Multimodal Embodied Conversational Agents: A discussion of architectures, frameworks and modules for commercial applications

Conference Paper · December 2022

DOI: 10.1109/AIVR56993.2022.00013

4 authors, including:

Kumar Shubham
Indian Institute of Science Bangalore
22 PUBLICATIONS **170** CITATIONS

SEE PROFILE

Laxmi Narayen Nagarajan Venkatesan
International Institute of Information Technology Bangalore
2 PUBLICATIONS **2** CITATIONS

SEE PROFILE

Dinesh Jayagopi
International Institute of Information Technology Bangalore
106 PUBLICATIONS **1,531** CITATIONS

SEE PROFILE

Multimodal embodied conversation agents: A discussion of architectures, frameworks and modules for commercial applications

Kumar Shubham*
International institute of information technology
Bangalore, India
kumar.shubham@iiitb.ac.in

Laxmi Narayen Nagarajan Venkatesan*
International institute of information technology
Bangalore, India
laxminarayen.nv@iiitb.ac.in

Dinesh Babu Jayagopi
International institute of information technology
Bangalore, India
jdinesh@iiitb.ac.in

Raj Tumuluri
Openstream.ai
Somerset, New Jersey, United States
raj@openstream.ai

***Abstract*—With the recent advancements in automated communication technology, many traditional businesses that rely on face-to-face communication have shifted to online portals. However, these online platforms often lack the personal touch essential for customer service. Research has shown that face-to-face communication is essential for building trust and empathy with customers. A multimodal embodied conversation agent (ECA) can fill this void in commercial applications. Such a platform provides tools to understand the user's mental state by analyzing their verbal and non-verbal behaviour and allows a human-like avatar to take necessary action based on the context of the conversation and as per social norms. However, the literature to understand the impact of ECA agents on commercial applications is limited because of the issues related to platform and scalability. In our work, we discuss some existing work that tries to solve the issues related to scalability and infrastructure. We also provide an overview of the components required for developing ECAs and their deployment in various applications.**

***Index Terms*—ECA, Avatar, Multimodal Conversational Interaction, Human-Robot Interaction, Conversational AI, Conversational Agents**

I. Introduction

Communication plays a vital role in many commercial and social exchanges. These interactions help create a sense of mutual trust and understanding between individuals. For example, in business, professionals often use persuasive tactics in person to drive and reach an agreement [3]. In customer services, face-to-face interactions help customers clear their doubts and conflict management with the organization [58]. Straus and McGrath [80] have further shown that the type of communication medium affects the outcome of a business process.

Face-to-face communication essentially involves multimodal sensory inputs [76], Non-verbal cues like body movements, facial expressions, and gestures of the interaction partner enrich the conversation [3], [76]. Studies [30], [50], [84] show that individuals use non-verbal cues from their interaction partner to determine whether or not their goals are met. Information related to frustration, depression, or anger can be predicted using verbal and non-verbal cues [17], and such information can hint at the need for assistance from the partner. Similarly, multimodal output generation for the postures and backchannels produced by an individual during a conversation also plays a critical role in developing shared trust [34] and is essential for an engaging conversation. Seibt et al. [72] have further shown that facial mimicry and facial expression dynamics often influence the interaction partner's cognitive and emotional state.

With the recent technological advancement, many businesses have shifted online. Taking business online has helped them reach an unprecedented scale with a global customer base and even helped customers benefit from competitive prices and world-class products. Many of these services rely on user-friendly websites or text-based chatbot services to meet the needs of their customers. Nevertheless, such services often lack the personal touch of face-to-face communication, which is essential for the satisfaction of customers.

Studies [53], [54] have shown that the rules and nuances of face-to-face conversation can be extended to human-computer interactions. Any medium resembling human-like qualities will get a human-like response [53], [54]. In many of these online services, companies provide artificial intelligence (AI) based chat-bot services to assist their customers. However, these services often lack the necessary multimodal interaction for engaging and satisfactory conversation. The agent's voice (Text-to-speech - TTS) can also significantly impact users' perception of cognitive and emotional trust [18], [64]. Many companies even commercialised these voice-based assistance

*These authors contributed equally to this work

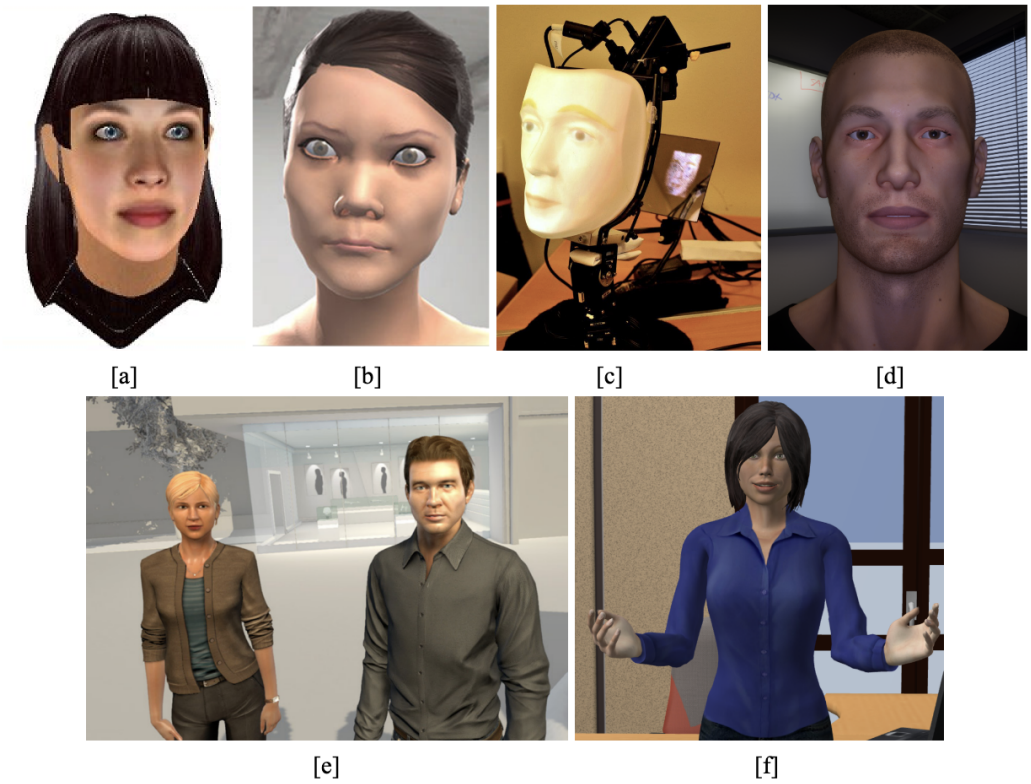

Fig. 1. Avatars from [a] LUCIA [43]. [b] FACSvatar [87]. [c] Furhat from IrisTK [77]. [d] Avatarsim [5]. [e] Smartbody [82]. [f] GRETA [19]

