GenECA: A General-Purpose Framework for Real-Time Adaptive Multimodal Embodied Conversational Agents

Santosh Patapati[1]*, Murari Ambati*[1]*, Aashrith Tatineni*[1]*, Trisanth Srinivasan*[1]

[1]Dept. of HCI, Cyrion Labs, United States

santosh@cyrionlabs.org, trisanth@cyrionlabs.org

Abstract

We present GenECA, a general-purpose framework for real-time multimodal interaction with embodied conversational agents. GenECA captures audio and visual signals from standard devices to analyze nonverbal features such as facial expressions, vocal tone, gaze, and posture. This information is used to generate context-aware dialogue and synchronize the agent's speech with dynamic gestures and backchannel facial animations in real time. GenECA provides the first ECA system able to deliver context-aware speech and well-timed animations in real-time without reliance on human operators. Through modular design, it can support a wide variety of applications, such as education, customer service, and therapy.

Index Terms: human-computer interaction, computational paralinguistics, multimodal interaction

1. Introduction

Embodied Conversational Agents (ECAs) are virtual characters that interact through speech and nonverbal cues, and they are increasingly used in a range of fields. Realistic multimodal interaction is critical for ECAs, as appropriate gestures and facial expressions greatly enhance user engagement. However, ECAs have been limited by rigid, scripted behavior. Many systems rely on pre-defined dialogue flows and animations. This restricts naturalness and adaptability in interactions. Although recent advances have enabled more flexible, context-aware dialogue and behavior, seamless real-time integration of these components remains a significant challenge. This has hindered deployment of advanced ECA techniques in interactive settings.

In response to these challenges, we developed GenECA, a unified framework to streamline the creation of ECAs with rich, real-time multimodal interaction. GenECA delivers autonomous interactions with dynamically timed gestures and empathetic speech by combining lightweight multimodal classifiers and Large Language Models (LLMs). All major components can be customized or replaced, allowing researchers to plug in their own models and content. GenECA's flexibility enables rapid development of ECAs for diverse applications.

2. GenECA Framework

GenECA employs a modular pipeline for embodied dialogue (Figure 1), where each component features a well-defined interface that allows independent replacement or modification.

1. **Multimodal Sensing**: GenECA captures audio and visual input in real time using a standard webcam and microphone. The video feed provides a live view of the user's face and upper body, while the audio captures speech. Real-time analysis of these inputs is used to generate backchannel behaviors, and the agent's conversational turn is segmented by detecting each time the user presses and then releases the speaking button.
2. **Real-Time Analysis and Understanding**: The captured multimodal data is immediately processed by GenECA's feature extraction and classification components [1]. By default, the framework uses Mediapipe [2] and OpenFace [3] to track facial landmarks, head pose, eye gaze direction, and body posture in every frame. These low-level signals are fed into lightweight, rule-based classifiers to infer the meaning behind a user's nonverbal cues. For example, consistent gaze aversion or fidgeting may indicate discomfort. On the audio side, the user's speech is processed by a configurable real-time Automatic Speech Recognition (ASR) model of the developer's choice. This produces a text transcription that is analyzed alongside paralinguistic features (tone, pitch, intensity, etc.) by modular classifiers to assess the user's emotional state and engagement. The outputs from both audio and video classifiers are combined into multiple weighted composite scores, each corresponding to a different action trigger. If any of these scores exceed their respective thresholds, the framework selects the appropriate response pathway (e.g., selecting a specific dialogue track or triggering nonverbal behaviors). The final output of this module yields a structured set of information, including the transcribed utterance and any detected conversational cues (e.g., "User exhibits abnormally high movement"). This is provided to the dialogue management module and used to inform the behavior generation module for empathetic gesture selection.
3. **Dialogue Management**: The Dialogue Manager determines the agent's response. It leverages a guided Large Language Model (LLM) interfaced via LangChain for simple modular support with a range of models. In our default configuration, an LLM (LLaMA 3.1 8B) generates the agent's verbal response using a few-shot chain-of-thought prompt template defined by the developer. For each user utterance, the manager constructs a prompt by combining elements of the conversation history (e.g., previously mentioned facts), the new input, and key nonverbal cues (e.g., tone of voice or facial expressions). This prompt may also incorporate system instructions or guardrails to ensure the agent's response is on role and appropriate. The final generated text is passed to the Behavior Generation module for speech synthesis.
4. **Behavior Generation**: Once the agent's next utterance and intent are determined, GenECA generates the behaviors to realize the response. This stage ensures that verbal and nonverbal outputs are synchronized with the intended communicative message.

