Building an Affective Input Framework for Virtual Humans

1st Cristoper Anderson
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
cristoper.anderson@binus.ac.id

2nd Felix Gozali
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
felix.gozali@binus.ac.id

3rd Andry Chowanda
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
achowanda@binus.edu

4th Jurike Moniaga
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
jurike@binus.edu

***Abstract*—In this paper, we propose a technology framework for human interaction with virtual humans that is more interactive and responsive to users. In the era of advanced technology, virtual humans are one of the areas that continue to grow in the field of human communication and interaction with technology. Although many frameworks are already available, there is still a need to develop a framework that is more natural and responsive in human interaction with virtual humans.**

***Index Terms*—Interactions; Input Signaling; Virtual Human; Artificial Intelligence;**

I. Introduction

In the modern world, virtual humans are being used more often, for example, virtual humans can be used to train people for job interviews [1], or the technology could even be used to fill a virtual world with virtual humans [2]. We can also see that papers that involve the creation of applications relating to IVA (Intelligent Virtual Agents) are becoming more prominent [3]. In order to effectively engage with users, virtual humans necessitate the fundamental ability to accurately understand the nuanced context underlying their conversations. To achieve this, virtual humans are equipped with an array of advanced capabilities, one of which notably encompasses their aptitude in skillfully processing and comprehending user input, particularly in the domain of natural language. By harnessing this ability, virtual humans can discern and interpret the meaning embedded within the content shared by users, facilitating more meaningful and contextually relevant interactions [4][5][6].

The main challenge in creating a virtual human is input signaling, where if done incorrectly, may frustrate the user while interacting with the virtual human. To improve the experience of users when using virtual humans, more research and development in this field is required. This research paper's goal is to explore input signaling for virtual humans, and to identify strategies to increase interaction. Doing this will increase virtual humans' capabilities in many area, and will have large impact in our society. To reach this purpose, we will examine the current condition in input signaling, and evaluate some technologies used in this field, such as, speech recognition, face recognition, and natural language processing. By evaluating these technologies, we hope to identify the most effective technologies to improve input signaling and improve the overall user experience.

The final goal of our research is to significantly contribute to the development of the technology of virtual humans, by providing knowledge and recommendations about how to improve input signaling and to create more interesting and effective interactions.

II. Recent Work

A. Virtual Human Framework

Research for Intelligent Virtual Agents (IVA) is a mix of different fields of science that use many advanced technologies. With the rate of growth of technology, it is considered difficult to intergrate these new technologies into a cohesive framework. Tools that help create IVAs emerge because of this. An example of a tool that has been built for this purpose are the Virtual Human Framework (VHToolkit) [7], that have a built in audio-visual detection, natural language processing, non-verbal behavior generation, and text-to-speech. Recently VHToolkit has been integrated with a tool called RIDE (Rapid Integration & Development Environment) which allows users to rapidly create digital worlds with virtual humans [2]. Another example of recent work made in this field is PRIMER, which is an emotionally aware virtual agent, that respond to direct virtual input and indirect emotional input. PRIMER acheives this by tracking the user's emotions through the user's text input [8]. Another example of work in this field is the SEMAINE (Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression) API, which aims to create a multimodal dialogue system with an emphasis on

nonverbal capabilities, such as detecting faces and producing sounds related to interactions with the user [9]. In the field of virtual humans, there is also something called the uncanny valley, where mismatches in realism of a virtual human's voice can result in the virtual human seeming unnatural, so a study is done to determine the optimal levels of realism in the virtual human's speech, appearance, and motion. The study recommends that for voice and motion, higher levels of realism should be used, and for visual appearance, a lower level of realism would be preferable [10]. Another example of work in the field of virtual humans is it's usage in training clinicians to understand their patient's issues and feelings, by developing a web-based training simulation, where trainee clinicians can train their empathic communication skills with a virtual human [11].

B. Speech Recognition

Speech recognition is a field of science which combines computer science and linguistics, where the goal is the understanding of human speech. Speech recognition is usually done using a device that can record human speech, and then transform that analog signal into a digital one, where that signal can be processed using a neural network, until that data is classified by that neural network [12]. The accuracy of speech recognition has been improved since deep neural network was adopted [13] . Some of the techniques that is used for speech recognition are CTC (Connectionist Temporal Classification) where the input sentences will be mapped into an output, and RNN transducer, which gives us a natural way of streaming speech recognition, where the current output depends on the previous one. A recent implementation of speech recognition is the Virtual Human Framework [7].