services, such as Siri [1], Cortana [2], and Alexa [3]. However, these services are missing two important components: a way to interpret and express nonverbal behaviours. Many socially meaningful discrete events like "smiling" and "nodding" play an important role during a conversation [30], [84]. A multimodal embodied conversational agent (ECA) with a prosodically rich synthesized voice can help in achieving the required human-like behaviour [64] by generating necessary backchannels [46] and socially meaningful events [31].

A multimodal embodied conversational agent generates the necessary multimodal output consisting of verbal and non-verbal behaviour of a 3D embodied avatar based on sensory input from the interaction partner's audio, visual, and textual features. Such a system continuously monitors the mental state [37], [71] of its counterpart and takes the necessary decisions as per the goal of the interaction. Recently, several studies [14], [63] have explored the application of such an agent for a dyadic conversation and have shown its effectiveness in engaging their interaction partner.

Although multimodal embodied conversational agents have been studied in the past, these platforms are used in very few commercial applications. One of the significant associated challenges has been a lack of appropriate architecture and infrastructure that can handle the video and audio data of hundreds of users and orchestrate the human-like conduct and TTS output. In this paper, we discuss different components and tools used by researchers to create multimodal embodied conversational agents and their limitations for commercial applications. A better understanding of these components can help in creating ECA agents that can serve multiple users on an online platform. Our work will also help the developers and business managers to get a holistic idea of the components and important modules associated with the ECA platform, before starting any commercial project

[1]https://www.apple.com/in/siri/
[2]https://www.microsoft.com/en-us/cortana
[3]https://developer.amazon.com/en-US/alexa

II. Application

Embodied conversational agents can be used in a variety of settings in business. A few examples of their application are as follows:

- **Customer service**: Customer support has been one of the most appealing use cases for embodied conversation agents, which has garnered much attention in recent years [26], [48], [49], [88]. In the given scenario, an agent interacts with the customers to understand their grievances, and attempts to resolve them. Such conversations often necessitate the use of highly trained professionals. Considering the scale of an online business, it becomes challenging and expensive for a business to hire a large pool of professionals to handle customer support. An intelligent embodied conversational agent can assist a company in scaling up its customer sup-

port team while also providing an opportunity to make these interactions more professional and personalized. Domains like insurance have a direct use case for such a system, where agents can interact with the customer to resolve claim-related issues. Along similar lines, **EVA**[4] and Yacoubi et al. [91] have demonstrated a framework that uses the emotional state as a driving factor for action in a conversational agent that interacts with the customer. Similar applications have been demonstrated in healthcare, where an agent can interact with a user either as a psycho-social companion [67], [75] or for healthcare support [9], [21], [68].

- **Recruitment**: Embodied conversational agents can also be used to interview a large pool of candidates [59], [61], [73]. Choosing an ECA for interviews has two advantages. First, interviewing an individual is often a time-consuming task that is limited by the interviewer's availability; second, ECA's are unbiased in judging the candidates [59]. An embodied conversational agent can speed up the interview process and make the interview process more flexible and comfortable for the candidate by allowing them to schedule the interview at their preferred time and location.
- **Marketing**: Companies have traditionally used such technology to showcase their technical capability to customers [45] and provide information about their work without solving any direct business process [88]. A common theme in marketing-related use cases has been using agents as guides to help people learn about any company's business solution.

Despite its applications, there have been several ethical questions around using ECA agents [97], particularly for healthcare and job interview-based applications. An ECA is often prone to design and algorithmic biases. For example, preference for a specific ethnicity or unfair assessment due to any algorithmic error can have serious consequences. Further questions have been raised about these agents' privacy, moral values, and inequitable access. However, recent studies have proposed ethical and practical guidelines [96] for using such agents.

III. Framework of embodied conversational agents

The major components of an Embodied Conversational Agent (ECA) are divided into three stages: Sense, Think, Act [9], [99]. To understand the user's mental state, an agent should first sense various multimodal features using primary sensory inputs such as a camera and a microphone. Then, following the collection of necessary and actionable features (Section - I), an agent should analyze these features based on the current context of the conversation and the user's mental state and generate appropriate multimodal output, i.e., verbal and nonverbal responses via a 3D embodied avatar.

[4]https://www.openstream.ai/gartner-mq/

A. Sense

Analysis of multimodal features associated with a user's behaviour plays an essential role in understanding the mental and emotional state of the user. In applications related to customer service and interview processes, such features provide an essential understanding of the user's satisfaction during the conversation. It also provides information about the user's level of frustration and engagement, which can be used to either change the agent's behaviour or involve a human to continue the conversation. In healthcare, ECAs are used to understand an individual's behaviour. SimSensei [21], for example, is a well-known ECA used for the identification and decision support of post-traumatic stress disorder (PTSD). It deals with sensitive content such as traumatic or stressful events. SimSensei uses MultiSense [79] to track the user's behaviour. MultiSense is a multi-component software; that tracks facial, verbal, and posture-related signals, representing these features using the perception markup language (PML) [70]. The agent then uses this PML representation to take appropriate verbal and nonverbal actions based on well-defined rules that the developers wrote down as per the application. For example, if a user is distracted during a conversation, or if any specific emotion like frustration is detected, then it can prompt the user by either changing its verbal response to encourage engagement or by non-verbal behaviour to facilitate empathy.

