Ph.D. Thesis

Scalable Hybrid Approach of Co-speech Text-to-Gesture Generation for Interactive Digital Humans

Ghazanfar Ali

AI Robotics

Korea Institute of Science and Technology (KIST) School

UNIVERSITY OF SCIENCE AND TECHNOLOGY

August 2023

Scalable Hybrid Approach of Co-speech Text-to-Gesture Generation for Interactive Digital Humans	**2023**	**Ghazanfar Ali**

Scalable Hybrid Approach of Co-speech Text-to-Gesture Generation for Interactive Digital Humans

Advisor: Dr. Jae-In Hwang

A Dissertation Submitted in Partial Fulfillment of Requirements
For the Degree of Ph.D.

August 2023

UNIVERSITY OF SCIENCE AND TECHNOLOGY
AI Robotics

Ghazanfar Ali

We hereby approve the Ph.D. thesis of "Ghazanfar Ali".

June 2023

Dr. Sang Chul Ahn
Chairman of Thesis Committee

Dr. Jae-In Hwang
Thesis Committee Member

Dr. Hwasup Lim
Thesis Committee Member

Dr. Suhyun Kim
Thesis Committee Member

Dr. Junho Kim
Thesis Committee Member

UNIVERSITY OF SCIENCE AND TECHNOLOGY

Acknowledgements

In the name of Allah, the Most Merciful, and peace be upon Prophet Muhammad (P.B.U.H), my journey to the culmination of my Ph.D. has reached its terminus. This voyage, steeped in personal and intellectual growth, would have remained a distant dream without the unwavering support of many cherished individuals.

I am forever indebted to my Grandfather, Abdul Sattar, and my Grandmother, Saleema Bibi, who nurtured my dreams and saw in me a potential that inspired me to strive for excellence. My parents, Iftikhar-ul-Haq and Farzana Kausar, who gave selflessly and loved endlessly, instilled in me a passion for learning and a dedication to hard work. I owe my accomplishments to their faith in me.

My heartfelt gratitude goes to my Uncle Zulfiqar Ahmed and Aunt Shehnaz who welcomed me into their home during my years in college and university, ensuring I was never away from family. The warmth and love they bestowed upon me served as a beacon during the challenging times.

My younger brother, Shahid Iftikhar, stood by me, providing support and sanctuary when I moved to Lahore, his unwavering belief in me was my motivation. My wife, Anam Shamhad, has been my stronghold, her resilience, and support are pivotal to my Ph.D. journey. My sons, Musa Ali and Abubakr Ali, whose innocent smiles have given me joy and inspiration.

I would like to extend a special thanks to Habibullah Khan, CEO of Digitalpenumbra. His financial assistance during the preliminary stages of my journey came when I was broke, for which I am eternally grateful. I thank my friend and classmate, Azeem Ahmad, for his recommendations and invaluable guidance during the admission process.

I also wish to thank my friend, Asif Mahmood, who was there at Incheon Airport to receive me and hosted me for many weekends to come.

My profound gratitude goes to Dr. Hwang Jae-In, an exemplary supervisor whose mentorship, patience, and unwavering support have been instrumental in bringing me to this pivotal moment in my academic journey. My lab mates, Le Hong Quan, Muhammad Firdaus Lubis, Hanseob Kim, Binhan, and HwangYoun Kim, deserve my sincerest thanks for their assistance during experiments and tasks integral to my research and graduation.

Thanks to my Pakistani community at KIST for their support and for creating loving memories.

To all my teachers, friends and family members I thank you all for being there for me.

To all, your contributions have made this journey not just bearable but enjoyable. You've ensured that I graduate not just with a degree, but with happiness and the best mental health. For this, I am forever indebted.

Abstract

This thesis chronicles developing and optimizing co-speech text-to-gesture generation for digital humans, underlining a significant leap towards enhancing user engagement and interaction in wearable mixed reality environments. The research narrative begins with the conception of a multimodal interaction framework for intelligent virtual agents and concludes with the advanced co-speech gesture generation system, ConGRets, capable of zero-shot speaker style adaptation.

The research began with creating an intelligent virtual agent system for mixed-reality environments. This system integrated features such as speech recognition, domain-specific chatbot capabilities, and manual mapping of body animations to speech content, highlighting the potential of digital humans as interactive tools. However, it also emphasized the need for a more automated and expressive body gesture generation system to enhance communication and engagement further.

Motivated by this requirement, the research introduced an automated, rule-based co-speech gesture mapping system. This system, leveraging large publicly available video datasets and word embeddings, generated a diverse and accurate set of gestures that bypassed the limitations of traditional human-created mappings.

A significant advancement was marked by the introduction of GestureCLR, a contrastive learning model that drastically improved the matching of 2D poses from videos to 3D gesture units, outperforming the previous mean cosine similarity method. The integration of GestureCLR and K-Means clustering on gesture units derived from English monologue videos and Korean speaker motion capture sequences led to the development of a cross-lingual method. This method effectively generated realistic, human-like gestures for digital humans, augmenting their non-verbal communication capabil-

ities across different languages without compromising user experience.

The final phase of the research introduced ConGRets, a unique framework optimizing real-time performance and resource usage. Addressing the limitations of previous methods, such as slow query speeds and averaged gesture generation, ConGRets incorporated zero-shot speaker style adaptation. This novel feature allowed the system to generate gestures in alignment with the text input and the individual speaker's motion style.

The thesis narrates the journey of this groundbreaking research, which is now permeating diverse applications, including medical screening kiosks, the virtual therapy system, the actor-focused pre-visualization system ASAP, the ChatGPT frontend, and the virtual presenter. This comprehensive exploration of the evolution of co-speech text-to-gesture generation for digital humans underscores a significant stride towards enriching their non-verbal communication abilities and enhancing user interaction and engagement in various environments.

국문 초록

이 논문은 웨어러블 혼합 현실 환경에서 사용자 참여 및 상호 작용을 향상시키는 중요한 도약을 강조하면서 디지털 휴먼을 위한 공동 음성 텍스트-제스처 생성을 개발하고 최적화하는 과정을 연대순으로 설명합니다. 연구 내러티브는 지능형 가상 에이전트를 위한 다중 모드 상호 작용 프레임워크의 개념으로 시작하여 제로 샷 화자 스타일 적응이 가능한 고급 공동 음성 제스처 생성 시스템인 ConGRets로 끝납니다.

이 연구는 혼합 현실 환경을 위한 지능형 가상 에이전트 시스템을 만드는 것으로 시작되었습니다. 이 시스템은 음성 인식, 도메인별 챗봇 기능, 음성 콘텐츠에 대한 신체 애니메이션 수동 매핑과 같은 기능을 통합하여 대화형 도구로서의 디지털 휴먼의 잠재력을 강조했습니다. 그러나 의사소통과 참여를 더욱 향상시키기 위해 보다 자동화되고 표현력이 풍부한 신체 제스처 생성 시스템의 필요성도 강조했습니다.

이 요구 사항에 동기를 부여하여 연구는 자동화된 규칙 기반 공동 음성 제스처 매핑 시스템을 도입했습니다. 이 시스템은 공개적으로 사용 가능한 대규모 비디오 데이터 세트와 단어 임베딩을 활용하여 사람이 만든 전통적인 매핑의 한계를 뛰어넘는 다양하고 정확한 제스처 세트를 생성했습니다.

비디오에서 3D 제스처 단위로의 2D 포즈 매칭을 대폭 개선하여 이전의 평균 코사인 유사성 방법을 능가하는 대조 학습 모델인 GestureCLR의 도입으로 상당한 발전이 이루어졌습니다. 영어 독백 비디오와 한국어 화자 모션 캡처 시퀀스에서 파생된 제스처 단위에 대한 GestureCLR 및 K-Means 클러스터링의 통합은 교차 언어 방법의 개발로 이어졌습니다. 이 방법은 디지털 휴먼을 위한 사실적이고 인간과 같은 제스처를 효과적으로 생성하여 사용자 경험을 손상시키지 않으면서 다양한 언어에 걸쳐 비언어적 커뮤니케이션 기능을 강화했습니다.

연구의 마지막 단계에서는 실시간 성능 및 리소스 사용을 최적화하는 고유한 프레임워크인 ConGRets를 도입했습니다. 느린 쿼리 속도 및 평균 제스처 생성과 같은 이전 방법의 한계를 해결하기 위해 ConGRets는 제로 샷 스피커 스타일 적응을 통합했습니다. 이 새로운 기능을 통해 시스템은 텍스트 입력 및 개별 화자의 동작 스타일에 맞춰 제스처를 생성할 수 있었습니다.

이 논문은 의료 검진 키오스크, 가상 치료 시스템, 배우 중심의 사전 시각화 시스템 ASAP, ChatGPT 프런트엔드 및 가상 발표자를 포함한 다양한 응용 프로그램에 현재 침투하고 있는 이 획기적인 연구의 여정을 설명합니다. 디지털 휴먼을 위한 공동 음성 텍스트-제스처 생성의 진화에 대한 이 포괄적인 탐구는 비언어적 의사 소통 능력을 강화하고 다양한 환경에서 사용자 상호 작용 및 참여를 향상시키는 데 상당한 진전이 있음을 강조합니다.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

A transformative narrative unfolds in the expansive theater of technological innovation, where digital humans have begun to take center stage. This thesis, ”Co-Speech Text-To-Gesture Generation For Digital Humans,” serves as a script, a meticulously detailed account of our journey to breathe new life into these digital entities.

We began our voyage with a formidable challenge: to endow an intelligent virtual agent, operating within the immersive domain of mixed-reality environments, with a more human-like ability to communicate. A lofty goal, indeed, but one we approached with the conviction of explorers embarking on a mission into uncharted territories.

The initial phase saw the integration of speech recognition, domain-specific chatbot capabilities, and manual mapping of body animations to speech content. Although this was a promising start, it was akin to a digital human learning its first words - significant, but the journey was far from complete. The next step was clear: our digital humans needed to express themselves through body language, the unspoken dialect that so often carries the heart of our communication.

With a pioneering spirit, we set forth on this new chapter. Rule-based co-speech

gesture mapping systems, contrastive learning models, and finally, the advent of ConGRets, our cutting-edge gesture generation system. Each breakthrough brought our digital humans closer to mirroring the complexity and richness of human communication.

From the initial steps to the final triumph of ConGRets, this thesis represents a chronicle of relentless innovation and rigorous exploration. It is an engaging tale of progress that marks a substantial stride towards enhancing user interaction and engagement with digital humans in various environments. So, let's delve into this journey and explore the extraordinary world of co-speech text-to-gesture generation.

1.1 Decoding the Title

In our "Co-Speech Text-to-Gesture Generation For Digital Humans" thesis title, each term carries substantial weight, contributing to a comprehensive understanding of our research focus.

Co-Speech Gestures: In the broad spectrum of gestures, co-speech gestures hold a specific place. These are physical movements typically made with hands or arms synchronized with speech. For example, akin to describing a large balloon while simultaneously using hands to demonstrate its size. This differs from gestures that a soccer referee employs, which are typically standalone and do not accompany speech.

Text-to-Gesture: This portion of the title represents our focus on interactive applications where text serves as the primary input. Given the occasional absence or complexity of acquiring audio input, we aim to convert text into corresponding gestures, much like translating a written command into a gestural indication in a game of charades.

Digital Humans: As the final component of our title, 'Digital Humans' refers to virtual avatars that are lifelike and engaging. These digital entities serve as mediums

of information dissemination, and through the generation of co-speech gestures, their communication can be enhanced, thus enriching user interaction.

In essence, our thesis aims to bridge the textual input and the non-verbal cues of digital human avatars, thereby making virtual interactions more intuitive and immersive.

1.2 Background and Motivation

The advent of mixed-reality environments and text-based systems has underscored the potential of digital humans as interactive tools [1, 2]. These intelligent virtual agents, embedded with speech recognition capabilities, domain-specific chatbot functionalities, and text-to-speech (TTS) technologies, have demonstrated promising results in enhancing user engagement [1, 2]. A notable use case is the transformation of AI models like ChatGPT into interactive virtual teachers, where the generated text can be utilized to produce corresponding audio and body gestures for a more immersive and engaging learning experience as shown in Fig1.1 This approach can be extended to other scenarios, such as pre-visualization of stories, where the text can generate audio and body animations, thereby reducing the cost and complexity associated with generating previz at the initial stage

However, relying on manual mapping of body animations to speech content exposes a fundamental limitation [3, 4]. While a substantial stride towards human-like communication, the present state of these digital entities lacks the subtleties and expressiveness of spontaneous body language, the cornerstone of authentic human interaction [5, 6, 7, 8].