C. Facial Expression Recognition

Emotion recognition can be obtained through various methods, including facial expressions, speech, EEG, and even text. Among these methods, facial expressions emerge as the most popular choice due to several advantages[14]. Facial expressions are readily visible and provide a wealth of cues that are highly informative for emotion recognition. Moreover, assessing someone's emotions based on their facial expressions is comparatively straightforward, making it a practical and widely applicable approach in various settings[15].

D. Affect Recognition

Affect recognition's goal is the understanding of emotions of humans. Computers usually use machine learning to understand human emotions, but machine learning is not the most efficient way for a computer to understand human emotions. Hence, researchers have moved on from using machine learning in affect recognition, and started using deep learning techniques instead [16][17][18]. A subcategory of affect recognition, which is context-aware emotion recognition doesn't only focus on the user's face or body, but also uses the background context of the user [19]. Affect recognition can be done in multiple ways, such as facial expression recognition, which a recent work does by using EmotionAPI emotion detector which uses a CNN (Convolutional Neural Network) which is based on VGG-13 [20][21], and or by analyzing the voice by analyzing the spoken language itself (Language Sentinment Analysis) by using a classifier [20], or by analyzing the tones of the spoken voice, which a recent work does by using a deep neural network [20], or by analyzing the subject's respiration rate by using ECG (Electrocardiography) or even by manually measuring it using a timer. The output of affect recognition is emotion, which are usually measured by two values, which are valence and arousal. Below is a diagram showing the mapping of valence and arousal values to emotions [22]. Some of the challenges

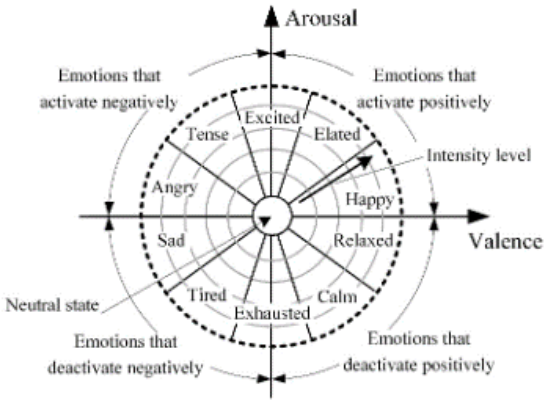

Fig. 1. Diagram of the Russel emotional model.

in affect recognition are odd/incorrect grammar, a lack of labeled datasets, the growing lexicon of the internet, detection of sarcasm, and the detection of multiple emotions in a single sentence [23].

E. Dialogue System

Dialogue system is a technology that allows users to converse with a computer system using natural language [24]. In the paper, the dialogue system consists of two components, namely natural language generation and natural language understanding. Natural language understanding is a technique in the dialogue system that is used to analyze user input and extract relevant information from it so that the virtual human can respond appropriately. Natural language generation is used to make the virtual human generate a response to the user. Natural language generation is usually done by using a GPT (Generative Pre-Trained Transformer) language model [25].An example of a recent engine for integration with virtual humans is Flipper 2.0 [26]. Another example of recent work in this field is VoxWorld, which is a multimodal dialogue system which allows users to communicate using language, gestures, facial expressions, and even gaze tracking [27].

F. SAIBA Framework

The SAIBA (Situation, Agent, Intention, Behavior, Animation) Framework is a model for generating behavior, distinguishing between communicative functions and behavior itself.

The SAIBA Framework was specifically designed to generate behavior for ECA (Embodied Conversational Agents), which are commonly referred to as social robots or robots capable of socializing. This framework divides the process of generating behavior into three smaller processes. The first process is intent planning, which provides input for the second process, behavior planning, resulting in a behavior. The final process involves the realization of this behavior. Some examples of the implementation of this framework are ERiSA, which aims to create an Intelligent Virtual Agent within a game (game agent) [28], PSA (Passive Sensing Framework) which is a virtual human coach, that collects data about the user's health using wearable gadgets, such as smartwatches [29], and FANTASIA, which is a framework that integrates a dialogue system, a graph database, a game engine, and a voice synthesis engine, to make the process of design and implementation of human interactive applications faster, for HCI (Human Computer Interaction) studies[30].

III. Research Methodology

A. Systematic Literature Review

The first phase of our research methodology involves conducting a systematic literature review. We gather relevant papers and articles related to our research topic and analyze them to draw meaningful conclusions. This review helps us establish a foundation of knowledge and identify gaps in existing research. Then comes developing the framework itself.

B. Research Question

Our research question is focused on two key aspects of developing and improving affection input framework for virtual humans:

RQ1. How effective is the proposed architecture in facial emotion recognition, combining face detection, extraction, and classification, and how does it enhance the Virtual Human's ability to comprehend and respond to user emotions?