PML [70] uses a hierarchical feature representation with two output levels. First, a sensing layer includes features like gaze, head, posture-related visual features, and speech-related features such as speech rate and variation. A behaviour layer consists of features associated with engagement, agreement, anxiety, attention, presence, and smile frequency.

MultiSense employs FACET to track facial action units, OKAO VISION[5] for the smile and Eye Gaze associated features, Microsoft Kinect[6] for body posture related features and CLNF [7] for the head pose detection. Huang et al. [36] developed a more robust alternative to multisense that uses modules like OpenFace 2.0 [8] and Openpose [16] to extract the facial and body signals. Although these modules provide better tracking of behavioural features, they are hard to scale for multi-user platforms. OpenPose [16] is a computationally expensive model and requires a GPU for real-time usage, whereas OpenFace has a costly commercial licence. A possible alternative to OpenFace2.0 [8] is open source software like PyFeat [7] and MediaPipe [8] which provide similar features. Table - I provides a summary.

For audio feature analysis, MultiSense [79] uses the COVAREP plugin [20] to analyze the prosody of the user's audio, while Huang et al. [36] make use of OpenSmile [24].

Another famous platform for developing embodied conversational agents is Agents United [9]. It uses Holistic Behaviour Analysis Framework (HBAF) which combines multimodal

[5]https://plus-sensing.omron.com/technology/index.html
[6]https://azure.microsoft.com/en-us/services/kinect-dk/
[7]https://py-feat.org/pages/intro.html
[8]https://google.github.io/mediapipe/

sensing technologies to analyze users' behaviour and profiling user context. The HBAF works with various data sources, including multimodal devices like smartphones and activity trackers plugins like AWARE[9] or UniversAAL[10]. The HBAF recognizes the user's physical, social, emotional, cognitive, and emotional behaviours such as walking and conversing with others.

Another key component of multimodal analysis is automatic speech recognition (ASR), to transcribe spoken audio input into its textual content. Speech or dialogues drive the overall conversation with the agent and provide the context for non-verbal behaviour generation. Some of the available commercial and open source ASR solutions: Sphinix[11] from CMU and Julius-speech[12] are Hidden Markov Model based ASRs. While Google ASR[13], Microsoft ASR[14], TensorFlowASR[15], DeepSpeech[16], and Kaldi[17] provides Neural-networks based solutions. ASRs have been traditionally developed on HMM and are nowadays designed using DNNs in a quest for better performance. VOSK[18] toolkit gives us both knowledge representation-based and Neural-net-based models.

B. Think

Conversations, non-verbal behaviour, and backchannels are chosen, improved, and given shape in the think stage. This stage makes decisions using the information gathered from the sense stage.

1) Dialogue Management:
The dialogue manager handles the text response generation and the conversation's context to meet the overall application's goal. In the ECA framework, a dialogue manager generates these responses using different natural language processing (NLP) techniques based on predefined rules and models trained on a large dataset.

A system-initiated conversation is a common approach used by the dialogue manager in many commercial applications. In this case, the agent takes the lead and steers the entire conversation, while the user's response is limited to a few phrases, such as saying yes/no or providing specific information. For example, customer care in the call centre converses with customers through predefined scripts. Such conversations are often task-driven to fulfil a specific goal. Some common task-driven dialogue managers are Ravenclaw [11], [12] and Disco [66]. Ravenclaw's dialogue management framework is built on a two-tier hierarchical architecture. The first tier (dialogue task specification) consists of a domain-specific dialogue tree. While the second tier, i.e., the dialogue engine, controls domain-independent aspects of the conversation. On the other hand, Disco [66] uses goals and sub-goals to drive the conversation. However, since the conversation is organized hierarchically, both approaches have problems with the quality of the conversation.

Contrary to a specific task-driven system, some dialogue managers [51] provide much more flexibility in what a user can say during a conversation, and an agent is supposed to generate responses that encourage the user to speak more freely. A system like this has a specific application in interview scenarios [59] or in healthcare for psychological distress management. SimSensei [21] and SimCoach [68] are two famous ECA agents for this task. These are healthcare agents designed to make the user feel comfortable while talking about information related to psychological distress.

Ideally, in a conversation, there is a need to balance the level of the system or user's control over the conversation, especially in marketing-based applications where there is no predefined goal that needs to be achieved. This balance motivates the creation of a mixed-initiative design that has the flexibility to handle both types of conversation. Many task-driven dialogue managers can also be based on mixed-initiative interactions. The newer versions of Ravenclaw attempt to solve mixed-initiative interactions [12] for specific goals.

FloRes [52] uses a similar approach for its dialogue management. It has a forward-looking, reward-seeking dialogue manager that supports a mixed-initiative dialogue. IrisTK [77], a similar approach, uses a state-chart approach. IrisTK [77] provides a multi-party interaction dialogue manager that handles two users at a time; dialogue authoring can be done using an XML-based language called IrisFlow. The dialogue states in IrisTK [77] consists of dialogue acts like AskQuestion or ReqElaborate, and if both users speak together, the system can also allow one user to hold and concentrate on the other user (using ReqHold—ask state). The Avatar's response is also governed by dynamic states like speaking, attending, and listening. However, unlike FloRes [52], IrisTK [77] does not provide a distributed framework for dialogue management, which makes it hard to scale. With a framework resembling that of RavenClaw [11], [12] and IrisTK [77], Flipper2.0 [99] is an open-source dialogue engine from Agents United[19]. Flipper2.0 can handle both straightforward turn-by-turn behaviour and more dynamic turn-taking. The documentation[20] and the research paper both provide quick implementation and prototyping design patterns.