Consequently, the need for an automated and expressive body gesture generation system has emerged. Although effective to a degree, traditional human-created mappings are restrictive in their scope and scalability [9, 10, 11]. Such an approach also fails

to account for the nuances of individual speaker styles, leading to a level of incongruity between the verbal and non-verbal elements of communication [12, 13]. This disconnect undermines the potential for seamless interaction with digital humans, marking it a pressing issue.

The motivation behind this research arises from the aspiration to augment the communication capabilities of digital humans. Introducing a rule-based co-speech gesture mapping system can leverage large publicly available video datasets and word embeddings to generate a diverse and accurate set of gestures [14, 15]. The aim is to bypass the limitations of traditional human-created mappings, thereby bringing the interaction with digital humans closer to natural human communication [16].

The ultimate goal is the development of ConGRets, a real-time co-speech gesture generation system capable of zero-shot speaker style adaptation. This advanced system can take text as input and generate corresponding body gestures, providing a more engaging and interactive user experience in various environments. Such a system can give life to virtual characters in educational, entertainment, or storytelling contexts, generating realistic body animations based on the provided text. This way, it allows for a more immersive and engaging user experience. [17].

1.3 Research Questions and Objectives

In light of the identified limitations and the proposed direction for improvement, the following research questions are formulated:

RQ1: How can we develop an automated system to map speech content to body gestures for digital humans more accurately and diversely, overcoming the limitations of traditional human-created mappings?

RQ2: How can we leverage large publicly available video datasets and word embeddings to develop this system to ensure a wide range of realistic gestures?

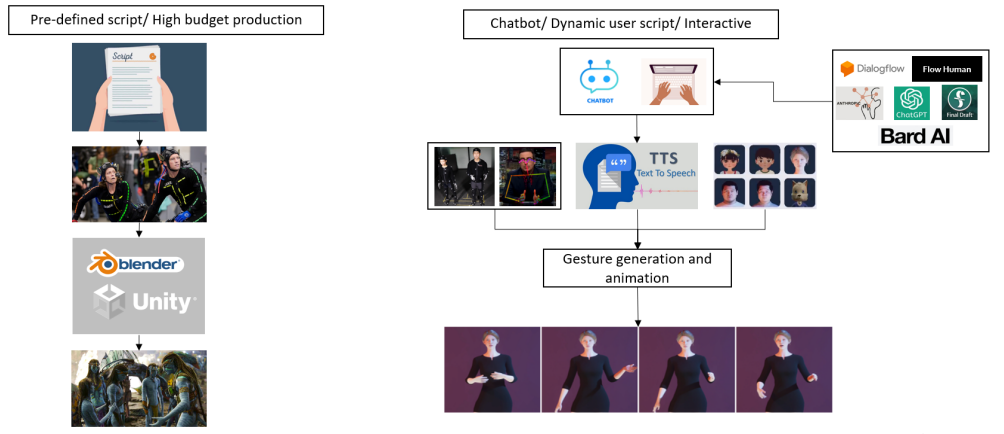


Figure 1.1: Requirements of using Virtual human in different scenarios

RQ3: Can a scalable, low latency gesture generation system be developed to adapt to individual speaker styles, thus enhancing the authenticity of the digital human's communication?

In response to these research questions, the objectives of this research are as follows:

O1: To develop an automated, rule-based co-speech gesture mapping system to generate a diverse and accurate set of gestures for digital humans.

O2: To utilize large publicly available video datasets and clean motion-captured data in developing this system, ensuring a comprehensive range of realistic gestures.

O3: To develop ConGRets, an advanced, scalable co-speech gesture generation system capable of adaptation of zero-shot speaker style, thus enhancing digital human communication's realism and authenticity.

These research questions and objectives guide the subsequent stages of this research, providing a clear path toward enhancing user interaction and engagement with digital humans across various environments.

1.4 Thesis Structure and Overview of Contributions

This thesis is structured into eight comprehensive chapters, each contributing unique insights to the field of co-speech text-to-gesture generation for digital humans.

Chapter 2, "Literature Review," provides an in-depth examination of the existing knowledge in the field. This chapter sets a robust theoretical foundation for the research, from embodied conversational agents and co-speech gesture generation data and techniques to speaker adaptation techniques.

Chapter 3, "Seamless Interaction Framework for Virtual Agents in Mixed Reality Environments", presents intelligent virtual agents' system architecture and multimodal interaction strategies. A specific application scenario in a botanical garden demonstrates the practical implementation of the proposed framework.

Chapter 4, "Automatic Rule-Based Gesture Generation for Conversational Agents", introduces an automated methodology for gesture generation, discussing data sources, rule generation process, and animation sequence generation in detail. The chapter concludes with an evaluation of the proposed method and a discussion of potential enhancements.

Chapter 5, "Cross-Lingual Text-to-Gesture Synthesis for Digital Humans: Generation and Evaluation", delves into the novel approach of GestureCLR, a method used to implement a cross-lingual gesture synthesis system. A detailed user study is presented to evaluate the effectiveness of the approach.

Chapter 6, "Rapid and Memory-Efficient Gesture Generation with Zero-Shot Speaker Adaptation", introduces the ConGRets system, an innovative gesture generation system capable of zero-shot speaker style adaptation. This approach's methodology, implementation, and results are discussed in detail, including implications and potential future work.

Chapter 7, "Discussion," offers an overview of the main findings, limitations, and

future work arising from the research. It also explores the broader implications of the research for virtual human interaction and various applications.

Chapter 8, “Conclusion,” summarizes the research succinctly, tying together the significant findings and their impact on digital human communication.

Each chapter encapsulates a distinct phase of the research journey, contributing to the overarching narrative of enhancing co-speech text-to-gesture generation for digital humans.

Chapter 2

Literature Review

2.1 Embodied conversational agents in Mixed reality

In the work of Liu et al. [18], they define the wide range of applications that MR technology can address. Xiao et al. [1] argued about the new methods of user interactions driven by current-generation commercial mixed reality devices. Agent-based systems and their potential for diverse applications have been researched extensively as described by McArthur et al. [19], McArthur et al. [20], Catterson et al. [21], Castano et al. [22] and Alencar et al. [23].

Xie et al. [24] emphasized the advantages of agent-based systems by defining their significant features. Anabuki et al. [25] introduced a new type of anthropomorphic agent that lives in a 3D space where the real and virtual worlds are seamlessly merged. In the research of Kopp et al. [26], a multimodal virtual humanoid agent assistant was created to have face-to-face discussions with users in a virtual environment. Cai et al. [5] worked on modeling human behavior for virtual agents, and Machidon et al. [27] provided a comprehensive review of the current state of the art in the use

of virtual humans in cultural heritage ICT applications. Vosinakis et al. [28] argue that virtual humans are more beneficial in disseminating intangible cultural heritage in virtual environments. Many studies have focused on creating believable characters or agents that partially show emotions and behavioral abilities when interacting with the environment or users, such as those by Arroyo-Palacios et al. [6] and Bogdanovych et al. [7].

Despite these developments, Richards [2] identified a significant issue: most solutions are geared towards PC-based systems and suffer from repetitive content delivery. To address this, current research focuses on using Wearable MR devices, which provide a more comfortable and engaging experience. These devices can be further enhanced with features such as an artistically crafted 3D virtual human, intelligent chatbot, object recognition, and intuitive interaction technique.

This body of research forms the foundation for our work, which aims to build upon these existing concepts and methodologies to create a more engaging and immersive MR experience. Our framework addresses key limitations in these studies and introduces new features to enhance user engagement and interaction.

2.2 Data for Co-speech Gesture Generation

The data used for co-speech gesture generation has evolved significantly over the years, influencing the quality and naturalness of the generated gestures.

Initial methods relied heavily on handcrafted animation data [29, 3]. These methods manually created gesture units or ”beat” gestures, which were then synchronized with speech. While such methods were pioneering, the handcrafted nature of the data meant that the scope of gestures was limited. This often resulted in gestures that could appear repetitive or less varied.

The development of motion capture technology presented an opportunity to gather

more natural, human-like gesture data. Researchers began using motion-captured data as a richer source of information for gesture generation [30]. Some methods convert this motion into gesture units, creating a form of ”vocabulary” of gestures [30]. Others use the motion and accompanying audio data for end-to-end learning of gesture generation [16, 31, 32]. These methods have the advantage of capturing the subtleties of human motion, which can result in more natural-looking gestures.

In recent years, the largest public motion capture data has been produced by Liu H et al. [33]. This dataset offers a vast range of human movement data, which could help generate various co-speech gestures.

Parallel to these efforts, another line of research has explored using open-source video data for co-speech gesture generation. These methods process videos using pose estimation techniques to extract gestural data and speech recognition to transcribe the accompanying audio [34, 35, 36]. While the data from these sources can be more noisy and variable than motion-captured data, the advantage lies in the diversity and volume of data from publicly available videos.

In short, the type of data used for co-speech gesture generation significantly impacts the quality and naturalness of the gestures produced. As technology and data collection methods advance, we anticipate that co-speech gesture generation will continue to improve.

2.3 Methods for Co-speech Gesture Generation

2.3.1 Rule-Based Systems

Rule-based systems for co-speech gesture generation have a rich history in Human-Computer Interaction (HCI) [37, 38, 39, 40, 41, 42]. These systems generate gestures based on predefined rules that establish a direct correspondence between the text input and the produced gestures.

One of the earliest and most influential works in this area is the embodied conversational agent developed by Cassell et al. [37]. This system introduced the concept of generating synchronized speech, intonation, facial expressions, and hand gestures based on typed text input. The system relied on rules derived from human conversational behavior research.

Building on this pioneering work, the BEAT system [3] was developed. This system generates nonverbal behaviors and synthesized speech based on typed text input, using a more sophisticated set of rules derived from research into human conversational behavior.

More recent rule-based systems, such as the one presented by Marsella et al. [40], have continued to refine this approach. Despite their success in contextually relevant gesture generation, these systems face challenges due to their memory-intensive nature and slower query speeds. These challenges are attributed to the complexity of the rule mapping process [43, 13]. Moreover, the adaptability of these systems to different speakers or speaking styles is often limited, necessitating manual tuning [44, 45].

Rule-based systems have laid a strong foundation for co-speech gesture generation despite these challenges. They provide a valuable point of comparison for newer, data-driven methods and continue to offer insights into how we can create more natural, engaging, and meaningful gestures.

2.3.2 Learning-Based Methods

To mitigate the limitations of rule-based systems, researchers have shifted their focus towards learning-based methods for co-speech gesture generation [46, 47, 48, 49, 50, 51, 52]. These methods employ data-driven models that learn gesture generation from large datasets, often yielding improved performance over rule-based systems.

One of the earlier attempts in this direction was made by Kipp et al. [46]. They proposed a system that uses annotated video data to learn the mapping between speech

and gesture, providing an initial step toward a data-driven approach. Similarly, Chiu et al. [49] employed machine learning techniques to predict co-speech gestures from the text.

As deep learning techniques advanced, they were quickly incorporated into gesture generation models. For instance, Yoon et al. [47] introduced an end-to-end neural network model that learns to generate co-speech gestures for robots based on TED talks. This model was trained to map between text and corresponding gestures, demonstrating the potential of deep learning for this task.

Kucherenko et al. [51] and Asakawa et al. [48] followed a similar approach but instead used speech input to generate gestures, showing that different modalities can be used for gesture generation.

More recently, Bhattacharya et al. [50] took this further with Text2Gestures, a transformer-based approach for generating emotive full-body gestures from textual inputs. This work demonstrated that complex models like transformers could be effectively used for gesture generation.

Using statistics-based techniques, Yang et al. [52] explored a different path. Their approach showed that data-driven methods are not limited to machine learning and deep learning models.

Contrastive learning has also emerged as a promising approach in the realm of co-speech gesture generation [15]. This method differentiates between similar and dissimilar gestures, allowing models to generate more nuanced and contextually appropriate gestures. GestureCLR [15] is a notable example that matches poses to gesture units using this technique.

In conclusion, learning-based methods for co-speech gesture generation have shown significant potential, leveraging the power of machine learning and deep learning to generate more natural and contextually appropriate gestures.

2.3.3 Speaker Style Adoption

Learning-based methods have proven effective in creating contextually suitable and natural co-speech gestures, but the ability to replicate the unique styles of specific speakers remains a significant challenge [35, 53].

Existing approaches to gesture generation typically either combine data from many speakers or customize a system for each speaker. The first method can lead to a generic style that may not fit all contexts or speakers, while the latter can be resource-intensive and less scalable due to the significant data requirement for each speaker [34].