RQ2. How well does the integration of data preprocessing, emotion classification, and the use of Convolutional Neural Networks (CNNs) perform in speech emotion recognition and understanding the content of speech?

By addressing these research questions, we aim to explore the integration of various technologies and methodologies to enable more effective communication and interaction between virtual humans and users. Additionally, we seek to investigate the ways in which the framework can be further developed and refined to improve its overall performance and user experience.

C. Literature Selection

During the literature selection process, we employ a systematic approach to identify and include relevant studies, papers, and articles that contribute to our research topic. The following criteria guide our selection:

1) Relevance: We prioritize literature that specifically addresses the integration of voice recognition, speech synthesis, natural language understanding, and dialogue systems in the context of creating an interactive input signaling framework for virtual humans.
2) Quality and Credibility: We select sources from reputable journals, conferences, and scholarly publications to ensure the reliability and validity of the information presented.

By adhering to these criteria, we ensure that our research is built upon a robust foundation of existing knowledge and expertise. This approach allows us to identify existing advancements, address gaps in the literature, and contribute to the further development and improvement of the input signaling framework for virtual humans.

IV. Proposed Architecture

The proposed input component comprises two elements: the user's face and speech. The user's face undergoes a three-step process involving face detection, face extraction, and face classification. Additionally, we collect two data points from the user's speech: the tone of speech and the message content of speech. Utilizing deep learning models, we analyze these inputs to predict the user's emotions and intentions. Finally, based on this analysis, Virtual human will generate a response for the user.

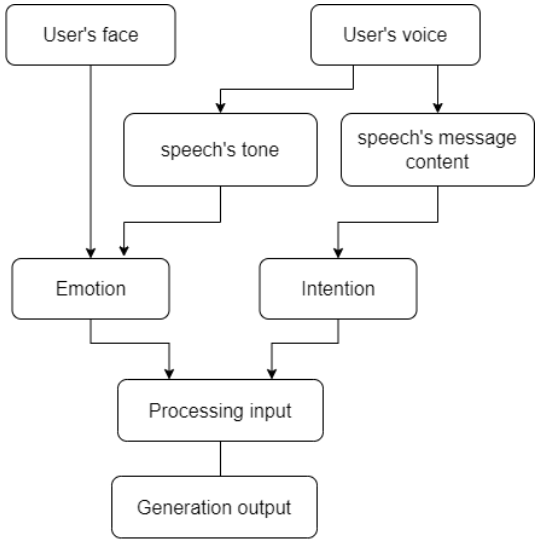

Fig. 2. Proposed Architect.

V. Result And Discussion

A. Facial Emotion Recognition

In the proposed input component, facial emotion recognition plays a crucial role in understanding the user's emotional state based on their facial expressions. The process involves three main steps:

1) Face Detection: The initial step is to detect the presence of a face within the input image or video frame. For this

purpose, we recommend utilizing using Haar cascades. Haar cascades offer efficient and accurate face detection capabilities, enabling subsequent processes to concentrate on relevant facial regions

2) Face Extraction: Once the face is detected, we extract the facial region and resize the image based on an established algorithm. This step is crucial for isolating the facial features effectively and ensuring consistency for further analysis.
3) Face Classification: In the final step, we employ pre-trained Convolutional Neural Networks (CNNs) for the task of facial emotion classification. These pre-trained CNN models have been extensively trained on large emotion-labeled datasets, enabling them to learn discriminative features crucial for accurate emotion recognition.

By implementing the recommended Haar cascades for face detection, we achieve reliable and accurate localization of facial features. Furthermore, the utilization of custom CNNs for face classification enhances the system's ability to discern emotions, accurately identifying expressions like happiness, sadness, anger, fear, and more. This integrated approach results in a robust and effective Facial Emotion Recognition system, empowering the Virtual Human to comprehend and respond to the user's emotions in a sophisticated and natural manner.