Ultes et al. [85] have even tried to create an end-to-end multi-domain dialogue system that is generic enough to be extended to a new task. One such approach is PyDial [85] which uses different modules like a semantic parser to get syntactic information associated with a sentence, a belief tracker to maintain the internal belief state, a topic tracker to identify the domain of the current input, and a policy based session tracker to map inputs to correct dialogues. These policies are

[9]https://awareframework.com/
[10]https://www.universaal.info/
[11]https://cmusphinx.github.io/wiki/
[12]https://github.com/julius-speech/julius
[13]https://cloud.google.com/speech-to-text
[14]https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/overview
[15]https://github.com/TensorSpeech/TensorFlowASR
[16]https://github.com/mozilla/DeepSpeech
[17]https://kaldi-asr.org/
[18]https://alphacephei.com/vosk
[19]https://www.agents-united.org/
[20]https://github.com/hmi-utwente/flipper-2.0

TABLE I
DETAILS ABOUT MULTIMODAL ANALYSIS TOOLS

Tool	AUs	Landmarks	Expression	Gaze	Head Pose	Skeleton	Realtime Processing on CPU
OKAO	✓	✓	✓	✓	✓	×	✓
OpenPose	×	✓	×	×	✓	✓	×
OpenFace	✓	✓	✓	✓	✓	×	✓
PyFeat	✓	✓	✓	✓	✓	×	×
MediaPipe	×	✓	✓	✓	✓	✓	✓

either hand-crafted for domain-independent sessions or can be learned using Gaussian process-based reinforcement-learning algorithms. The main advantage of such a system is that the same platform can be extended to different use cases [42].

2) Non-verbal behaviour Management:
In many health-related queries or insurance claim-based calls, a customer is often in a distressed emotional state, and an agent who can show empathy and is trustworthy becomes essential in such a conversation. However, creating nonverbal behaviours in an ECA platform [15], [74] presents its own set of difficulties. One of the significant issues faced by the developers is the lack of a standard process across studies to generate these non-verbal behaviours.

To standardize the non-verbal behaviour generation process in ECA platforms, Kopp et al. [39] proposed the SAIBA framework, which subdivides the overall behaviour generation process into three parts: Intent planning is where an agent decides the intent it wants to communicate through non-verbal behaviour like agreement, emphasis, sadness and excitement. The intent is, in many cases, represented using a functional markup language (FML). Behaviour planning, where the intent is represented with appropriate animation as per the posture and body of the avatar; is often represented using a behaviour markup language (BML). In terms of behaviour planning, three significant features of the body need to be controlled: (a) Avatar's facial expression, (b) Avatar's head and body movement, and (c) Lip sync associated with a given verbal sentence.

For facial expression, platforms like Greta [19], SmartBody [82], SimSensei [21], MAX [38], SimCoach [68], SARA [47], and ARIA [86] uses FACS based coding scheme [23] to animate the facial expressions of the avatar in accordance with the intent that needs to be communicated. These platforms allow users to define hand-crafted rules using a set of action units. Contrary to hand-crafted rules, FACSvatar [87] uses a machine learning algorithm to mimic the user's facial expression on the avatar. Researchers have even used the psycho-biological simulation of a human child to create animation for an avatar automatically (BabyX) [98]. Brain Language (BL) is a unique framework used by BabyX for psycho-biologically driven animation representation. Table - II further enlist details about different features in these ECA platforms.

Other common non-verbal behaviours include head movements (nods, tilts), eyebrow movements (raise, lower), eye movements (Gaze, look up, down) and body part movements like hand gestures and shoulder shrugs. These behaviours often accompany verbal speech to either emphasize certain aspects or convey the intent [4]. ECA agents like Furhat [77] use the behaviours and animations associated with head motion and gaze to communicate with two users simultaneously. They propose a priority conflict resolution scheme during the behaviour realization stage. NVBG [41] can also control the avatar's emotional state and corresponding facial expression. However, intent analysis is limited to the semantics available in the surface texts. This limitation is overcome by Cerebella [1] in Simsensei [21]. Unlike NVBG [41], Cerebella [1] uses multimodal information (pre-recorded audio and textual data) associated with the spoken sentence to describe the intent that needs to be communicated. The input (audio and text) is processed in separate pipelines, and the information is maintained in the working memory to derive a spoken sentence's communicative functions or intent. These communicative functions are then mapped to corresponding non-verbal behaviours. Behaviours are scheduled with the start and end times of the word utterances.

Huang et al. [36] have proposed an end-to-end data-driven model for reactive behaviour generation using recurrent neural networks (RNN). The data used to train the model was collected from a dyadic Skype-based conversation between speakers and listeners. The participants' facial expressions are extracted using OpenFace [8], in accordance with FACS [23] and posture information is extracted using OpenPose [16]. The author used OpenSmile [24] to extract the prosody-based features of the user's audio data.

Xu et al., [90] has proposed a rule to generate realistic behaviour using ideational units based on intent planning. The author uses the G-unit to realize natural behaviour, which consists of the preparation phase, stroke, holding, and relaxing phases over multiple coupled gestures. Like this, Matej Rojc et al. [69] proposed an EVA capable of generating non-verbal and verbal co-occurrence behaviours. The generated gestures are not individual animations but rather a sequence of gestures.

3) Backchannel management:
In addition to generating verbal queries or responses for the user, an effective and engaging ECA agent should also generate verbal and nonverbal backchannels during a conversation at appropriate moments. Backchannels are brief visual and verbal cues used in conversation to express interest, attention, and comprehension to the speaker [62]. It ensures that the user is comfortable during the conversation and encourages them to speak freely about a given topic. A few common types of backchannel used during the conversation are as follow [21]:

- **Verbal backchannels**:

TABLE II
DETAILS ABOUT DIFFERENT VERBAL AND NON-VERBAL MODULES USED BY ECA PLATFORM.

Tool	Gesture Control	Body Type	Approach	Interaction Type	Dialogue Control	Design	Approach	Backchannel
VHToolKit	NVBG	Full-Body	Rule-Based	One-One	NPC Editor, Scripting	Single/Distributed	Mixed-Initiative, Data-Driven	Yes
Flipper2.0	-	-	-	-	Information state update	Rules, scripting	Mixed-Initiative, Data-Driven	Yes
IrisTK	Dialogue and Attention Manager	Head and Neck	Rule-Based	One-Two	IrisFlow - Statecharts	Single	Mixed-Initiative, Data-Driven	Yes
RAVENCLAW	-	-	-	-	Dialogue Task Specification	Single/Distributed	Mixed-Initiative, Data-Driven	Yes
Simsensei Kiosk	Cerebella	Upper-Body	Rule-Based	One-One	Flores, Scripting	Single/Distributed	Mixed-Initiative, Data-Driven	Yes
PyDial	-	-	-	-	Ontologies, User simulation	Single	Generative,Data-Driven	No
GECA Framework	RNN	Full-Body	Model-Predictions	One-One	GECA-Dialogue Manager	Distributed	Supports Any	Yes

These are responses used in conversation to depict understanding of the user's response or queries. Eg., "yeah", "uh-huh", "hmm", "right"

- **Non-verbal backchannels**: It involves nodding and smiling at appropriate conversation points to show the user's agreement.

