 3: The box plot of the correct perception score for the four model conditions. 'X' denotes the average score.

AC model is considered. All driving affects that are the same as the perceived affect are rated significantly higher than the driving affects that are different from the perceived affects. However, when the perceived affect is surprise, the Likert score for the surprise driving affect is not rated significantly different from happiness. According to Table 5, the average Likert score for happiness driving affect and perceived surprise is 3.73 and the perceived surprise score for surprise driving affect is 4.11. This suggests that the surprised multimodal behavior is not significantly different from the happy behavior in terms of the perception of surprise.

As we observe the interaction effect of *driving affect * perceived affect*, we are also interested in the comparisons of the *perceived affects* on the Likert score under each *model* condition. Ideally when the perceived affect is the same as the driving affect, the Likert score should be significantly higher than other perceived affects. We find most comparison pairs follow the above expectation for AC model. However, when the driving affect is boredom, the perception of boredom is not significantly different from sadness. The result suggests that there is confusion between the perception of boredom and sadness when the bored behavior is provided. We believe this is due to the proximity of boredom and sadness on their valence-arousal values [51]. When the driving affect is surprise or boredom, we find no significant differences between the perception of the same affect and neutrality, which indicates that these generated emotional behaviors are not as expressive as others.

5.3 Effects on Correct Perception of Affects

We are also interested in the interaction of the model and the driving affect on the correct perception of Likert score. The correct perception of an expressive behavior means the perception of the same affect as the driving affect. In other words, all the diagonal entries in Table 5 are considered. We further conduct the 2-way ANOVA test to report the effect of the two independent variables, *model* and *driving affect*, on the correct affect perception in Likert scores. The result shows that there is no significant interaction of the two independent variables on the correct perception, with $F(15, 2815) = 1.640$ and $p < 0.056$. However, we observe the main effects of *model* and *driving affect*, with $F(3, 2815) = 27.773$, $p < 0.001$ and $F(5, 2815) = 21.578$, $p < 0.001$ respectively. We apply Tukey HSD for the post hoc test for the comparisons of the conditions in the two variables.

5.3.1 Difference in Models. The differences between the AC model and the IV, IF, and IB models across all driving affects represents the discrepancy in the correct affect perception between an affect-consistent behavior and an inconsistent behavior. The result indicates that across all driving affects, the AC model receives a significantly higher Likert score than the IV, IF, and IB model conditions, with average differences = 0.822, 0.928, 0.660 and $p < 0.001$. Adding a consistent affect to one of the voice, face, and body modality can significantly improve the perception of the driving affect. Specifically, the average score differences in voice and face modality are higher than the body modality, which suggests that the voice and face modality are more important than body when expressing affects. We can observe the same finding from Figure 3 that IB model condition has higher average score than IV and IF, and still receives a similar 75th percentile as AC model condition. Our findings are aligned with the previous studies [15, 22, 57] that all modalities do contribute to affect perception and combining the expressive modalities is more helpful. The perception, however, is not mainly judged by body expressions.

5.3.2 Difference in Driving Affects. The comparison of the conditions in *driving affect* on the correct perception score reflects how easily the affects can be recognized by the participants. Our result indicates that when the driving affect is sadness, the correct perception score is significantly higher than all other affects, across all *model* conditions. The three affects, happiness, boredom, and surprise, however, are hard to recognize as they receive relatively lower scores than the other three affects. We observe that the removal of the three consistent affects in a modality would largely decrease their correct perception scores, while the other three are somewhat more resilient. The low expressiveness of the three emotional behaviors could be the limitation of our collected methods. Furthermore, when looking specifically at every model condition, we can still tell that sadness is the most recognizable among all the affects, as it receives the highest score for correct perception. However, we can only see significant differences in the correct perception between the two pairs, sadness-happiness, and sadness-surprise, where the affects have opposite valence and arousal values. This shows that the sad face is a major indicator of the recognition of sadness.

6 DISCUSSION

The results of our user study indicate that affect consistency maximizes the perception of the driving affect and making a single modality's affect inconsistent decreases the perception of the correct affect. The conclusion is supported by previous studies [15, 22], which investigated the effect that the presence of face and body modalities had on perception. They found that both modalities individually contribute to affect perception and that the perception is maximized when both modalities are present. Furthermore, our experimental analysis suggests that the voice and face modalities contribute more than the body modality to the perception of affects. This finding accords with prior works [22, 57] regarding the perception of emotions for virtual characters, although the experimental designs are different. While their stimuli contain all combinations of the presence of each modality with consistent affect, ours include all modalities with both consistent and inconsistent affects. The

differences in the modalities in our study need to be analyzed by comparing the differences between the affect consistent model and inconsistent models. Nonetheless, our comparisons of modalities can be done in a more natural conversation scenario where all modalities are present.