B. Speech Emotion Recognition

In the proposed input component, Speech Emotion Recognition (SER) serves as a critical component in understanding the user's emotional state based on their speech patterns. The process involves two key steps, each contributing to a comprehensive and accurate emotion recognition system. The process involves two main steps:

1) Data Preprocessing: To build a robust SER system, we start by collecting audio samples from users, capturing a diverse range of emotional expressions. The audio data is then segmented into smaller frames, typically lasting between 20ms to 50ms, effectively capturing dynamic variations in speech that reflect various emotional states. We extract meaningful features from the raw audio, including Mel-Frequency Cepstral Coefficients (MFCCs) to represent the spectral characteristics of speech, and prosody features such as pitch, energy, and duration that provide valuable emotional content information. Additionally, we utilize Automatic Speech Recognition (ASR) to transcribe the speech content, converting audio signals into textual representations of the speech's message content. This textual information complements the acoustic features, providing a holistic view of the emotions conveyed through speech.
2) Emotion Classification:item *Emotion Classification with Neural Networks:* The extracted MFCCs, prosody features, and transcribed textual content are combined to create a unified representation for each speech segment, forming the input data for the emotion classification phase. For this classification, we employ a combination of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). RNNs effectively model sequential data and capture temporal dependencies in speech, facilitating a contextual understanding of emotional nuances expressed over time. At the same time, CNNs capture local patterns in the speech data, learning hierarchical features that identify salient emotional patterns in short segments of the audio frames and text. The integration of these neural network architectures results in a comprehensive and accurate emotion recognition model.

VI. Conclusion

In this research, we addressed two critical research questions to enhance emotion recognition in our proposed input component. Firstly, we investigated the effectiveness of various algorithms in both facial emotion recognition and speech emotion recognition. Secondly, we explored which algorithm proves to be more effective in understanding the speech's content.

Our proposed architecture consists of two main components: facial emotion recognition and speech emotion recognition. For facial emotion recognition, we employed a three-step process involving face detection, face extraction, and face classification. Utilizing Haar cascades for face detection and custom Convolutional Neural Networks (CNNs) for face classification, we achieved accurate emotion recognition by identifying various expressions, such as happiness, sadness, anger, and fear. This integrated approach resulted in a robust Facial Emotion Recognition system, empowering the Virtual Human to comprehend and respond to the user's emotions in a sophisticated and natural manner.

For speech emotion recognition, we introduced a two-step process comprising data preprocessing and emotion classification. By collecting users' audio samples, segmenting them into smaller frames, and extracting relevant features like Mel-Frequency Cepstral Coefficients (MFCCs) and prosody features, we obtained valuable information about emotional content in speech. Utilizing a combination of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) for emotion classification, our Speech Emotion Recognition system effectively captured temporal dependencies and local patterns in speech data, leading to comprehensive emotion recognition results.

References

[1] A. Hartholt, S. Mozgai, A. S. Rizzo, Virtual job interviewing practice for high-anxiety populations, Association for Computing Machinery (ACM), 2019, pp. 238–240. doi:10.1145/3308532.3329417.

[2] A. Hartholt, S. Mozgai, Creating virtual worlds with the virtual human toolkit and the rapid integration & development environment, Vol. 69, AHFE International, 2023. doi:10.54941/ahfe1002856.

[3] N. Norouzi, M. Lee, K. Kim, S. Daher, G. Welch, J. Hochreiter, G. Bruder, A systematic survey of 15 years of user studies published in the intelligent virtual agents conference, Association for Computing Machinery, Inc, 2018, pp. 17–22. doi:10.1145/3267851.3267901.

[4] A. Galassi, M. Lippi, P. Torroni, Attention in natural language processing, IEEE Transactions on Neural Networks and Learning Systems 32 (2021) 4291–4308. doi:10.1109/TNNLS.2020.3019893.

[5] S. Ruder, M. Peters, S. Swayamdipta, T. Wolf, Transfer learning in natural language processing tutorial (2019).

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. V. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing (2021).
URL https://github.com/huggingface/

[7] A. Hartholt, E. Fast, Z. Li, K. Kim, A. Leeds, S. Mozgai, Re-architecting the virtual human toolkit: Towards an interoperable platform for embodied conversational agent research and development, Association for Computing Machinery, Inc, 2022. doi:10.1145/3514197.3549671.

[8] C. Gordon, A. Leuski, G. Benn, E. Klassen, E. Fast, M. Liewer, A. Hartholt, D. Traum, Primer: An emotionally aware virtual agent acm reference format (2019).

[9] M. Schröder, The semaine api: Towards a standards-based framework for building emotion-oriented systems, Advances in Human-Computer Interaction 2010 (2010). doi:10.1155/2010/319406.

[10] Y. Ferstl, S. Thomas, C. Guiard, C. Ennis, R. McDonnell, Human or robot?: Investigating voice, appearance and gesture motion realism of conversational social agents, Association for Computing Machinery, Inc, 2021, pp. 76–83. doi:10.1145/3472306.3478338.