For embodied conversational agents, Dirk Heylen [95] has compiled a list of recent studies that attempts to extend the standard scheme of listening behaviour in face-to-face communication to human agent conversation in ECA platforms.

C. Act

In this stage, the necessary behaviors are realized over the body of the avatar. This module interpret the signals from the thinking stage and use appropriate animation, voice and body of the avatar to realize a given response.

1) Text to speech engine:

Once the dialogue manager has generated a verbal response, the next critical task is to generate the associated audio file. An ideal text-to-speech engine should not only generate audio for the given text but should also provide prosodic variations associated with features like pitch, loudness, articulation rate, and pause to make the generated audio more realistic and human-like.

For this task, Simsensei [21] used a pre-recorded audio of a human actor as the Elle avatar's voice. An actor ensures that a sentence is spoken with natural variation in prosody and different emotional tones so that the user perceives it as more natural. However, such an approach is often limited to scenarios where the avatar's dialogue is predefined and cannot be used in applications that require the generation of new sentences in real-time.

To ensure that associated audio can be generated for any new spoken sentences in run-time scenarios, VHToolkit [32] provides flexibility to use custom text-to-speech (TTS) engines like FESTIVAL[21]. While, IrisTK [77] uses CereVoice system developed by CereProc[22] a cloud-based TTS service. However, such custom voices often lack a human voice's naturalness and prosodic variation.

Recent research has attempted to create speech synthesis with all human-like prosodic features. Models like Tacotron 2 [78] can generate realistic audio output with the prosodic variation of a human voice. Other state-of-art TTS models are Transformer TTS [44], FastSpeech [65].

A custom TTS engine can also be built using open-source toolkits such as ESPNET-TTS [33], which supports cutting-edge models such as Tacotron 2, Transformer TTS, and is easily reproducible.

Some Other commercial cloud based neural TTS engines are Amazon Polly[23], Google text-to-speech[24], Azure Speech services[25] which provides naturalistic sounding TTS services.

2) Lip Sync:

Lip-Sync of an avatar is another critical component in the overall nonverbal behaviour generation process. The phoneme associated with the audio file must be mapped to appropriate visual animations to ensure effective animation generation. While phonemes are single units of sound-related information, Visemes are visual counterparts of these phonemes. Information about the mouth and face position is conveyed via these Visemes. One way to achieve lip-syncing is by effective phoneme detection and mapping it to corresponding Viseme-based animations. Some Phoneme to Viseme maps in literature includes [25], [40].

Researchers have even tried to use end-to-end approaches like JALI [22] and VisemeNet [92] to generate realistic lip-syncing face animation. Various commercial speech engines like Amazon polly[26], Azure Neural TTS[27] also provide phoneme to Viseme maps for different languages.

3) The body or Avatar:

The look of the avatar is an important component of human-avatar communication. However, several issues surrounding the avatar's realism have been discussed [64], [81], with only a few studies supporting its usage [27], [64]. Realistic characters might make people feel uneasy, especially in healthcare settings involving the treatment of post-traumatic stress disorders (PTSD), even if they can foster attractiveness and pleasantness in commercial applications like customer service and marketing [49], [88]. However, research on the effects of realistic avatars on online services has frequently been constrained [64] due to network and technological difficulties.

Researchers have recently developed several platforms for creating animations and associated asset files for low-poly avatars. These low-poly avatars use fewer polygons for the mesh structure of the avatar and support little facial and behavioural details in the animation. LUCIA [43] an open-

[21]https://github.com/festvox/festival/

[22]http://www.cereproc.com/

[23]https://aws.amazon.com/polly/

[24]https://cloud.google.com/text-to-speech/

[25]https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

[26]https://aws.amazon.com/polly/

[27]https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

TABLE III
DETAILS ABOUT THE DIFFERENT EMBODIED AVATARS.

Avatar	Gender	behaviour animation scheduler	Embodiment	Avatar creation platform
LUCIA	Female	XML	Face	Interface
FACSvatar	Custom	Generative data-driven	Face	Makehuman
SmartBody	Custom	BML	Full body	Unity
Avatarsim	Male	Python	Face	Unreal engine
GRETA	Female	BML	Full body	Unity or Ogre
ASAP	Custom	BML	Full body	Unity or Ogre

source 3D platform that allows developers to control the posture and behaviour of the avatar. It can imitate humans by tracking the movements of passive markers placed on a person's face. This process was further extended by FACSvatar [87] by tracking markerless action units detected using OpenFace [8]. Such platforms enable animators to create detailed facial animations using facial mimicry, which would otherwise take a significantly longer time to create. In a parallel direction, FACSHuman [28] created a 3D modelling software to create assets based on the facial action coding system (FACS) [23]. Other platforms like Smartbody [82] enable developers to control and generate seamless non-verbal behaviours and lip sync on the low-poly avatar. Similarly, Aneja et al. [5] proposed Avatarsim an embodied avatar that can produce lip-syncing and facial expressions using FACS and bone position movement via a python interface. Agents United [9] provide two open-source SAIBA-compliant multimodal behaviour realizers GRETA [56] and ASAP [89]; which are compatible with and able to accept as input files in both Function Markup Language and Behavior Markup Language formats. Table - III enlist details about different Avatars.

Numerous 3D platforms have started to provide tools for building realistic avatars in terms of look[28]. However, creating a realistic avatar is time-consuming, expensive, and requires specific artistic skills. Commercial gaming engine companies such as Unreal Engine [29] and Unity[30] have begun to offer readymade assets of realistic-looking high-poly avatars. MetaHuman [31] and facial mimicry tools like facial mocap [32] are two examples of tools that can be used to create high-poly avatars and detailed facial animations associated with them. Even more recent work [94] has shown ways to produce realistically detailed avatars using DeepFake [55] and GAN [29]. However, there is little study on how these new tools and avatar forms may affect the ECA framework.