   The generated text response is fed to a text-to-speech (TTS)

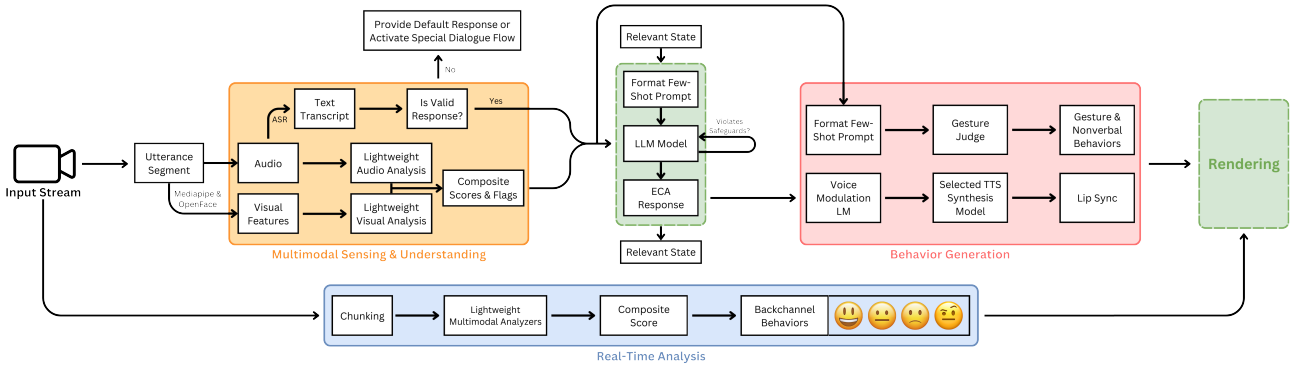

Figure 1: *Diagram of GenECA's real-time interaction pipeline. Each time the user presses and releases the speaking button, an 'Utterance Segment' is recorded and processed. Real-time analysis is performed on short segments of audio captured at each interval.*

model to synthesize the final speech output. Developers can configure different emotion styles for the agent using one of two methods: 1) a lightweight language model selects one of several pre-trained synthesis models to match the desired tone, or 2) a language model with a regression head predicts appropriate valence, arousal, and dominance values for dynamic voice modulation. The first method offers higher quality synthesis at the cost of speed, while the second provides faster but lower quality results.

In parallel, the framework selects nonverbal gestures to complement the verbal response. GenECA supports two approaches: (a) choosing from a library of pre-authored animations using an LLM-based selector, or (b) generating behavior using the Behavior Markup Language [4]. For the pre-authored option, developers supply detailed descriptions of each animation. In our sample implementation (Figure 2), a library of common conversational gestures (e.g., nodding, hand spreads, head tilts) and facial expressions is defined, and a secondary LLM-based classifier acts as a few-shot gesture judge. This classifier reviews the recent dialogue context and the planned utterance to select an appropriate gesture label before the response is delivered.

As the TTS generates speech, the framework performs real-time lip synchronization for the avatar. An efficient and lightweight algorithm analyzes the audio waveform in real-time and maps it to a set of predefined mouth shape blend-shapes using a heuristic based on frequency levels.

Finally, developers may also incorporate idle animations (such as blinking or swaying) and design complex animation cycles using the provided Unity animation controller.

5. **3D Rendering and Output**: In the final stage, GenECA renders the agent's performance and delivers it to the user. The system employs the Unity 3D game engine for real-time animation and rendering. The avatar is a rigged 3D model with predefined articulation points, blend-shapes, and facial animations. Unity receives animation commands from the behavior generation module and applies them to the avatar, while the synthesized speech audio is played concurrently. A queue system ensures that gestures and facial movements are synchronized with the spoken words. As a result, the user experiences a virtual agent that responds naturally, speaking clearly and engaging with coordinated gestures and expressions. Additionally, the Unity-based rendering allows the agent to inhabit different virtual environments as needed.

Each major component of GenECA runs as an independent server on a shared device. These servers are orchestrated via ZeroMQ [5] for direct peer-to-peer communication.

Figure 2: *Sample frame from an interaction between an ECA and a user. The ECA maintains a neutral listening position as the user speaks. Mediapipe tracks real-time visual features.*

2.1. Configuration

Developers can plug in custom models and classifiers, adjust the dialogue policy (e.g., using a conversational flow or dialogue tree), or connect external data sources.

To make GenECA accessible to non-technical users, we are currently developing an early-stage interactive GUI tool which will provide a high-level environment to configure the agent's behavior logic and appearance without writing code. Through the GUI, a developer will be able to map detected user patterns to agent responses or actions, customize the 3D model, define a set of dialogue prompts, develop decision trees, and toggle safeguards. We are designing this interface to lower the barrier to creating specialized ECAs.

3. References

[1] S. V. Patapati, "Integrating large language models into a tri-modal architecture for automated depression classification on the daic-woz," 2024. [Online]. Available: https://arxiv.org/abs/2407.19340

[2] C. Lugaresi *et al.*, "Mediapipe: A framework for building perception pipelines," arXiv preprint, June 2019.

[3] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Openface: An open source facial behavior analysis toolkit," in *IEEE WACV*, 2016.

[4] S. Kopp *et al.*, "Towards a common framework for multimodal generation: The behavior markup language," pp. 205–217, Jan. 2006.

[5] P. Hintjens, *ZeroMQ*. O'Reilly Media, Inc., 2013.