To address these issues, various techniques have been proposed to incorporate multi-speaker styles into a single gesture generation system. Some models use one-hot encoding of speaker identity in training, which allows the system to learn and generate unique style embeddings for multiple speakers, but limits it to the speakers on which it was trained [35, 50]. Others employ style transfer techniques, which can adopt the style of a new speaker but require retraining [36]. Zero-shot speaker style adoption has also been proposed but relies on both audio and gestures, making it audio-dependent [54].

While these techniques show promise, they underline that speaker style adoption in co-speech gesture generation remains an underexplored area. Our primary focus is to advance in this field by developing a speaker style adoption mechanism that does not require retraining and relies solely on gestures.

Chapter 3

Seamless Interaction Framework for Virtual Agents in Mixed Reality Environments[1]

3.1 Introduction

Mixed reality (MR) is a technology that merges the real and virtual worlds, creating unique environments where users can interact in real-time [18]. Commercial MR devices, such as Microsoft HoloLens, Magic Leap, and Meta 2, propel innovative user interaction methods through hand gestures, voice, and gaze [1].

Agent-based systems have also received significant attention due to their potential in diverse applications [19, 20, 21, 22, 23]. An intriguing development in this field is the emergence of anthropomorphic agents in a 3D space, where the real and virtual worlds are seamlessly merged [25]. Ideally, these agents should mimic human behavior, showing emotions and behavioral abilities when interacting with the environment

[1]This work was published in 32nd International Conference on Computer Animation and Social Agents (CASA 2019) as Design of Seamless Multi-modal Interaction Framework for Intelligent Virtual Agents in Wearable Mixed Reality Environment

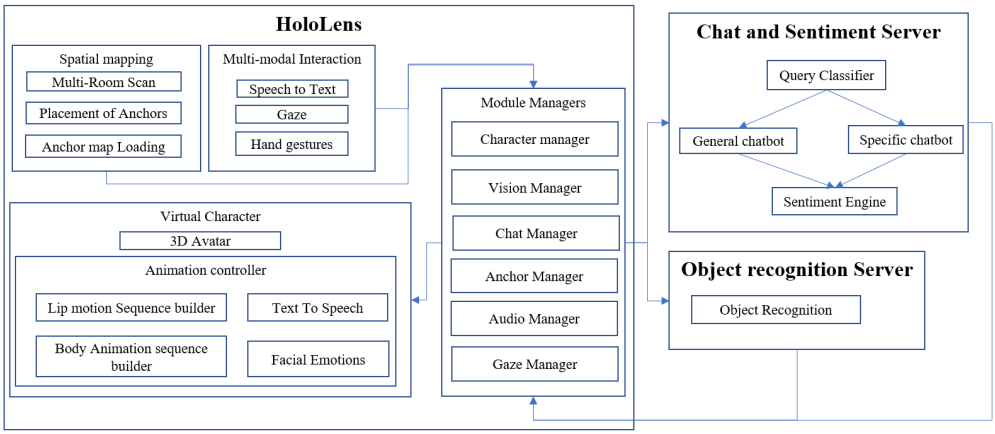


Figure 3.1: Overall framework for an intelligent virtual agent in wearable mixed reality.

or users [26, 6, 7].

However, to further enhance the interactive user experience, a more engaging and less repetitive form of content delivery is needed [2]. The current generation of Wearable MR devices offers a more immersive experience that can be further enhanced by integrating a 3D virtual human, an intelligent chatbot, object recognition, and intuitive interaction techniques.

We propose a multimodal interaction framework for more than one agent in an MR environment specifically designed for museums and botanical gardens. Our framework incorporates a virtual character, multi-modal interaction, a domain-specific chatbot, and an object recognition system, aiming to provide a seamless, interactive, engaging, and non-repetitive experience. The framework is highly adaptable and flexible due to its modular approach, allowing users to explore the real world with digital information delivered by a friendly virtual human. Our contributions are as follows:

- Design and implementation of the virtual agent framework for wearable MR.
- Interaction design and implementation on the intelligent virtual agent.

- Expression and body gesture mapping of intelligent virtual agent.
- Scalable anchor system for wearable MR.

The framework is explained in section 2. In section 3, we briefly introduce the development platform. In section 4, a scenario is presented and discussed.

3.2 System Architecture for intelligent virtual agents in MR Environments

To build a useful and meaningful system by integrating all required components, requirements in the wearable mixed reality system are:

- **Interaction:** Virtual agent which recognizes user gazing objects and converse about it.
- **Expression:** Virtual agent which shows emotion and body gestures appropriate to the conversation.
- **Scalability:** Virtual agent who shows up at an infinite working area.

The overall architecture is shown in Figure 3.2.

3.2.1 Interaction: Gazing object recognition and conversation.

The interaction process aims to provide a more natural interaction method with the virtual character. The major problem was asking the virtual character about any object for the first time. The problem arises when the user needs to ask about some object. If the user gazes at the character and asks about the object, the user must quickly look at the object to capture it. This sometimes does not give enough time to react. This problem was solved by including limited voice commands, i.e., ”What is this,” ”Tell me about this,” etc., in our framework. Figure 3.3 shows the interaction process with and without anchors in sight.

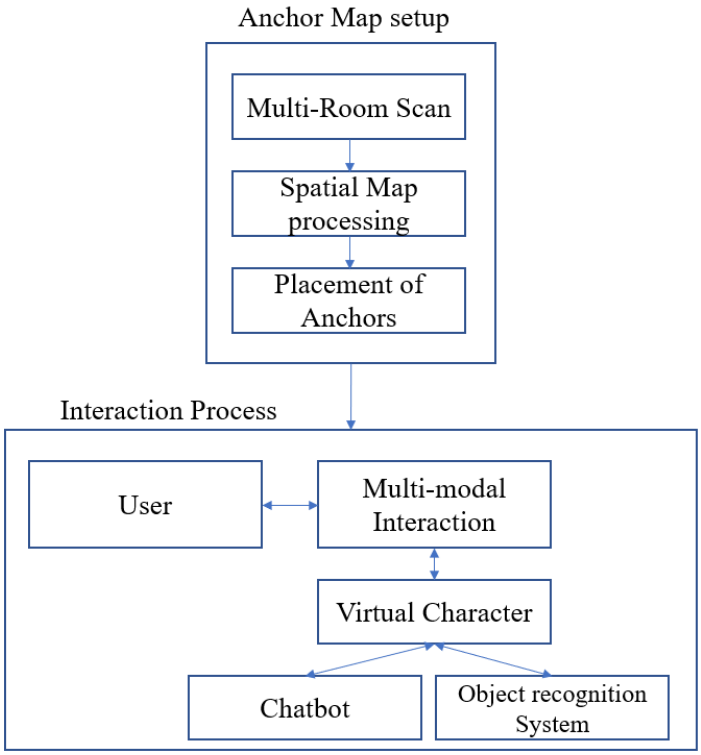


Figure 3.2: Overall architecture.

3.2.1.1 Gaze and speech based multimodal interaction:

There has been plenty of research about creating natural methods for human-agent interaction. For example, this study by Freigang et al. [55] attempted to emphasize the role of speech and hand gestures in conveying pragmatic information between users and virtual humans. Arroyo-Palacios et al. [6] proposed an approach for interactive virtual characters for MR applications. Gaze and gestures were applied by users so that they could feel comfortable with a natural way to interact with agents. According to Kolkmeier et al. [56], gaze and proxemics behaviors can establish and maintain a level of intimacy for human-human interaction and for human-agent interaction. Results indicated that agents exhibiting higher proximity caused participants to step away from

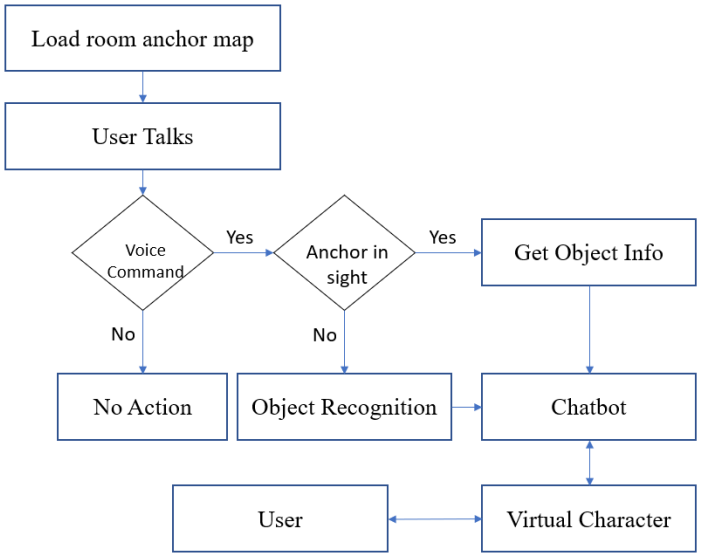


Figure 3.3: Process to get recognized object information

more than agents with low proximity.

In a study by Yumak [57], a gaze behavior model was presented for an interactive virtual character in the real world. Uchino et al. [58] and Rebman et al. [59] showed the importance of speech recognition in interaction. The study of Avramova et al. [60] presented a toolkit with full-body animation for the virtual character, which offered MR interaction. We utilized gaze and speech recognition to interact with the virtual character. HoloLens provides the API for gaze and speech recognition. HoloLens Toolkit provides high-level speech recognition functions like dictation manager, which enables us to say long sentences. As shown in Figure 3.4, our method does not require any trigger keyword. The speech recognizer starts when the user gazes at the virtual character for 4 seconds. If the user talks, then it will wait to finish the user talking and then send the query to a chatbot. If the user does not speak, then the virtual character will try to start the conversation by saying, ”Hello, do you need help?” or similar sentences. The conversation will end if the user gazes out while the speech recognizer is running and

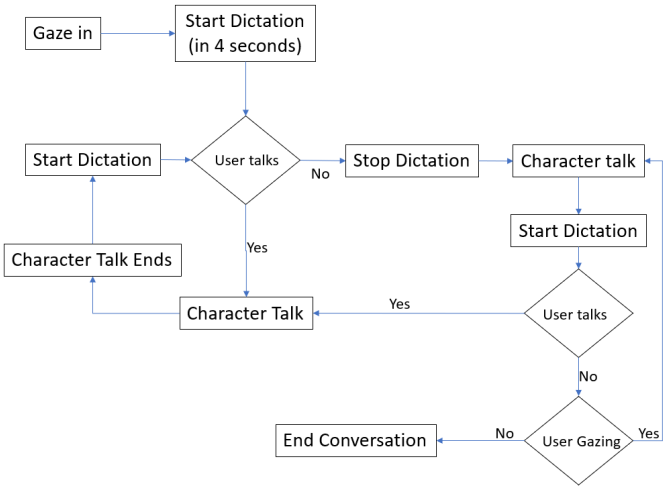


Figure 3.4: Gaze and Speech interaction architecture

the user is silent.

3.2.1.2 Gazed object recognition:

The object recognition system works as the eyes of the virtual character. It enables the virtual character to be aware of its soundings. Any state-of-the-art or custom-trained object recognition or image classification system can be used if it can handle web requests. The vision manager module handles object recognition-related tasks. It only needs a valid API Endpoint. The user can make their API Endpoint using any available system or use an existing API Endpoint. Vision manager is flexible to handle as it can also be connected to the on-device system. By design, this can cover any number of objects if the chatbot has the required knowledge base.

3.2.1.3 Conversation of intelligent virtual agent:

For a virtual character to interact with the user, it needs a knowledge base and associated program to give out the required information. Those programs are called chatbots.

Chatbots play a vital role in Human-Computer Interactions. Chatbots are also being used to automate customer services. In our system, the chatbot serves two purposes, i.e., 1) It provides domain-specific information. 2) General conversation options. Our chatbot is equipped with a sentiment engine that analyses the response and attaches a sentiment class and level of sentiment to it, as shown in Figure 3.5. This chatbot is deployed on the cloud and accessed through web requests. Input parameters to this chatbot are "(*Query, Object*)." *Query* can be a question, comment, or answer to a virtual character's question. *Object* is any virtual or real object detected through object recognition about which we want to ask a question. *Query* can be a general query like asking about the character itself or a specific query about the objects around you. A prerequisite for a specific query about any object is that the user must look at the object for the first time and ask about it. This will trigger the object recognition component and pass the information to a chatbot. Follow-up queries don't require looking at the object. Reply to the *Query* is "(*Reply, Sentiment Class, Sentiment Level*)." Sentiment class and the level are used for appropriately animating the character. The user can ask both general or specific queries in one conversation and any order. This style of conversation mimics human-human interaction. This increases engagement and improves the quality of experience.