IV. PLATFORM AND CUSTOMER SERVING

From a commercial standpoint, any ECA platform or application developed to serve customers needs the following requirements:

- **Scalability**: The platform should be scalable enough to serve multiple concurrent users at a given time.
- **Maintenance**: The platform should be easy to maintain and provide flexibility to add new features to the existing architecture.
- **Deployment**: The final application should be deployable to mobile and other low-memory devices.
- **Realism**: The platform should provide flexibility to use a realistic-looking avatar.

A. Scalability

Commercial applications require the capability to handle multiple customers at the same time. However, scaling the modules associated with the ECA platform to handle multiple users is hard. Many limitations come from the high computing requirements of the models used to analyze users' behaviour or generate associated verbal and non-verbal behaviours. Furthermore, these platforms are not flexible enough to handle multiple data streams from different users simultaneously to provide a near real-time experience.

One possible approach to solving the scalability issue is to effectively parallelize the working of individual components in the ECA framework. In this direction, Bohus et al. [13] have proposed a platform for situated intelligence (PSI) which allows developers to use different AI technologies and Azure services for multimodal analysis and synthesis. PSI is an open-source framework that provides real-time temporal data visualization and debugging tools. Bernstein et al. [10] also discussed similar approaches to scaling up the platform.

Researchers have utilized cloud-based platforms and services to scale the platform vertically as per the requirement. SIVA – Socially Intelligent Virtual Agent [6] uses PSI to build a scalable platform. The pipeline makes use of the Microsoft Speech API. CMU-based PocketSphinx[33] to recognise the phonemes and map them to Visemses. The text sentiment is analyzed by feeding the text to Bing Speech API[34] and performing sentiment analysis on the converted text. The conversational style manager supplements the dialogue using intents and responses from the user's utterances via the Language Understanding and Intent Service (LUIS – Azure)[35]. The platform provides a holistic overview of a scalable platform. However, the author mentions the delay in the agent's response as a limitation of the existing platform.

[28]https://zivadynamics.com/

[29]https://www.unrealengine.com/en-US/

[30]https://unity.com/

[31]https://www.unrealengine.com/en-US/metahuman-creator

[32]https://mocap.reallusion.com/iclone-motion-live-mocap/iphone-live-face.html

[33]https://cmusphinx.github.io/wiki/

[34]https://azure.microsoft.com/en-in/pricing/details/cognitive-services/speech-api/

[35]https://azure.microsoft.com/en-us/services/cognitive-services/conversational-language-understanding/

B. Maintenance

To ensure efficient maintenance and scalability of the platform, the developer has to ensure that changes to the existing system do not break the system. One way of doing so is by building modular systems.

As discussed in the previous section (section - III), many existing frameworks follow a modular approach. While VHToolKit [32] uses separate modules for dialogue management, multimodal analysis, non-verbal behaviour generation, and behaviour realization with communication protocols defined using PML, FML and BML scripting language [70]. GECA Framework [35] uses a middleware framework that integrates different ECA components to work as a single unit. The three main parts of the GECA framework are the naming service, message forwarding management, and subscription. Such a modular architecture ensures that only a specific component of the entire architecture is modified per requirement.

C. Deployment

Many classic ECA platforms, like VHToolKit [32] and IrisTK [77] provide a monolithic desktop application, where all the dependencies and modules are packaged into a single desktop application. A user is expected to install the complete software on their local machine. Although such a process prevents any delay associated with network latency or other bandwidth issues, the complete software package size is often in the gigabyte range, which makes it hard to use in low-memory devices like tablets and creates issues associated with its maintenance and updates.

Compared to a desktop-based application, a web-based solution provides greater flexibility regarding its deployment on low-end devices or serving multiple users at a given time. Polceanu et al. [60], and Web-ECA [57] are examples of recent work which provides a web-based solution for ECA agents. Most notably, Polceanu et al. [60] propose EEVA, which operates on three layers. The application layer is a JavaScript mainframe coordinating the multimodal user interface on the client side. Logic-layer provides a state of machine-based logic, and the data layer holds information like phrases and multi-media content. On the other hand, Web-ECA [57] does all the processing on the server, including the coordination of modalities. Such systems are relatively easier to maintain and test as they provide flexibility to change components directly on the server. Also, a web-based platform is easier to use on low-memory devices. However, many of these platforms use the WebGL[36] framework to create and deploy avatars for interaction.

D. Realism

Studies to understand the impact of avatar realism on commercial applications have frequently been limited due to network bandwidth, and other infrastructure-related issues [64]. However, with the recent advancements in gaming engines and deep learning, new solutions have emerged that allow cheaper and more flexible solutions to improve realism.

One of the recent developments for realistic avatar deployment has been the emergence of pixel-streaming service[37]. These platforms allow developers to run a high-quality and realistic-looking avatar created using a gaming engine platform like Unity or Unreal Engine on a cloud-based GPU server and to stream the audio and video frames directly to a local web browser. To ease up the overall deployment process, many companies are offering customized pixel streaming services, such as Furioos[38] and Vagon[39]. In addition, companies like soulMachines [40] and Uneeq [41] further provide solutions to choose from different sets of realistic-looking avatars and control the verbal and non-verbal features using APIs.

Another emerging field is generative model-based realistic avatar creation [29], [55]. Rather than relying on custom game engines, these models use neural rendering approaches to create realistic-looking avatars with similar controls available as gaming counterparts [83], [93].

Developing an embodied conversational agent for commercial applications that can comprehend a user's mental state and generate necessary verbal and nonverbal behaviours is challenging. Such a platform necessitates appropriate behaviour by social norms and requires architectures that facilitate scalability, maintenance, deployment, and realism. Such an approach helps generate high-quality behaviours similar to the most advanced and realistic ECA agents like Simsensei, **EVA**[42] and others over a high poly avatar. Cohen et al. [2] further illustrate other advanced features of the platform.