From the statistical analysis of the main and simple effects of *perceived affect* variable, we can observe that overall, neutrality is strongly perceived, despite it not being the driving affect. The observation still holds true regardless of the *model* condition, but the perception of neutrality is stronger when a modality has an inconsistent affect. Our study shows that the emotion dilution occurs when the one modality has an inconsistent affect and the participants tend to rate a higher neutrality perception score. The issue, however, is mitigated under the AC model condition where all the modalities are generated with consistent emotion conditioning. Evidenced by the experimental results and analyses, we can conclude that our conceptual framework for ECAs with multimodal emotion conditioning and affect consistency successfully addresses emotion dilution and enhances the correct perception of affect.

The correlations of the affects are investigated in our study through the analyses of the interaction effect between the *driving affect* and *perceived affect* as well as the difference in the effect of the *driving affects* on the correct perception. We do observe certain links that explain the correlations of affects in the confusion matrix with the valence-arousal circumplex. For example, boredom can be confused with sadness, and happiness is correlated with surprise. However, we do not observe that the body modality helps the participants discriminate emotions based on their valence values, as suggested by the previous study [3]. There are some confusions that cannot be explained by their valence-arousal values. For instance, the neutrality perception scores are also high when the driving affects are boredom and surprise, which indicates that the two emotional behaviors are not as expressive as other affective behaviors. The reason could result from the limitation of the collected data-driven models. Currently, the expressiveness of the framework is dependent on the methods we use to generate the modalities. We can see the dependence from our results that sadness is the most recognizable affect among all, but the previous works [15, 22] on the perception of virtual characters revealed that anger and happiness are more easily recognized, and sadness is the most difficult to tell. We acknowledge a gap in expressiveness between the real human behaviors that were used as stimulus by previous studies and the generated human behaviors in our experiments. Nonetheless, the affect-consistent framework is able to generate emotional behaviors with increased correct perception.

Overall, our framework combines the existing methods for the voice, face, and body modalities and creates a shared affect space for generating affect consistent behaviors. We could see several limitations. First, the quality of the affect consistent framework is dependent on those methods. For instance, the use of gesture generation from audio and text by the framework could potentially lead to a loss of communication efficacy when compared with rule-based methods [5, 50, 58]. According to the classification of gesture types proposed by [44], our generative approach could only generate beat and metaphoric gestures from the input audio and text. Second, the mappings of the 6 affects and the latent parameters are only specific to our implementation, which decreases the generalizability of the proposed framework. However, the methods can be easily substituted and the mappings can be obtained in different ways accordingly. For example, the affect mappings can be heuristics-based, manually selected and verified through user study, or achieved via supervised training. The framework itself can be improved as a better method for modality generation is available. Still, the importance of the two properties, emotion conditioning and affect consistency, holds true.

On the other hand, when it comes to real-time interaction, our framework does require longer process time compared with pre-authored behaviors. The latency comes from the interdependence of the modalities, so all the modalities cannot be generated in parallel. The incorporation of separate generative methods into the framework also introduces additional latency as the generated facial and body animations have to be retargeted to the same character at runtime.

7 CONCLUSION

In this work, we propose a conceptual framework, ACTOR, with affect-consistent multimodal behaviors for ECAs that aims to enhance the user's perception of affects. We conduct the main user study for the evaluation of our framework. The result indicates that the multimodal behavior with consistent affect receives the highest correct perception score and removing a consistent affect from the voice, face, and body modalities can significantly decrease the perception of the driving affect. Our statistical analysis also suggests that emotion conditioning and affect consistency are helpful for mitigating the emotion dilution issue.

ACKNOWLEDGMENTS

The research was supported in part by NSF awards: IIS-1703883, IIS-1955404, IIS-1955365, RETTL-2119265, and EAGER-2122119. This material is based upon work supported by the U.S. Department of Homeland Security[1] under Grant Award Number 22STESE00001 01 01. This publication is based upon work supported by King Fahd University of Petroleum & Minerals. Author(s) at KFUPM acknowledge the Interdisciplinary Research Center for Intelligent Secure Systems for the support received under Grant Number INSS2305.

REFERENCES

[1] Hervé Abdi and Lynne J Williams. 2010. Tukey's honestly significant difference (HSD) test. *Encyclopedia of research design* 3, 1 (2010), 1–5.
[2] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.
[3] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* 338, 6111 (2012), 1225–1229.
[4] Farnaz Badiee and David Kaufman. 2015. Design evaluation of a simulation for teacher education. *Sage Open* 5, 2 (2015), 2158244015592454.
[5] Kirsten Bergmann and Stefan Kopp. 2009. Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. 361–368.
[6] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning. In *Proceedings of*

[1]Disclaimer. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

the 29th ACM International Conference on Multimedia (MM '21). Association for Computing Machinery, New York, NY, USA.