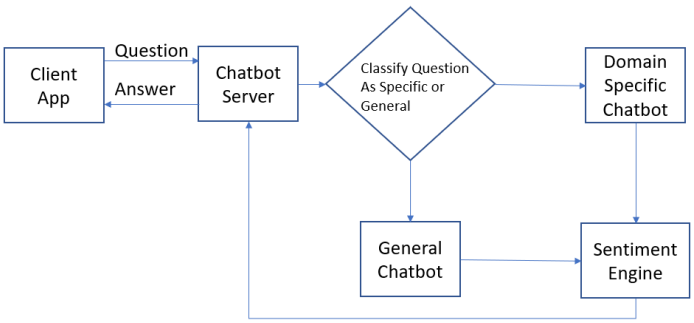


Figure 3.5: Chatbot Architecture

3.2.2 Expression: Virtual character emotion and body gesture

Our virtual human is a 3D model. To make a realistic virtual human, it must have rich body animations and facial emotions. Body animations are for verbal and nonverbal activities. As virtual humans are required to move around and respond to verbal cues, they must have a large set of animations available. In our system, we have 30 different animations. Facial emotions are for the expression of different emotions. Our system has four emotions, i.e., Joy, Angry, Sad, and Fear. Each emotion has three levels, i.e., High, Medium, and Low, as shown in Figure 3.6.

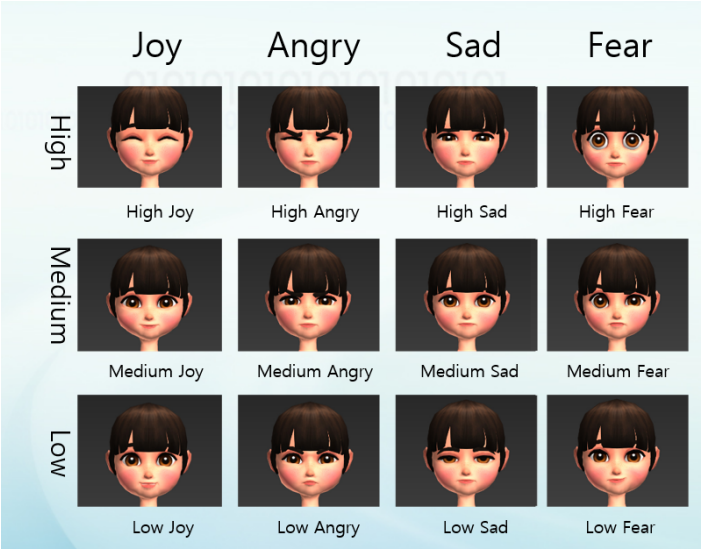


Figure 3.6: Facial emotions types and levels

Body animations and facial emotions are applied based on the content of the speech. Speech is generated from text using the Text-to-Speech system. Figure 3.7 shows the architecture for building the animation sequence. This architecture requires a mapping table for mapping text to animations. Human experts create this map. It has an animation for phrases and individual words utilized by the animation builder, as shown in Figure 3.7. Facial emotions are predicted from the text by sentiment engine integrated

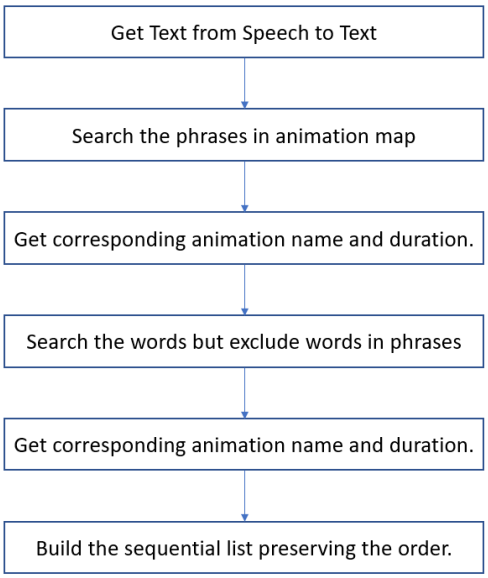


Figure 3.7: Architecture for building body-animation sequence

into the chatbot and applied for the complete duration of the text. Speech from the text is lip-synced by converting text to phonemes sequence. A Phonemes map is used for this purpose. The lip shape for each phoneme is shown in Figure 3.8. Speech, body animation, and facial emotions are applied in parallel, as shown in Figure 3.9.

3.2.3 Scalability: Anchor map setup

HoloLens provides a mechanism known as anchors through which static objects can be tracked. By combining image recognition and anchors, we created a process by which we can map infinite working areas. This process can save information about real objects which the virtual agent can access. This process minimizes the use of object recognition for each object. Figure 3.10 a) shows the anchor placement process. The initial two steps are part of the framework's anchor manager. Since there is no mechanism to load room-specific anchors in HoloLens and for identical rooms, wrong

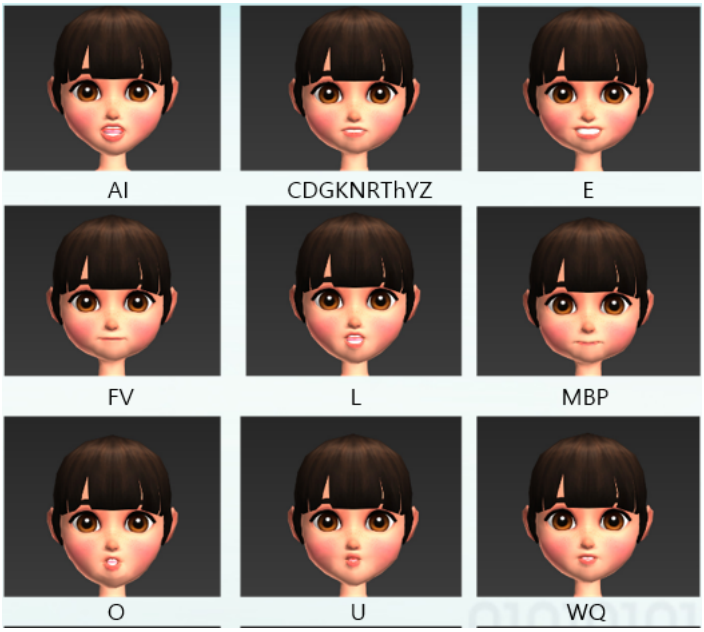


Figure 3.8: Lip shape for phonemes

anchors may be loaded. The anchor management process utilizes object recognition to load room-specific anchors to avoid this issue. The process is shown in Figure 3.10 b).

3.3 Development environment and implementation

The platform comprises Unity3D, Microsoft HoloLens, HoloLens Toolkit, and cloud computing. HoloLens is a spatial MR holographic device developed by Microsoft. With HoloLens, we scan the room to create a 3D space through which users, virtual objects, and virtual characters are tracked. HoloLens supports Unity3D as a development platform. The HoloLens toolkit provides an easy way of setting up the initial project and settings. It also offers numerous helping functions for mapping, gaze tracking, and speech recognition. HoloLens is a self-contained device with limited resources. It is impossible to efficiently run a chatbot and object recognition from the device without

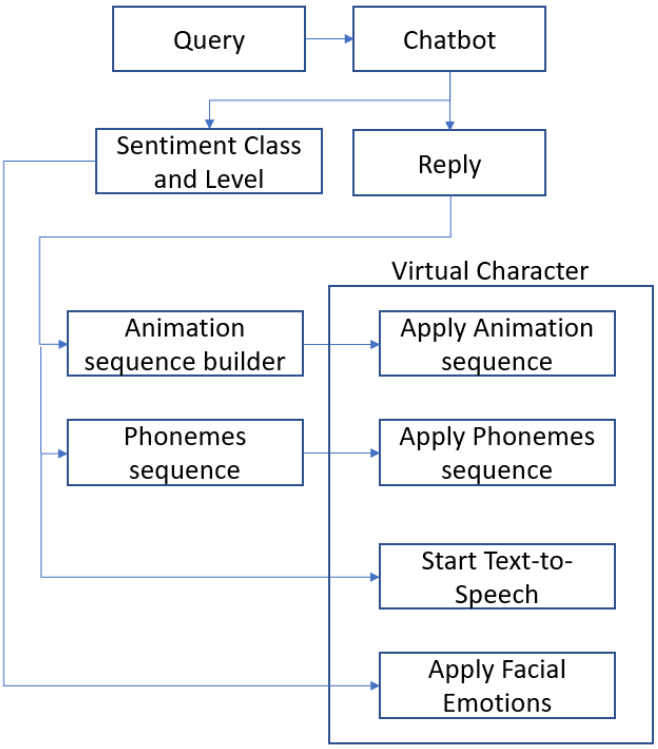


Figure 3.9: Applying speech, animation, and facial emotions on the character

compromising the quality of the user experience. Our framework access a cloud-based chatbot and object recognition system via module manager to solve this.

3.4 Application scenario: A botanical garden

The framework was utilized to create a demo system. The scenario was a tour at a Botanical garden. In this scenario, the user can interact with the virtual character to ask about different flowers. This scenario needed the following items:

- 3D humanoid model
- Flower dataset for the object recognition system

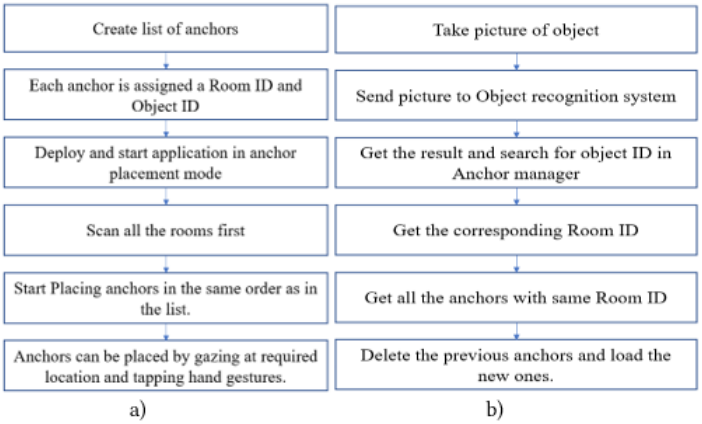


Figure 3.10: a) Anchor placement process. b) Anchor loading

- Knowledgebase of flowers for chatbot

3D artists created 3D Models. It included body design, clothes, and facial emotions. Body animations were acquired from Mixamo[https://www.mixamo.com]. There were three models, i.e., a Girl, a Boy, and Old Man. Different voices from Text-To-Speech systems were assigned to them. Body animation also varies for each character. The framework required no customization to add multiple characters in one application. Categories of the flower were nine. We used artificial flower models. A dataset of images was prepared from those models. Our object recognition system was trained and deployed on the custom vision API of Microsoft Azure. We trained the system using our dataset. Knowledgebase for the chatbot was built with extensive research. The chatbot's goal was to include as much information as possible so that the user should not be forced to ask only specific questions.

For this scenario, two rooms were set up with five flowers in 1 room and 4 in 1. Scanning was performed, and anchors were placed. The girl model was assigned room 1, and the Boy model was assigned room 2. Then we successfully demonstrated the framework. The outcome was a seamless interaction with a response time ranging be-

Figure 3.11: Models

Figure 3.12: Nine flowers used in the scenario.

tween 2-4 seconds. This time includes speech-to-text conversion, sending text queries to a chatbot, receiving a response, and building an animation list. This response time, however, does not include the time of silence to detect the end of user speech. The response time will be higher if object recognition is also triggered, typically 5-8 seconds total. As this delay is more significant, so during this time, our character shows the animation of thinking about the object and says something like, ”Let me see, let me think about it.” This way, the user does not feel a time delay. Our modules were deployed on a cloud platform and accessed through the internet, so most of the time were utilized in web requests. In Table 3.1, we summarize the average results for three queries. Query A was loading the anchor map, and it required object recognition. Query B was a general conversation. Query C was about flowers, and it needed object recognition. The above table proves the usability of the anchor map to reduce response time.