V. Conclusion

This paper presents different frameworks and architectures for embodied conversational agents(ECA), particularly the platform requirements and studies addressing ECA agents' commercial applications and deployment. In the current information era, many businesses have traditionally relied on face-to-face communication but have now moved onto online platforms to serve a more extensive customer base. In such a scenario, providing the same level of customer service and support traditionally available is essential. An ECA agent can fill in this void associated with the user experience. However, such processes have often been limited because of the issues related to scalability, deployment, maintenance and realism in the existing platform. Through this work, we have tried to provide a holistic overview of the necessary architectures and recent research on platform and customer services.

VI. Acknowledgement

This work was funded by Openstream.ai

[36]https://www.khronos.org/webgl/

[37]https://docs.microsoft.com/en-us/gaming/azure/reference-architectures/unreal-pixel-streaming-in-azure

[38]https://www.furioos.com/

[39]https://vagon.io/

[40]https://www.soulmachines.com/

[41]https://digitalhumans.com/

[42]https://www.openstream.ai/eva/

REFERENCES

[1] Lhommet at al. Gesture with meaning. In *International Workshop on Intelligent Virtual Agents*, pages 303–312. Springer, 2013.

[2] Cohen el al. Commercialization of multimodal systems. *The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions-Volume 3*, pages 621–658, 2019.

[3] Adair et al. The display of "dominant" nonverbal cues in negotiation: The role of culture and gender. *International Negotiation*, 16(3):451–479, 2011.

[4] Alibali et al. Gesture and the process of speech production: We think, therefore we gesture. *Language and cognitive processes*, 15(6):593–613, 2000.

[5] Aneja et al. A high-fidelity open embodied avatar with lip syncing and expression capabilities. In *2019 International Conference on Multimodal Interaction*, pages 69–73, 2019.

[6] Aneja et al. Understanding conversational and expressive style in a multimodal embodied conversational agent. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2021.

[7] Baltrusaitis et al. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 354–361, 2013.

[8] Baltrusaitis et al. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[9] Beinema Tessa et al. *Agents United: An Open Platform for Multi-Agent Conversational Systems*, page 17–24. Association for Computing Machinery, New York, NY, USA, 2021.

[10] Bernstein et al. Orleans: Distributed virtual actors for programmability and scalability. *MSR-TR-2014–41*, 2014.

[11] Bohus et al. Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda. 2003.

[12] Bohus et al. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, 23(3):332–361, 2009.

[13] Bohus et al. Platform for situated intelligence. *arXiv preprint arXiv:2103.15975*, 2021.

[14] Buisine et al. The influence of user's personality and gender on the processing of virtual agents' multimodal behavior. *Adv. Psychol. Res*, 65:1–14, 2009.

[15] Cafaro et al. Nonverbal behavior in. *The Handbook of Multimodal-Multisensor Interfaces, Volume 3: Language Processing, Software, Commercialization, and Emerging Directions*, page 219, 2019.

[16] Cao et al. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[17] Cohn et al. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009.

[18] Craig et al. The impact of virtual human voice on learner trust. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 63, pages 2272–2276. SAGE Publications Sage CA: Los Angeles, CA, 2019.

[19] De Rosis et al. From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. *International journal of human-computer studies*, 59(1-2):81–118, 2003.

[20] Degottex et al. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014.

[21] DeVault et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014.

[22] Edwards et al. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4):1–11, 2016.

[23] Ekman et al. Facial action coding system. 1978.

[24] Eyben et al. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[25] Finn et al. Automatic optically-based recognition of speech. *Pattern Recognition Letters*, 8(3):159–164, 1988.

[26] Flavián et al. The impact of virtual, augmented and mixed reality technologies on the customer experience. *Journal of business research*, 100:547–560, 2019.

[27] Garau et al. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 529–536, 2003.

[28] Gilbert et al. Facshuman a software to create experimental material by modeling 3d facial expression. pages 333–334, 2018.

[29] Goodfellow Ian et al. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[30] Grahe et al. The importance of nonverbal cues in judging rapport. *Journal of Nonverbal behavior*, 23(4):253–269, 1999.

[31] Gratch et al. Virtual rapport. In *International Workshop on Intelligent Virtual Agents*, pages 14–27. Springer, 2006.

[32] Hartholt et al. All together now. In *International Workshop on Intelligent Virtual Agents*, pages 368–381. Springer, 2013.

[33] Hayashi et al. Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 7654–7658. IEEE, 2020.

[34] Hjalmarsson et al. Gaze direction as a back-channel inviting cue in dialogue. In *IVA 2012 workshop on realtime conversational virtual agents*, volume 9. Citeseer, 2012.

[35] Huang et al. The design of a generic framework for integrating eca components. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pages 128–135, 2008.

[36] Huang et al. Development of a platform for rnn driven multimodal interaction with embodied conversational agents. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 200–202, 2019.

[37] Kocaballi et al. The personalization of conversational agents in health care: systematic review. *Journal of medical Internet research*, 21(11):e15360, 2019.

[38] Kopp et al. Max-a multimodal assistant in virtual reality construction. *KI*, 17(4):11, 2003.

[39] Kopp et al. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*, pages 205–217. Springer, 2006.

[40] Kricos et al. Differences in visual intelligibility across talkers. *The Volta Review*, 1982.

[41] Lee et al. Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*, pages 243–255. Springer, 2006.

[42] Lee et al. Dialport, gone live: an update after a year of development. In *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue*, pages 170–173, 2017.

[43] Leone et al. Lucia: An open source 3d expressive avatar for multimodal hmi. In *International Conference on Intelligent Technologies for Interactive Entertainment*, pages 193–202. Springer, 2011.

[44] Li et al. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713, 2019.

[45] Liew et al. Exploring the effects of specialist versus generalist embodied virtual agents in a multi-product category online store. *Telematics and Informatics*, 35(1):122–135, 2018.

[46] Maatman et al. Natural behavior of a listening agent. In *International workshop on intelligent virtual agents*, pages 25–36. Springer, 2005.

[47] Matsuyama et al. Socially-aware animated intelligent personal assistant agent. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*, pages 224–227, 2016.

[48] Matthews et al. Individualised product portrayals in the usability of a 3d embodied conversational agent in an ebanking scenario. In *International Workshop on Intelligent Virtual Agents*, pages 516–517. Springer, 2008.