[7] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents** This work has been supported in part by ARO Grants W911NF1910069 and W911NF1910315, and Intel. Code and additional materials available at: https://gamma. umd. edu/t2g. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 1–10.

[8] Pieter A Blomsma, Guido M Linders, Julija Vaitonyte, and Max M Louwerse. 2020. Intrapersonal dependencies in multimodal behavior. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[9] Jacky Casas, Timo Spring, Karl Daher, Elena Mugellini, Omar Abou Khaled, and Philippe Cudré-Mauroux. 2021. Enhancing conversational agents with empathic abilities. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 41–47.

[10] Susana Castillo, Philipp Hahn, Katharina Legde, and Douglas W Cunningham. 2018. Personality analysis of embodied conversational agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 227–232.

[11] Che-Jui Chang. 2020. Transfer Learning from Monolingual ASR to Transcription-free Cross-lingual Voice Conversion. https://doi.org/10.48550/ARXIV.2009.14668

[12] Che-Jui Chang, Sen Zhang, and Mubbasir Kapadia. 2022. The TeamName entry to the GENEA Challenge 2022 – A Tacotron2 Based Method for Co-Speech Gesture Generation With Locality-Constraint Attention Mechanism. *in press* (2022).

[13] Che-Jui Chang, Long Zhao, Sen Zhang, and Mubbasir Kapadia. 2022. Disentangling audio content and emotion with adaptive instance normalization for expressive facial animation synthesis. *Computer Animation and Virtual Worlds* 33, 3-4 (2022), e2076. https://doi.org/10.1002/cav.2076 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cav.2076

[14] Chaona Chen, Oliver GB Garrod, Philippe G Schyns, and Rachael E Jack. 2020. Dynamic Face Movement Texture Enhances the Perceived Realism of Facial Expressions of Emotion. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–3.

[15] Céline Clavel, Justine Plessier, Jean-Claude Martin, Laurent Ach, and Benoit Morel. 2009. Combining facial and postural expressions of emotions in a virtual character. In *International Workshop on Intelligent Virtual Agents*. Springer, 287–300.

[16] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-Driven Dialog Generation. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT* (2019-01-01). 3734–3743. https://aclweb.org/anthology/papers/N/N19/N19-1374/

[17] Alban Delamarre, Cédric Buche, and Christine Lisetti. 2019. Aimer: Appraisal interpersonal model of emotion regulation, affective virtual students to support teachers training. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 182–184.

[18] Steve DiPaola and Özge Nilay Yalçin. 2019. A multi-layer artificial intelligence and sensing based affective conversational embodied agent. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 91–92.

[19] Funda Durupinar, Mubbasir Kapadia, Susan Deutsch, Michael Neff, and Norman I Badler. 2016. Perform: Perceptual approach for adding ocean personality to human motion using laban movement analysis. *ACM Transactions on Graphics (TOG)* 36, 1 (2016), 1–16.

[20] Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion* 98, 45-60 (1999), 16.

[21] Unreal Engine. 2021. MetaHuman Creator.

[22] Cathy Ennis, Ludovic Hoyet, Arjan Egges, and Rachel McDonnell. 2013. Emotion capture: Emotionally expressive characters for games. In *Proceedings of motion on games*. 53–60.

[23] Jessica Falk, Steven Poulakos, Mubbasir Kapadia, and Robert W Sumner. 2018. Pica: Proactive intelligent conversational agent for interactive narratives. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 141–146.

[24] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Understanding the predictability of gesture parameters from speech and their perceptual importance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[25] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2021. ExpressGesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds* 32, 3-4 (2021), e2016.

[26] Ylva Ferstl, Sean Thomas, Cédric Guiard, Cathy Ennis, and Rachel McDonnell. 2021. Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 76–83.

[27] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[28] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 101–108.

[29] Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An end-to-end conversational style matching agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 111–118.

[30] IBM. 2015. IBM Text to Speech. https://www.ibm.com/watson. Accessed: 2022-03-05.

[31] Ryo Ishii, Chaitanya Ahuja, Yukiko I Nakano, and Louis-Philippe Morency. 2020. Impact of personality on nonverbal behavior generation. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[32] Sepehr Janghorbani, Ashutosh Modi, Jakob Buhmann, and Mubbasir Kapadia. 2019. Domain authoring assistant for intelligent virtual agents. *arXiv preprint arXiv:1904.03266* (2019).

[33] Charles M Judd, Gary H McClelland, and Carey S Ryan. 2017. *Data analysis: A model comparison approach to regression, ANOVA, and beyond*. Routledge.