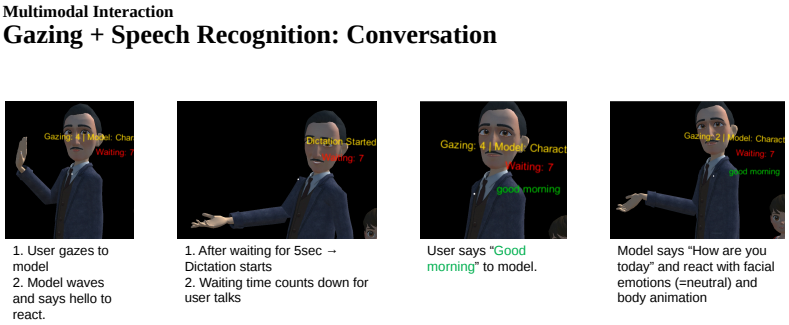


Figure 3.13: Interaction with model

Table 3.1: Total average response time for each query

Query	Query A	Query B	Query C
Total queries	30	30	30
Object recognition	5.4s	n/a	5.3s
Chatbot	n/a	2.1s	2.0s
Processing time	0.5s	1s	1s
Total Time	5.9s	3.1s	8.3s
Std Dev	0.99	0.98	0.99

3.5 Conclusion

We studied and conceptualized a framework by integrating spatial mapping, virtual character, multi-modal interaction, chatbot, and object recognition for wearable mixed reality. We successfully demonstrated that such a framework could reduce the time to develop an augmented reality environment with an intelligent virtual agent. It can prove to be more engaging and entertaining. The approach presented theoretical and experimental aspects emphasizing the importance of using multimodal interactions between humans and agents. The potential of spatial mapping, virtual character, multi-

modal interaction, chatbot, and object recognition was effectively integrated to enhance human-agent interaction and to stimulate the realism of the mixed reality environment.

Chapter 4

Automatic Rule-Based Gesture Generation for Conversational Agents[1]

4.1 Introduction

Communication is a vibrant tapestry, meticulously woven with threads of words, expressions, and gestures, creating a vivid panorama of thought and sentiment. These non-verbal cues, particularly gestures, lend depth and clarity to our verbal messages, painting them with layers of emotion, emphasis, and context. It is well-documented in the works of [61] that during verbal exchanges, individuals frequently resort to gestures to relay their oral messages clearly or to manifest their emotional intent. A particularly fascinating class of gestures, the co-speech gestures, synchronize with speech to articulate the spoken content and emotion in a visually engaging manner. Hand motions contribute significantly among the different co-speech gestures, capturing a large portion of the co-speech gesture space as outlined by [8].

As we embark on the era of virtual and augmented reality (VR/AR), we encounter

[1] This work was published in Journal of Computer Animations And Virtual Worlds 2020 as Automatic Text-to-Gesture Rule Generation for Embodied Conversational Agents

an intriguing advent of embodied conversational agents (ECAs) that contribute significantly to various virtual scenarios, ranging from entertainment and education to training and interactive learning [62, 63, 64, 65, 66, 67, 68, 69]. These ECAs, digital incarnations mimicking human-like behavior, necessitate the incorporation of believable gesture animations to convincingly represent the dynamism of conversation [70, 71]. However, generating meaningful and believable gestures for these ECAs is a daunting challenge, largely due to the complexity and subtlety of human non-verbal communication.

Our work seeks to navigate this labyrinth of gesture generation, proposing a novel approach that significantly enhances the current rule-based methods. We introduce two fundamental improvements in the existing framework. First, we automate the rule generation procedure, seamlessly integrating the process into our model. Second, we refine the rule-searching algorithm to identify semantically rich entries from the rule base. Therefore, our primary contributions can be encapsulated as follows:

- An innovative, automated rule generation method for ECAs that leverages the immense potential of a large publicly available dataset, thereby democratizing access to high-quality gesture generation.
- A sophisticated enhancement of rule searching by deploying semantic-aware word embedding for text-based co-speech gesture generation, enhancing the capability of ECAs to generate more natural and contextually appropriate gestures.

4.2 Automatic rule generation methodology

Automating complex processes in this digital transformation era is more crucial than ever. Our approach aims to automate rule generation, extracting co-speech gestures from a vast video data repository. This method, devoid of human intervention, produces

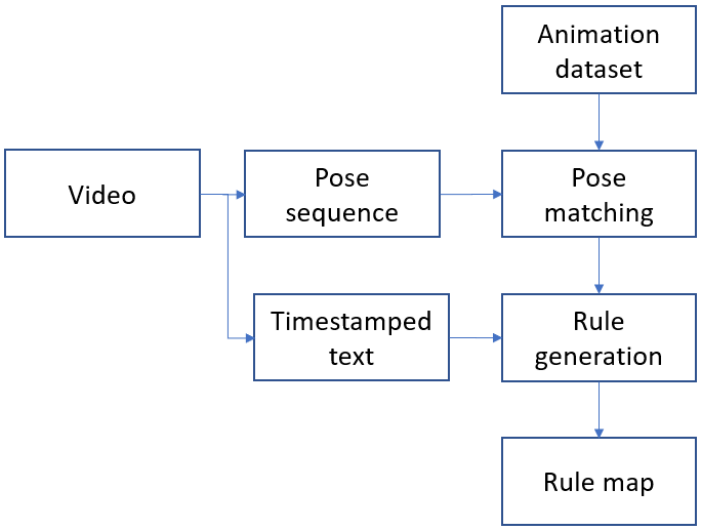


Figure 4.1: Overview of our approach pipeline for rule-map generation.

a valuable mapping table between text and corresponding co-speech gestures.

Traditionally, rule-based systems depend heavily on experts who use an assortment of pre-defined animation clips, manually mapping them to specific words or phrases. These animation clips are generally grouped into various categories such as beat, deictic, iconic, and metaphoric. The mapping is predominantly based on their extensive expertise and experience. Even in the case of some data-driven methods, manual annotations are incorporated to mirror these categories.

Advancements in pose estimation methods have drastically simplified the process of generating pose sequences from RBG videos. These innovative techniques can generate either a 2D or 3D pose sequence. However, as of now, 2D pose sequence estimation has shown superior results compared to its 3D counterpart.

OpenPose [72], a widely recognized open-source 2D pose estimation system, has made significant contributions in this domain. Despite its commendable speed, OpenPose may occasionally produce artifacts or inaccurate results, rendering the output unsuitable for direct usage. Nevertheless, these pose sequences prove highly beneficial

for pose matching and classification.

Our methodology exploits this potential by using a collection of clearly defined 3D gesture animations to generate a corresponding set of 2D projections. Following this, we implement 2D sequence-to-sequence matching. As a result, we achieve a comprehensive mapping without human experts. This innovation in automatic rule generation methodology signifies a major leap forward, paving the way for more efficient, accurate, and automated co-speech gesture mapping in the future.

4.3 Data Sources

The dual-pronged dataset that forms the backbone of our approach comprises of two distinct elements. The first segment involves videos culled from publicly accessible platforms such as streaming websites. These rich and diverse sources facilitate the extraction of both pose sequences and the corresponding speech-to-text transcription for each clip.

This process is made possible through the automated pipeline developed by Yoon et al.[47], designed to efficiently extract pertinent data from an extensive range of videos. They have provided a substantial dataset, the *TedTalk dataset*, comprising 106 hours of video data. Each clip in this dataset includes pose frames and corresponding text uttered by the speaker, as illustrated in Figure4.2.

The textual data is meticulously aligned with the pose sequence, achieved by converting time into frame numbers. Thus, each word is assigned a starting and ending frame number. However, as the text is derived from subtitles and converted into a frame interval, occasional discrepancies in alignment may occur.

The second element of our dataset introduces the 2D projections of pre-defined gesture animations sourced from the ICT Virtual Human Toolkit [10]. This additional dataset will henceforth be referred to as the *Gesture dataset*. The *Gesture dataset* en-

compasses animations of varying lengths. For uniformity, all animations were processed to match the length of the longest animation by padding zero frames on either side of each animation.

We initially selected only the upper body joints from the *TedTalk* and *Gesture datasets* for optimization purposes. These selected joints were normalized using the neck joint as a reference point. This meticulous data sourcing and preparation ensure the robustness and reliability of our automatic rule generation methodology, contributing significantly to the quality of the result.

Figure 4.2: Dataset sample with pose and corresponding text

4.4 Rule generation process

Our approach's primary innovation revolves around utilizing the utilization of the *Gesture dataset* as a strided sliding window over the *TedTalk dataset*. This unique methodology aims to derive a cohesive {gesture, text} pair. The algorithm initiates by selecting a clip from the *TedTalk dataset*, over which the *Gesture dataset* is progressively slid.

At each step of this process, we compute the frame-by-frame cosine similarity for each gesture and subsequently derive its average. Following this, we select the gesture with the closest similarity score. A threshold of 0.92 was established for this selection process. The selection is made randomly in scenarios where more than one gesture exceeds this threshold. The rationale behind establishing this specific threshold value is elaborated in the discussion section.

Upon selecting the gesture, we extract the text from the corresponding location in

the *TedTalk dataset*. It's important to note that the maximum length of the corresponding text is confined to five words. If the text length surpasses this limit, it is strategically trimmed from both ends to maintain uniformity.

Subsequently, the derived {gesture, text} pair is stored in the mapping table. The stride, or the shift between each sliding window step, is equivalent to the length of the gesture sequence. This sophisticated algorithm, instrumental to our automatic rule generation methodology, is further detailed in Algorithm 1.

4.5 Animation sequence generation

Upon the successful derivation of the rule-map, it is primed for use. At runtime, the text intended for animation is relayed to the rule-map driver algorithm. This algorithm takes the given text, tokenizes it into five-word chunks, and subsequently searches for each chunk within the rule-map. The corresponding animation for the best match is retrieved, and the animator executes the resulting animations sequentially. In our instance, this was achieved using the Unity3D animator.

Given that the search for the rule essentially involves string matching, the potential for further refinement exists. Traditional methods primarily employ direct string matching, which, despite being straightforward, exhibits a key shortcoming - it disregards the context of the string.

To address this, we took inspiration from advancements in Natural Language Processing (NLP) research. We leveraged the concept of word embedding or word2vec, specifically using the GloVe model [14] pre-trained with 300 dimensions. During the matching step, we obtained vector representations of the five-word chunk from the input text and the text from the rule, followed by a cosine similarity measurement. The vector representation of a given string was achieved by summating all word vectors of each word in the string, resulting in a vector of 300 floating-point values. The com-

prehensive algorithm is detailed in Algorithm 2, and a comparative analysis between word embedding and string match is discussed in the subsequent section.

An integral aspect of animation sequence generation is preserving time consistency. Given that gesture animations vary in length, ranging from 1 to 3 seconds, we designated a fixed time slot for each animation. The empirical analysis led us to conclude that a 5-word chunk generates suitable time slots. Thus, we partition the input text into 5-word chunks, discarding any chunk smaller than five words. If the input text comprises fewer than five words, it is processed as a single slot. For instance, a 15-word sentence would translate into 3 animation slots.

The time allocated for each slot is dynamically calculated at runtime by dividing the audio length from the Text-To-Speech (TTS) system by the number of animation slots. This ensures higher time consistency for the gestures, regardless of the TTS system used.

4.6 Evaluation: Objective and subjective assessments

4.6.1 Objective Evaluation

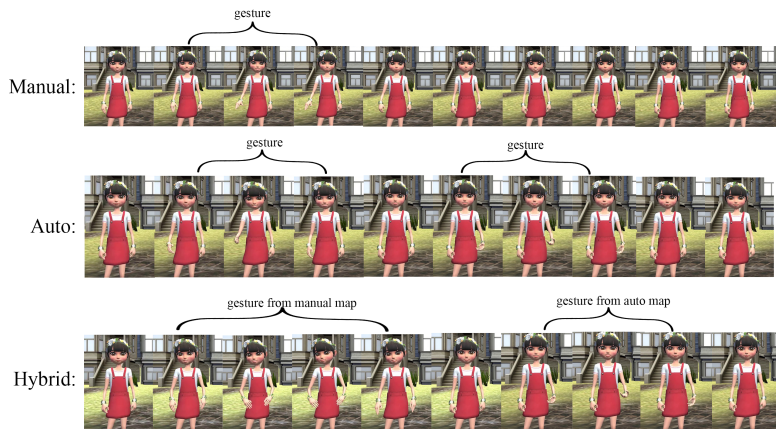


Figure 4.3: Gestures for a sample text “The lion is not a trained pet. It can hurt you”

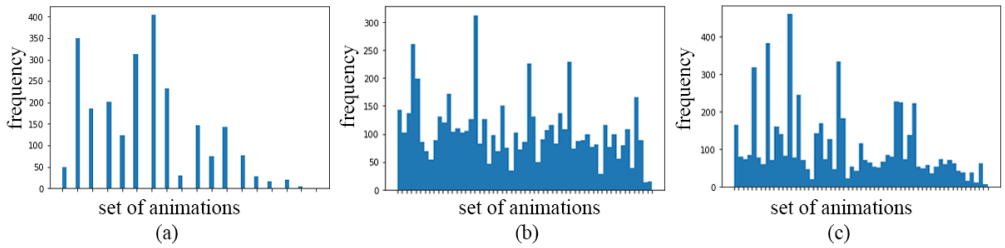


Figure 4.4: Evaluation of gesture distribution. x-axis is set of animations and y-axis is frequency of each animation. Left to Right: a) Manual b) Auto c) Hybrid

To critically evaluate our methodology, we employed three distinct rule-based mapping tables:

- **Manual:** This is the default rule-base in the ICT Virtual Human Toolkit. Each rule within the *Manual map* encompasses multiple words and corresponding multiple animations.
- **Auto:** This map is generated through our automatic rule generation approach. Each rule in the *Auto map* comprises a sequence of words and a single animation. Word embedding technique is employed for rule retrieval.
- **Hybrid:** This map represents a combination of the *Manual map* and the *Auto map*, with the *Manual map* assigned a higher priority.