[49] McBreen et al. Experimental assessment of the effectiveness of synthetic personae for multi-modal e-retail applications. In *Proceedings of the fourth international conference on Autonomous agents*, pages 39–45, 2000.

[50] McNeill et al. Gesture and thought. In *Gesture and Thought*. University of Chicago press, 2008.

[51] Morbini et al. A mixed-initiative conversational dialogue system for healthcare. In *Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue*, pages 137–139, 2012.

[52] Morbini et al. Flores: a forward looking, reward seeking, dialogue manager. In *Natural interaction with robots, knowbots and smartphones*, pages 313–325. Springer, 2014.

[53] Nass et al. Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, 43(2):223–239, 1995.

[54] Nass et al. Can computers be teammates? *International Journal of Human-Computer Studies*, 45(6):669–678, 1996.

[55] Nguyen Thanh Thi et al. Deep learning for deepfakes creation and detection: A survey. *arXiv preprint arXiv:1909.11573*, 2019.

[56] Niewiadomski et al. Greta: an interactive expressive eca system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 1399–1400. Citeseer, 2009.

[57] Nihei et al. Web-eca: A web-based eca platform. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 835–836, 2021.

[58] Olekalns et al. Communication processes and conflict management. 2008.

[59] Pickard et al. Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions? *Computers in Human Behavior*, 65:23–30, 2016.

[60] Polceanu et al. Time to go online! a modular framework for building internet-based socially interactive agents. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 227–229, 2019.

[61] Pooja Rao et al. Automatic follow-up question generation for asynchronous interviews. In *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, pages 10–20, 2020.

[62] Poppe et al. Backchannels: Quantity, type and timing matters. In *International workshop on intelligent virtual agents*, pages 228–239. Springer, 2011.

[63] Provoost et al. Embodied conversational agents in clinical psychology: a scoping review. *Journal of medical Internet research*, 19(5):e6553, 2017.

[64] Qiu et al. Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International journal of human-computer interaction*, 19(1):75–94, 2005.

[65] Ren et al. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32, 2019.

[66] Rich et al. Using collaborative discourse theory to partially automate dialogue tree authoring. In *International conference on intelligent virtual agents*, pages 327–340. Springer, 2012.

[67] Ring et al. Addressing loneliness and isolation in older adults: Proactive affective agents provide better support. In *2013 Humaine Association conference on affective computing and intelligent interaction*, pages 61–66. IEEE, 2013.

[68] Rizzo et al. Simcoach: an intelligent virtual human system for providing healthcare information and support. 2011.

[69] Rojc et al. The tts-driven affective embodied conversational agent eva, based on a novel conversational-behavior generation algorithm. *Engineering Applications of Artificial Intelligence*, 57:80–104, 2017.

[70] Scherer et al. Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In *International Conference on Intelligent Virtual Agents*, pages 455–463. Springer, 2012.

[71] Segedy et al. Supporting student learning using conversational agents in a teachable agent environment. 2012.

[72] Seibt et al. Facial mimicry in its social setting. *Frontiers in psychology*, 6:1122, 2015.

[73] Shubham et al. Conventional and non-conventional job interviewing methods: A comparative study in two countries. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 620–624, 2020.

[74] Shvo et al. 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*, pages 65–72, 2019.

[75] Sidner et al. Creating new technologies for companionable agents to support isolated older adults. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(3):1–27, 2018.

[76] Siegman et al. *Nonverbal behavior and communication*. Psychology Press, 2014.

[77] Skantze et al. Iristk: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 69–76, 2012.

[78] Skerry-Ryan et al. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR, 2018.

[79] Stratou et al. Multisense—context-aware nonverbal behavior analysis framework: A psychological distress use case. *IEEE Transactions on Affective Computing*, 8(2):190–203, 2017.

[80] Straus et al. Does the medium matter? the interaction of task type and technology on group performance and member reactions. *Journal of applied psychology*, 79(1):87, 1994.

[81] Thaler et al. Agent vs. avatar: Comparing embodied conversational agents concerning characteristics of the uncanny valley. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2020.

[82] Thiebaux et al. Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pages 151–158, 2008.

[83] Thies et al. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*, pages 716–731. Springer, 2020.

[84] Tickle-Degnen et al. The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293, 1990.

[85] Ultes et al. Pydial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, 2017.

[86] Valstar et al. Ask alice: an artificial retrieval of information agent. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 419–420, 2016.

[87] Van Der Struijk et al. Facsvatar: An open source modular framework for real-time facs based facial animation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 159–164, 2018.

[88] Vasiljevs et al. Application of virtual agents for delivery of information services. *New Challenges of Economic and Business Development*, pages 702–713, 2017.

[89] Welbergen et al. Asaprealizer 2.0: The next steps in fluent behavior realization for ecas. In *International Conference on Intelligent Virtual Agents*, pages 449–462. Springer, 2014.

[90] Xu et al. Compound gesture generation: A model based on ideational units. In *International Conference on Intelligent Virtual Agents*, pages 477–491. Springer, 2014.

[91] Yacoubi et al. Teatime: A formal model of action tendencies in conversational agents. In *ICAART (2)*, pages 143–153, 2018.

[92] Zhou et al. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 37(4):1–10, 2018.

[93] Zielonka et al. Towards metrical reconstruction of human faces. *arXiv preprint arXiv:2204.06607*, 2022.

[94] Nicole Janette Hertel. *Trust and Behavioral Intention Toward Generative Adversarial Network (GAN)-Derived Avatar Healthcare Provider (HCP) in Simulated Telehealth Setting*. PhD thesis, The Florida State University, 2021.

[95] Dirk Heylen. Multimodal backchannel generation for conversational agents. In *MOG 2007 Workshop on Multimodal Output Generation*, page 81. Citeseer, 2007.

[96] David D Luxton. Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial intelligence in medicine*, 62(1):1–10, 2014.

[97] David D Luxton. Ethical implications of conversational agents in global public health. *Bulletin of the World Health Organization*, 98(4):285, 2020.

[98] Mark Sagar. Babyx. In *ACM SIGGRAPH 2015 Computer Animation Festival*, pages 184–184. 2015.

[99] van Waterschoot et al. Flipper 2.0: A pragmatic dialogue engine for embodied conversational agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 43–50, 2018.