[11] H. Yao, A. G. D. Siqueira, A. Bafna, D. Peterkin, J. Richards, M. L. Rogers, A. Foster, I. Galynker, B. Lok, A virtual human interaction using scaffolded ping-pong feedback for healthcare learners to practice empathy skills, Association for Computing Machinery, Inc, 2022. doi:10.1145/3514197.3549621.

[12] V. Vashisht, A. K. Pandey, S. P. Yadav, Speech recognition using machine learning, IEIE Transactions on Smart Processing and Computing 10 (2021) 233–239. doi:10.5573/IEIESPC.2021.10.3.233.

[13] J. Li, X. Chen, Y. Gaur, Y. He, Y. Huang, N. Kanda, Z. Meng, Y. Shi, E. Sun, X. Wang, Y. Wu, G. Ye, R. Zhao, J. Li, Recent advances in end-to-end automatic speech recognition, APSIPA Transactions on Signal and Information Processing 11 (2022) 8. doi:10.1561/116.00000050_supp.
URL http://dx.doi.org/10.1561/116.00000050_supp

[14] S. Minaee, M. Minaei, A. Abdolrashidi, Deep-emotion: Facial expression recognition using attentional convolutional network, Sensors 21 (5 2021). doi:10.3390/s21093046.

[15] S. Zhang, X. Pan, Y. Cui, X. Zhao, L. Liu, Learning affective video features for facial expression recognition via hybrid deep learning, IEEE Access 7 (2019) 32297–32304. doi:10.1109/ACCESS.2019.2901521.

[16] P. V. Rouast, M. T. Adam, R. Chiong, Deep learning for human affect recognition: Insights and new developments, IEEE Transactions on Affective Computing 12 (2021) 524–543. doi:10.1109/TAFFC.2018.2890471.

[17] Mustaqeem, S. Kwon, A cnn-assisted enhanced audio signal processing for speech emotion recognition, Sensors (Switzerland) 20 (1 2020). doi:10.3390/s20010183.

[18] D. Issa, M. F. Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomedical Signal Processing and Control 59 (5 2020). doi:10.1016/j.bspc.2020.101894.

[19] D. Yang, Z. Chen, Y. Wang, S. Wang, M. Li, S. Liu, X. Zhao, S. Huang, Z. Dong, P. Zhai, L. Zhang, Context de-confounded emotion recognition.

[20] D. McDuff, K. Rowan, P. Choudhury, J. Wolk, T. Pham, M. Czerwinski, A multimodal emotion sensing platform for building emotion-aware applications (3 2019).
URL http://arxiv.org/abs/1903.12133

[21] B. J. Abbaschian, D. Sierra-Sosa, A. Elmaghraby, Deep learning techniques for speech emotion recognition, from databases to models (2 2021). doi:10.3390/s21041249.

[22] A. Dzedzickis, A. Kaklauskas, V. Bucinskas, Human emotion recognition: Review of sensors and methods (2 2020). doi:10.3390/s20030592.

[23] P. Nandwani, R. Verma, A review on sentiment analysis and emotion detection from text (12 2021). doi:10.1007/s13278-021-00776-6.

[24] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, M. Cieliebak, Survey on evaluation methods for dialogue systems, Artificial Intelligence Review 54 (2021) 755–810. doi:10.1007/s10462-020-09866-x.

[25] D. W. Otter, J. R. Medina, J. K. Kalita, A survey of the usages of deep learning in natural language processing (7 2018).
URL http://arxiv.org/abs/1807.10854

[26] J. V. Waterschoot, D. Reidsma, M. Bruijnes, D. Davison, D. Heylen, J. Flokstra, M. Theune, Flipper 2.0 a pragmatic dialogue engine for embodied conversational agents, Association for Computing Machinery, Inc, 2018, pp. 43–50. doi:10.1145/3267851.3267882.

[27] J. Pustejovsky, N. Krishnaswamy, Embodied human-computer interactions through situated grounding, Association for Computing Machinery, Inc, 2020. doi:10.1145/3383652.3423910.

[28] A. Chowanda, P. Blanchfield, M. Flintham, M. Valstar, Erisa: Building emotionally realistic social game-agents companions.

[29] S. Mozgai, A. Hartholt, A. Rizzo, The passive sensing agent: A multimodal adaptive mhealth application, in: 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 2020, pp. 1–3. doi:10.1109/PerComWorkshops48775.2020.9156177.

[30] A. Origlia, F. Cutugno, A. Rodà, P. Cosi, C. Zmarich, Fantasia: a framework for advanced natural tools and applications in social, interactive approaches, Multimedia Tools and Applications 78 (2019) 13613–13648. doi:10.1007/s11042-019-7362-5.