The hypothesis underpinning our evaluation posits that ”for a fairly large and general text input, a gesture generation system should be capable of retrieving the maximum number of gestures from a given map.” To test this, we fed the complete novel ”Animal Farm” by George Orwell into our system.

As depicted in Figure 4.4, the output consisted of the frequency of each animation used throughout the entire text by each map. The statistics revealed that the *Manual map* only utilized 18 out of 57 animations. In contrast, the *Auto map* and the *Hybrid*

map used all available animations. This indicates that our method yields a more generalized and less repetitive animation sequence.

One significant observation is that a mapping specific to an application can generate a more plausible animation sequence. Our method is designed for general conversation scenarios where gestures serve a complimentary role, enhancing the overall effectiveness of the conversation.

Figure 4.3 exhibits a sample of animation sequences generated by the three maps. From the *Manual map*, the system produced a sequence with one animation, whereas the *Auto map* and the *Hybrid map* each generated sequences with two animations. This underlines the greater versatility and adaptability of our automatic rule-generation approach.

Algorithm 1: Automatic rule generation

```
Result: {gesture,text}
Initialize: TeldTalk-dataset, Gesture-dataset, stride, pairList, clip, nGesture, subclip, gesture,
  dist, text
stride=Length(Gesture-dataset[0])
pairList=[]
while !EOF TedTalk-dataset do
    clip=nextClip()
    while !EOF clip do
        nGesture=[]
        subclip=clip[i:i+stride]
        while !EOF Gesture-dataset do
            gesture=nextGesture()
            dist=averageCosSimilarity(subclip,gesture)
            if dist ≥ 0.92 then
                nGesture.append(gesture)
            end
        end
        i=i+stride
        if size(nGesture) ≥ 1 then
            gesture=randomSelect(nGesture)
            text=getText(subClip)
            pairList.append((gesture,text))
        end
    end
end
```

Algorithm 2: Animation sequence generation

```
Result: {animation-sequence}
Initialize: rule-map, text, sequence, text-tokens, token, vecToken, nSim, rule, vecRule, sim,
  mSim
rule-map //pair-list
text //to be animated
sequence=[]
text-tokens=tokenize(text,5) //5 word token
while !EOF text-tokens do
    token=nextToken()
    vecToken=ConvertToVec(token)
    nSim=[]
    while !EOF rule-map do
        rule=nextRule()
        vecRule=ConvertToVec(rule)
        sim=CosSimilarity(token,rule)
        nSim.append((sim,rule))
    end
    mSim=max(nSim.sim)
    sequence.append(mSim.rule)
end
```

4.6.2 Subjective Evaluation

In this section, we describe the experiment we conducted to see how human users perceive the generated gestures of ECA. Similar to the objective evaluation presented in the previous section, we compared gestures generated by the three methods: *Manual*, *Auto*, and *Hybrid*. However, here we measured users' subjective ratings on three categories: *naturalness*, *time consistency*, and *semantic consistency*. We used the questionnaire in [73]. Each category consists of three questions, and we did not modify the questions.

For this study, we prepared five texts of various lengths and converted them into audio using Google Text-to-Speech. To generate co-speech gestures, we implemented a virtual environment with a female human character standing at the center using the Unity3D game engine (see Figure 4.3). The lip animations were achieved using the SALSA LipSync [2]. The co-speech gestures were rendered using the three-generation methods per each text and recorded. Thus a total of 15 videos were generated.

We made an online survey form using the pre-generated videos. A within-subjects design was used in making the survey form, in which participants watched each video and answered the questions asking their subjective impression of the co-speech gesture of the ECA. Videos of the same text were grouped, and the order was randomized. We asked the participant to read the text before watching the video. Participants were asked to rate each question using a 5-Likert scale (1: totally disagree, 5: totally agree). We additionally asked participants to rank the three methods based on their preferences.

Participants were recruited from Amazon Mechanical Turk. A total of 29 participants completed the survey. However, data from 9 participants were removed as they did not pass our screening criteria. To identify insincerely completed surveys, we embedded an attention question among the actual questions in the middle of the survey. We also measured the completion time. Participants who failed to pass the attention question or spent less than the threshold time—members of our group completed the survey to set the threshold time—were rejected. Finally, there were 20 valid participants (4 females, 16 males, age 22–67, average 36.6). Fourteen were native speakers, four were fluent or proficient users, and two marked their English proficiency as conversational.

We aggregated the ratings per category for each generation method to perform statistical analysis and calculated mean ratings per participant. We treated preference rank as ratings—we assigned 1 to the most preferred and 3 to the least preferred method.

[2] https://crazyminnowstudio.com

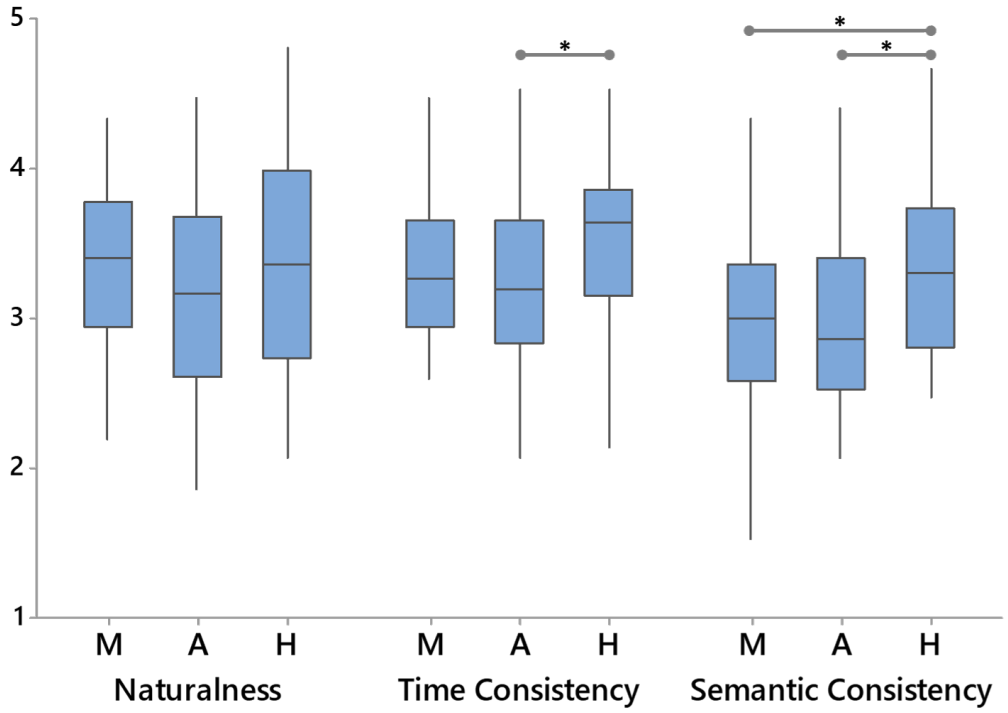


Figure 4.5: Results of subjective measures. Whiskers in the box plots are extended to represent the data points less than 1.5 interquartile range (M: *Manual*, A: *Auto*, H: *Hybrid*, *: $p < .05$).

Due to the ordinal aspect of the original ratings, we performed non-parametric Friedman tests with a 5% significance level on each category. The results are summarized in Figure 4.5.

There were significant main effects of the gesture generation methods on *time consistency* ($\chi^2 = 8.72$, $p < .05$), *semantic consistency* ($\chi^2 = 10.31$, $p < .01$), *preference* ($\chi^2 = 7.80$, $p < .05$), but not on *naturalness* ($\chi^2 = 2.36$, $p = .31$). Pairwise comparisons using Dunn's post-hoc tests with Bonferroni correction revealed that participants gave higher ratings on *time consistency* for Hybrid compared to Auto ($p < .05$). For *semantic consistency*, ratings for Hybrid were higher than those of Manual ($p < .05$) and those of Auto ($p < .05$). Finally, participants favored Hybrid compared to Manual ($p < .05$),

but the tendency was insignificant when compared with those of Auto ($p < .08$).

4.7 Discussion: Enhancements and synergistic effects

4.7.1 Ensuring Gesture Diversity

One of the key challenges associated with rule-based systems is the requirement to formulate rules for all possible gestures that humans naturally use [73]. To navigate this issue, we have leveraged a data-driven approach, in which models are refined through extensive datasets. Rather than training models, we chose to extract rules from these datasets.

The motivation for employing such large datasets is the assumption that if diverse gestures are covered within the dataset, the models trained on these datasets will reflect the same variety. In deep learning, two recent research studies share a similar focus to our project.

Kucherenko et al.[73] utilized a large set of data captured in a studio environment. However, their method predominantly generated beat gestures. Similarly, Yoon et al.[47] employed the *TedTalk dataset*, using the vector representation of text input as the input to their network. Their results also indicated a predominance of beat gestures, with only a limited number of other gestures.

Upon observing the failure of data-driven approaches to generate a diverse range of gestures, we decided to leverage predefined gesture animations, which we refer to as the *Gesture dataset*. Our approach was designed to generate or learn rules from a substantial dataset of real human gestures.

Rather than learning gestures directly from the dataset (as with other data-driven approaches), our methodology utilizes the diverse gesture animations in the *Gesture dataset*. We intentionally designed our automatic rule generation to avoid bias towards a subset of gesture animations in the *Gesture dataset*.

In the automatic rule generation process, we adjusted our similarity check threshold

to 0.92. This threshold was determined empirically, guided by the principle that a good rule map is one where the distribution is evenly spread across all gestures without bias towards one or two particular gestures.

In addition to the threshold, we also tested different selection criteria. We established three distinct criteria to select a gesture animation that matches the human gesture performed in the *TedTalk dataset*. The first criterion involved random selection above the cut-off threshold. The second criterion focused on maximum similarity above the cut-off threshold, and the third criterion involved maximum similarity without any cut-off threshold.

The cut-off threshold was empirically determined by analyzing a small sample from the *TedTalk dataset* with different similarities. We discovered that co-speech gestures generated from the first criterion, with 0.92 as the similarity check threshold, were closer to the ground truth.

Figure 4.6 depicts the distribution of gesture animations used. The distributions for the second and third criteria appear similar (with a threshold of 0.92), as both criteria focus on maximum similarity, albeit the second criterion employs a cut-off threshold. The third criterion generated many rules, albeit of lower quality. The second criterion provided quality, but the distribution was biased. The first criterion struck a balance, yielding a better distribution.

However, this approach highly depends on the diversity of gesture animations in the *Gesture dataset*. The resulting co-speech gestures will also be limited if the *Gesture dataset* is limited. We utilized 57 gesture animations, but the results will improve with the addition of more gestures.