[34] Mubbasir Kapadia, Fabio Zünd, Jessica Falk, Marcel Marti, Robert W Sumner, and Markus Gross. 2015. Evaluating the authoring complexity of interactive narratives with interactive behaviour trees. *Foundations of Digital Games* (2015).

[35] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

[36] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*. Springer, 205–217.

[37] Jina Lee and Stacy Marsella. 2006. Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 243–255.

[38] Tao Li, Shan Yang, Liumeng Xue, and Lei Xie. 2021. Controllable emotion transfer for end-to-end speech synthesis. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 1–5.

[39] Meng Liu, Yaocong Duan, Robin AA Ince, Chaona Chen, Oliver GB Garrod, Philippe G Schyns, and Rachael E Jack. 2020. Building a generative space of facial expressions of emotions using psychological data-driven methods. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–3.

[40] Yu Liu, Gelareh Mohammadi, Yang Song, and Wafa Johal. 2021. Speech-based Gesture Generation for Robots and Embodied Agents: A Scoping Review. In *Proceedings of the 9th International Conference on Human-Agent Interaction*. 31–38.

[41] Max M Louwerse, Rick Dale, Ellen G Bard, and Patrick Jeuniaux. 2012. Behavior matching in multimodal communication is synchronized. *Cognitive science* 36, 8 (2012), 1404–1426.

[42] Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60, 2 (1992), 175–215.

[43] Rachel McDonnell, Sophie Jörg, Joanna McHugh, Fiona Newell, and Carol O'Sullivan. 2008. Evaluating the emotional content of human motions on real and virtual characters. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*. 67–74.

[44] David McNeill. 1992. Hand and Mind: What Gestures Reveal About Thought. (1992).

[45] Rajmund Nagy, Taras Kucherenko, Birger Moell, André Pereira, Hedvig Kjellström, and Ulysses Bernardet. 2021. A framework for integrating gesture generation models into interactive conversational agents. *arXiv preprint arXiv:2102.12302* (2021).

[46] Nvidia. 2021. Omniverse Audio2Face.

[47] Delphine Potdevin, Céline Clavel, and Nicolas Sabouret. 2018. Virtual Intimacy, this little something between us: a study about Human perception of intimate behaviors in Embodied Conversational Agents. In *Proceedings of the 18th international conference on intelligent virtual agents*. 165–172.

[48] Qualtrics. 2021. *Qualtrics*. Qualtrics, Provo, Utah, USA. http://www.qualtrics.com

[49] Tanmay Randhavane, Aniket Bera, Kyra Kapsaskis, Rahul Sheth, Kurt Gray, and Dinesh Manocha. 2019. Eva: Generating emotional behavior of virtual agents using expressive features of gait and gaze. In *ACM symposium on applied perception 2019*. 1–10.

[50] Brian Ravenet, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella. 2018. Automating the production of communicative gestures in embodied characters. *Frontiers in psychology* 9 (2018), 1144.

[51] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[52] Pejman Sajjadi, Laura Hoffmann, Philipp Cimiano, and Stefan Kopp. 2019. A personality-based emotional model for embodied conversational agents: Effects on perceived social presence and game experience of users. *Entertainment Computing* 32 (2019), 100313.

[53] Marc Schröder. 2001. Emotional speech synthesis: A review. In *Seventh European Conference on Speech Communication and Technology*. Citeseer.

[54] Alexander Shoulson, Nathan Marshak, Mubbasir Kapadia, and Norman I Badler. 2013. Adapt: the agent developmentand prototyping testbed. *IEEE Transactions on Visualization and Computer Graphics* 20, 7 (2013), 1035–1047.

[55] Maayan Shvo, Jakob Buhmann, and Mubbasir Kapadia. 2019. An interdependent model of personality, motivation, emotion, and mood for intelligent virtual agents. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 65–72.

[56] Samuel S Sohn, Xun Zhang, Fernando Geraci, and Mubbasir Kapadia. 2018. An emotionally aware embodied conversational agent. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 2250–2252.

[57] Sinan Sonlu, Uğur Güdükbay, and Funda Durupinar. 2021. A conversational agent framework with multi-modal personality expression. *ACM Transactions on Graphics (TOG)* 40, 1 (2021), 1–16.

[58] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. , 209–232 pages.

[59] Bernard L Welch. 1947. The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika* 34, 1-2 (1947), 28–35.

[60] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2022. A review of evaluation practices of gesture generation in embodied conversational agents. *IEEE Transactions on Human-Machine Systems* (2022).

[61] Özge Nilay Yalçın. 2020. Empathy framework for embodied conversational agents. *Cognitive Systems Research* 59 (2020), 123–132.

[62] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.

[63] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots. In *Proc. of The International Conference in Robotics and Automation (ICRA)*.

[64] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing-Qian Liu. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *AAAI*.