4.7.2 Improving Rule Matching

Semantic mismatches between spoken words and co-gestures are a well-known issue of data-driven approaches. On the other hand, human experts created rule-based ap-

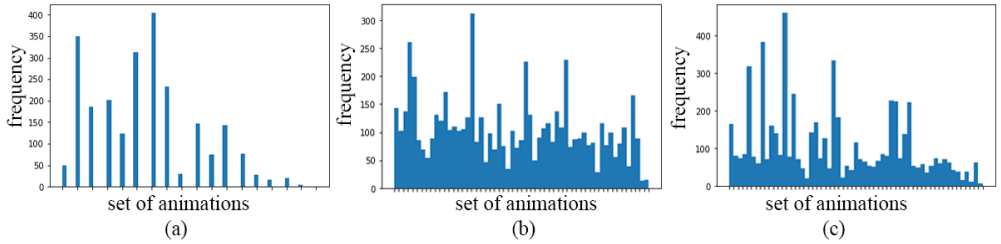


Figure 4.6: Gesture distribution with different criteria. x-axis is set of animations and y-axis is frequency of each animation. Left to Right: a) Random selection above cut-off threshold. b) Maximum similarity above cut-off threshold. c) Maximum similarity.

proaches that can generate semantically sounding co-speech gestures but lack the frequency of gesturing or repeating the same gestures. To preserve the semantic consistency and increase the frequency of gesturing, we focused on the rule-matching algorithm in the animation sequence generation procedure described in Section Animation sequence generation. To make semantic-aware matching, we used the GloVe word embedding matrix. We compared the word embedding method with the Levenshtein distance for the same task. As shown in table 4.1, we took a short passage from Geroge Orwell's novel "Animal Farm." The passage is "*Even when you have conquered him, do not adopt his vices. No animal must ever live in a house or sleep in a bed or wear clothes or drink alcohol or smoke tobacco or touch money or engage in trade. All the habits of Man are evil*". The result of the comparison indicates that GloVe preserved the semantics and context in a much better way in 1,6,7,9 then the simple string matching. GloVe performed worst on 4, and others are insignificant.

4.7.3 Synergistic Effects of Hybrid Rules

In Section 4.6.2, we conducted a comparison between co-speech gestures generated by human-crafted rules (*Manual*) and those generated by our automatically created rules (*Auto*). The comparison also incorporated a blend of these two rulesets (*Hybrid*), capitalizing on the benefits of the text-driven approach, namely the ease of expanding

1	Input: Even when you have conquered GloVe: you have invaded me in LD: see when you lose your
2	Input: him do not adopt his GloVe: that did not keep him LD: him to do so the
3	Input: vices No animal must ever GloVe: every animal they see including LD: species of animals out there
4	Input: live in a house or GloVe: them and live in a LD: live in a world of
5	Input: sleep in a bed or GloVe: a book in bed my LD: been in a cave for
6	Input: wear clothes or drink alcohol GloVe: or to wear anything or LD: we create our own echo
7	Input: or smoke tobacco or touch GloVe: or smoking or that people LD: go home today or to
8	Input: money or engage in trade GloVe: have to trade part of LD: us to engage in a
9	Input: All the habits of Man GloVe: habits of mind and these LD: or the arms of an

Table 4.1: GloVe and Levenshtein distance(LD) comparison

rules.

In the objective evaluations, both *Manual* and *Hybrid* generated a more diverse range of gestures compared to *Auto*, with a higher frequency of gesturing. However, compared to *Manual*, *Hybrid* appeared to generate a more biased use of gestures. Interestingly, this biased yet diverse use of gestures produced synergistic effects in the subjective evaluation. Co-speech gestures produced by *Hybrid* were perceived as more timely and semantically accurate.

This might be because *Hybrid* rules incorporate the best features of both rulesets. *Manual* provides a limited but strong association of gestures with text, while *Auto* offers a comparatively weak but broad association of gestures. *Hybrid* prioritizes the strongly associated gestures from *Manual*.

As depicted in Figure 4.3, *Manual* generated a single gesture, while both *Auto* and *Hybrid* generated two gestures each. The gesture from *Manual* is a *negative* animation that strongly associates with the text. In the *Auto* sequence, the first animation is an *emphasize* animation, followed by a *you* animation. *Hybrid* combines the *negative* animation from *Manual* and the *you* animation from *Auto* to generate a superior animation sequence compared to the other methods.

4.8 Conclusion and future work

We have presented an innovative method for automating the generation of rule-based co-speech gesture mappings, leveraging a publicly available video dataset and predefined gesture animations. This approach synthesizes rules without the need for human expert intervention. By utilizing pre-trained word embeddings, our method effectively preserves the context of input, while remaining independent of any specific Text-to-Speech system. Furthermore, our method ensures a general time consistency of the animation sequence by considering the length of the generated audio.

Our evaluation demonstrates that our method can significantly enhance gesture

generation by augmenting the limited manual mappings currently in use. Despite its promise, our approach does have its limitations. A key constraint is the *Gesture dataset* itself, coupled with the exclusion of hand movements in selecting upper-body joints. Some animations in the *Gesture dataset* only vary in hand motion, leading to potential misclassification due to this exclusion.

Moreover, the large number of rules and word embeddings substantially increase memory and computation costs. This makes the method less feasible for deployment on mobile and wearable devices. However, cloud deployment could potentially mitigate these constraints [74].

Looking forward, future work could incorporate full-body pose estimation, including hand pose, in the video dataset, and perform sentiment analysis of the text. This could further enhance the user experience. By expanding the number of available gestures, mappings could evolve from text-to-single-animation to text-to-multiple-animations, taking into account the sentiment of the text. Another interesting avenue for exploration is the generation of personalized automatic rules, which would require the inclusion of more personality-related tags such as gender, age, and personality traits in the pose and text data.

In conclusion, our method represents a significant advancement in the automatic generation of a vast set of text-to-gesture rules derived from pose and text data, employing predefined gesture animations. The text-to-gesture rules generated by our approach have the potential to dramatically improve upon the limited manual mappings currently in use, paving the way for more dynamic and engaging virtual interactions.

Chapter 5

Cross-Lingual Text-to-Gesture Synthesis for Digital Humans: Generation and Evaluation[1]

5.1 Introduction

The use of digital humans in virtual environments for various applications, such as acting as guides in virtual museums[74] or as actors in pre-visualizations [17], is gaining popularity. Digital humans require co-speech gestures to appear natural [12], similar to how humans use speech and gesture to convey their viewpoints. Traditional methods for generating co-speech gestures involve labor-intensive motion capture and manual animation mapping. Rule-based animations offer control and flexibility but suffer from subjective expert definitions [3]. On the other hand, data-driven methods use probabilistic models and intensive and elaborate annotation schemes[4]. Recently there has been a significant increase in the use of end-to-end deep-learning techniques to generate co-speech gestures, but given the non-deterministic nature of co-speech gestures,

[1] This work was partly published in SIGGRAPH ASIA 2022 as Improving Co-speech gesture rule-map generation via wild pose matching with gesture units

the results are underwhelming so far[13].

In this study, we address the problem of generating co-speech gestures for digital humans by proposing a cross-lingual text-to-gesture synthesis method that leverages both English and Korean data. Our specific aims are to develop a more accurate gesture-matching system, diversify the output to create a more natural communication experience, and demonstrate the effectiveness of our approach through a user study.

While numerous studies have explored the generation of gestures using audio-to-gesture methods[75, 76], Audio-only modality systems generate better gesture flow than text-only methods but lack semantic understanding[13]. Text-only systems provide better output control and semantic awareness due to pre-trained LLMs [12]. However, these systems may face challenges in capturing the subtleties of human motion and may not generalize well across languages and cultures. To address these limitations, we focus on text-based input, which provides richer semantic information and aligns with the current trend of text-only chatbot technologies. This shift is particularly relevant as text-to-speech (TTS) systems continue improving, moving away from robotic-like voices to more natural and expressive human-like ones. As a result, there is a growing need for digital humans to exhibit appropriate gestures that complement the sophisticated TTS output, thereby enhancing the overall user experience. By focusing on text-based input, our method enables digital humans to generate more contextually relevant and meaningful gestures, going beyond the simple rhythm provided by audio data. This enhances the non-verbal communication capabilities of digital humans and provides a more engaging and immersive experience for users.

The quality of co-speech gestures generated by any method strongly depends on the data used [13]. Methods that use motion capture data have better motion flow[16, 31, 32] than wild data from videos[**?**, 36]. However, the quality can decrease for systems trained on motion capture data when the input is semantically different from the training data [32]. Public videos provide diverse text corpus but may have motions of lower

quality, resulting in the model learning averaged motions[12]. To address these issues, a hybrid approach was proposed that combines diverse text data from public videos and clean motion data from motion capture, using intermediate models via deep learning for text and motion matching[15].

Regarding language, most co-speech gesture-generation methods use English data[75, 76]. Other languages are rarely explored in this context, presenting further research opportunities. Studies have shown that language and culture influence gesture production[77, 78, 79], suggesting language-specific models are needed. Our study aims to address this gap by developing a cross-lingual text-to-gesture synthesis method that leverages both English and Korean data.

Ali et al. [12] proposed a method to address the issue of subjectivity in rule-based animations. Their approach involved using transcript and 2D pose data extracted from monologue-style public videos of human presenters. They then used a set of predefined gesture units as sliding windows over the 2D pose sequences and extracted corresponding text for the best matches. This process resulted in a comprehensive text-to-gesture mapping. In addition, the authors used the mean cosine similarity for the pose-to-gesture matching, which performs poorly on temporal misalignment and noise. However, this novel approach can help mitigate the subjectivity problem of rule-based animations. The gesture units were 54, and the rules were 84k. We used this system as a baseline system, seeking to improve upon it by incorporating cross-lingual data and developing a more accurate gesture matching system.

Our approach is based on GestureCLR[15] and automatic text-to-gesture rule generation[12] with cross-lingual data, specifically videos of English speakers and motion capture (mocap) data of Korean speakers, to create an automatic rule-map that can be used with input data in both English and Korean. To achieve this, we first needed to identify gesture units and develop a more effective method for matching 2D pose data to 3D mocap gesture units.

To generate gesture units from the mocap data, we designed an expert system that increased the number of gestures to 2035. This expert system allowed us to create a more nuanced set of gestures that could capture the subtleties of human motion. Additionally, we used deep contrastive learning to train our model to match 210k rules from a set of 350k pose-text pairs. This method helped us to improve gesture matching and generate more accurate rule-maps.

We employed gesture clustering techniques to diversify the output further and avoid excessive gesture repetition. This allowed us to slightly change the output every time for the same input while maintaining relevance.

Recognizing the need for multilingual capabilities in the digital human domain, we integrated translation technology into our co-speech text-to-gesture generation method. This allows the core system, built in English, to easily adapt to support multiple languages without compromising user experience.

Since the rule map is in English, we needed to translate Korean input data into English for rule matching at runtime. Despite this additional step, our study with Korean speakers demonstrates that users interacting with digital humans in their native language experienced no degradation in the quality of the digital human's gestures compared to those in English.

The user study evaluated the effectiveness of co-speech gestures generated by different methods on a digital human's naturalness, human-likeness, time consistency, semantic consistency, and social presence. Participants watched 20 videos with different combinations of speech languages (Korean and English) and gesture conditions (No Gesture, Ali et al. [12], and Ours) and rated their experience on a 7-point Likert scale. The results showed that the proposed method generated more natural, human-like, temporally and semantically consistent, and socially present co-speech gestures compared to the other methods. There were no significant differences in evaluations when the speech-language was Korean or English, suggesting that the perception of

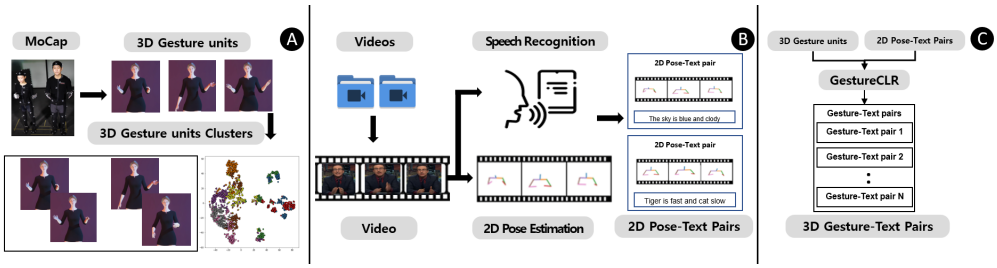


Figure 5.1: System overview. A) Gesture unit extraction and gesture clustering. B) 2D Pose-Text pairs extraction C) Rulemap generation by matching pose and gesture units with GestureCLR.

co-speech gestures can be relatively consistent across different languages.

Our approach demonstrates the potential of using cross-lingual data and expert systems to develop automatic rule-maps for gesture generation. By improving the accuracy of gesture matching and diversifying the output, we can create a more natural and human-like communication experience for users. Furthermore, our method showcases the ability to accommodate multiple languages without compromising the overall user experience, making it a promising solution for the diverse global audience interacting with digital humans. Future research can explore incorporating more languages and developing a more comprehensive understanding of cultural nuances in gesture generation to enhance the effectiveness of our method further.

In the following sections, we present the approach explaining the system, implementation, and details of the user study.

5.2 Approach

Our approach breaks the problem into three parts as shown in Fig.5.1, i.e., 1) Gesture units and clustering, 2) 2D Pose-Text pairs extraction 3) Rulemap generation with GestureCLR. We choose videos on a wide variety of topics with different presenters. The topic ranges from current affairs talk show to lectures in universities.

5.2.1 Gesture units and clustering

In this study, we defined a gesture unit as a short sequence of upper body motion during speech. To extract gesture units, we obtained two hours of motion capture data of a male and a female discussing life experiences and general topics. The conversation type is irrelevant, as we only need the body motion to create gesture units, but diverse conversation topics help ensure a diverse set of gestures.

As manual extraction of gesture units can be time-consuming and subjective, we used a set of rules, as described in Algorithm 3, to extract the units automatically from any motion capture data for co-speech gestures. The first rule was that the unit length should be between 2 and 3 seconds, corresponding to a minimum of 30 frames and a maximum of 45 frames at 15 fps. The second rule required that the starting and ending positions of the unit should have a minimum Euclidean distance to ensure similar poses at the start and end of the gesture. Finally, we defined a third rule for optimal variance, which is expert-dependent and determined through testing to find the best value for the given dataset. A high variance would yield fewer but highly dynamic gesture units, while a low variance would produce more but less dynamic units, including units with slight motion.

We extracted 2035 new gesture units using these rules (Algorithm 3), with 1025 for male and 1010 for female styles, from the two hours of data. We observed that gestures could be clustered to improve output diversity and relevance. We used the Bisecting K-Means algorithm to cluster the gesture units, eliminating the need for TopK sampling rules. Now, sampling is done from the cluster. Our findings suggest that these automatic extraction rules and gesture clustering techniques can significantly reduce the time and effort required to extract gesture units and improve the diversity and relevance of generated gestures.

Algorithm 3: Extracting motion clips from a motion sequence

Input: motion sequence

Output: list of extracted motion clips

```
while motion sequence has more than 3 seconds do
    for each possible starting position s do
        for each possible ending position e do
            if 30 ≤ (e − s) ≤ 45 at 15 fps and Euclidean distance between s
              and e is minimal then
                clip ← extract motion clip from s to e; if clip has optimal
                  variance then
                    add clip to the list of extracted motion clips;
                end
            end
        end
    end
    Remove the longest clip from the motion sequence;
end
```

5.2.2 2D Pose-Text pairs extraction

In the field of natural language processing, grounding refers to the process of connecting language to the real world. Grounding is the process of linking words or phrases in natural language to specific entities or concepts in the real world, such as physical objects, locations, or actions. In the context of analyzing public monologue-based videos with human presenters, we obtained timestamped speech transcripts and 2D human pose data with OpenPose[80]. The pose sequences were then processed to make 3-second chunks (Text, Pose) pairs. However, the pose sequences obtained from OpenPose were usually noisy and not directly usable. Some common noise sources in pose

data include incomplete or missing pose data, drift, jitter, and false positives.

To address the limitations of OpenPose, such as the lack of person tracking, we used facial recognition to identify the person of interest in the videos. A dictionary of faces was built to identify the face present in the majority of frames. By tracking the person of interest's head position, we could create segments where the person was continuously present. Speech recognition was then employed to find segments containing speech, and the text was extracted from those segments. Segments without speech were discarded.

Pose estimation was run on the segments obtained from the previous step, and only those poses whose head position matched the head position from facial recognition were selected. The noisy pose sequences were not cleaned up, but instead, a model was trained to link them with predefined animations, as detailed in the section titled "Rulemap Generation with GestureCLR."

The process of extracting the 2D Pose-Text pairs can be summarized as follows:

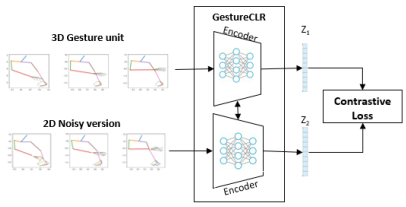
1. Get a set of videos where ideally one person is the main subject, such as a talk show host or lecturer delivering lectures.
2. Take a video from the set of videos.
3. Run facial recognition on the video and build a dictionary of faces. Identify the face which is present in the majority of frames as the person of interest.
4. Track the person of interest's head position and create segments where the person is continuously present.
5. Run speech recognition to find the segments containing speech. Obtain text from these segments and discard segments without speech.
6. Run pose estimation on the segments obtained from the previous step, selecting only those poses whose head position matched the head position from facial

recognition.

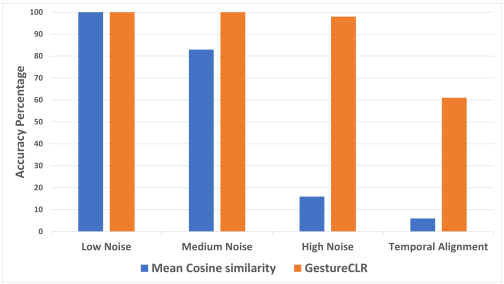
7. Save the pose-text pairs from the given segment and continue with other segments.

8. Repeat the process for the remaining videos in the set.

By following this process, we successfully built a dataset of accurate and usable (Text, Pose) pairs extracted from public monologue-based videos.



(a) gestureCLR model.



(b) Comparative accuracy performance of GestureCLR and Mean Cosine Similarity.

Figure 5.2: gestureCLR model and evaluation

5.2.3 Rulemap Generation with GestureCLR

In this study, we developed GestureCLR, a deep contrastive learning model based on the SimCLR architecture proposed by Chen et al. [81], for recognizing gestures from noisy and varied input data.

We trained our model on motion capture data from the Talking with hands dataset [82]. We first processed the data to create sequence units of 2-3 seconds, or 30-45 frames, as the frames per second are set to 15. The input included a clean 3D unit and a 2D version of the same unit. The 2D version was augmented with Gaussian noise and temporal shifts to mimic the variability and noise present in real-world 2D pose data. Gaussian noise was generated with a variable variance of 0.001, 0.01, and 0.1

for low, medium, and high noise. Temporal shifts were introduced by creating two sequences of 45 frames, one filled with the mean pose of the original sequence and the other with zero poses. Contiguous segments of 30 frames were randomly selected from the original sequence and inserted into the new sequences at a random position between 1 and 15, contiguously occupying the next 30 frames.

Our GestureCLR model is built on top of a Transformer encoder architecture [83]. The model takes as input the 3D and 2D body gesture data and learns to map them to a latent space, where the representations of corresponding 3D and 2D gestures have high similarity, and non-corresponding pairs have low similarity. The main components of the model are a Positional Encoding module, a Transformer Encoder, and a fully connected layer followed by a batch normalization and a leaky ReLU activation function.

The Positional Encoding module generates sinusoidal positional encodings for the input data, enabling the model to capture the sequential nature of the gesture data. The Transformer Encoder consists of 3 layers, each having 5 self-attention heads and a feedforward network with a hidden dimension of 512. The input and output dimensions of the encoder are set to 150. The fully connected layer maps the output of the Transformer Encoder to a latent space of dimension 10. We experimented with latent space dimensions of 10, 16, 32, and 64 and found that a dimension of 10 resulted in the best accuracy.

We trained the model with a contrastive learning approach using the Normalized Temperature-Scaled Cross Entropy (NT-Xent) loss. The loss function is defined as:

$$L_{i,j} = -\log\left(\frac{\exp(\text{cos\_sim}_{i,j})}{\sum_k \exp(\text{cos\_sim}_{i,k})}\right) \tag{5.1}$$

where $\text{cos\_sim}_{i,j}$ is the cosine similarity between the i-th 2D motion sequence and the j-th 3D motion sequence, computed as:

$$\text{cos\_sim}_{i,j} = \frac{2D_i \cdot 3D_j}{\|2D_i\|_2 \|3D_j\|_2} / T \tag{5.2}$$

In these equations:

- $2D_i$ and $3D_j$ are the latent representations of the i-th 2D motion sequence and the j-th 3D motion sequence, respectively. - $\cdot$ denotes the dot product. - $\| \cdot \|_2$ denotes the L2 norm. - T is the temperature hyperparameter, controlling the concentration of the distribution. - The index k runs over all possible negative samples.

The final loss is the mean over all pairs (i, j):

$$L = \frac{1}{N^2} \sum_{i,j} L_{i,j} \tag{5.3}$$

where N is the total number of samples in the batch.

We used the AdamW optimizer [84] with a Cosine Annealing learning rate scheduler for training. The model was trained with a learning rate of 5×10^{-4}, a weight decay of 10^{-4}, and a maximum number of epochs set to 1000. The training and validation batch sizes were set to 512.

To implement the GestureCLR model and training pipeline, we used PyTorch Lightning [85], a high-level library for PyTorch that simplifies the training process. The model checkpoints were saved periodically based on the validation loss, and the learning rate was logged during training.

Using the GestureCLR model, we can match 2D poses from OpenPose with 3D poses from motion capture. It can also be used as a motion encoder to encode gesture units and cluster them using the latent representations obtained from the GestureCLR.

Finally, we applied our model to poses from (Text, Pose) pairs and gesture units from our mocap data to obtain (Text, Gesture) pairs. To obtain an embedding for the text input, we used SentenceBERT [86]. This gave us (Text_{Emb}, Gesture) pairs as rulebase.

5.3 Implementation

We employed the server-client model to implement our gesture system, with Unity as the platform for visualizing the gestures. Our gesture units were initially in BVH format, but as described in the section ”Gesture units and clustering,” we applied an algorithm on the BVH files to extract the start and end of each gesture unit and saved the information in a file. We then used Blender to convert the BVH files to FBX format. The FBX files were loaded into Unity, and a script used the gesture unit information from the file to create clips in the FBX, which were then loaded into the Unity animator.

We chose 6-gram tokenization (6-word chunks) for the text input based on the assumption of 2.5 words per second for an average text-to-speech system. As our gestures have a minimum duration of 2 seconds and a maximum of 3 seconds, this segmentation accurately represents the text. The tokens were sent to the server's gesture system, which converted them into embeddings using SentenceBERT.

To find the best match in the rule map, we used cosine similarity on text embeddings and ranked them to identify the gesture unit ID with the lowest distance. As described in the section ”Rulemap Generation with GestureCLR,” this process was repeated for each token, and the resulting IDs were returned to the Unity animator. The implementation guarantees that an input text will always match with some text in the rules, even if the matching distance might be low in some cases. A minimum threshold can be applied if needed, and an idle gesture can be played in instances with low matching distances.

We employ clustering to ensure that the generated gestures are diverse and do not become repetitive, as explained in the ”Gesture units and clustering” section. The rules have a cluster ID instead of a gesture ID, and given a cluster ID, we randomly pick an animation. This approach balances the relevance of the animation with the input text, creating a slightly different animation sequence each time while maintaining overall similarity.

We also addressed the timing and synchronization between the text and the corresponding gestures. The text is segmented into 6-word chunks, which helps handle the timing issue effectively. The runtime of each gesture unit is calculated by dividing the total length of the audio by the number of gestures to be played.

Overall, our gesture system relies on a combination of server-side and client-side technologies to provide a seamless experience for users. Using a rule map and SentenceBERT, we can accurately match text inputs with appropriate gesture units, allowing for clear and effective communication through the Unity platform.

5.4 Evaluation

5.4.1 Objective Evaluation

To assess the effectiveness of our model, we conducted experiments using a test set of 500 randomly selected gesture units extracted from 3D motion capture data. We created 2D noisy versions of the gesture units using Gaussian noise with variable variance levels of 0.001, 0.01, and 0.1 for low, medium, and high noise, respectively. Additionally, we created a temporally shifted 2D version by randomly selecting contiguous segments of frames and inserting them into new sequences.

We evaluated our model's output by comparing it with the frame-by-frame mean cosine similarity metric proposed by Ali et al. [12]. While this metric has limitations in that it is not robust for noise and temporal shifts, our model outperformed the metric in all scenarios. Specifically, for low noise, the accuracy of our model and the metric was 100%, whereas, for medium noise, our model achieved 100% accuracy compared to 82% for the metric. For high noise, our model achieved an accuracy of 97% compared to only 18% for the metric. Finally, for temporal shifts, our model achieved an accuracy of 62% compared to only 6% for the metric. The results are presented in Figure 5.2b.

It is important to note that these results may be limited by the specific test set used in the experiment. Further evaluation using different data types is necessary to fully as-

sess the model's effectiveness. Additionally, it is worth considering other evaluation metrics in future work to provide a more comprehensive assessment of the model's performance. Finally, randomly selecting contiguous segments of frames for the temporally shifted 2D version may not fully capture the variability present in real-world data, which is a potential source of bias.

5.4.2 User Study

5.4.3 Participants and Experimental Settings

A total of 51 Korean university students (33 males, 18 females) aged 20–27 (M=23.7, SD=1.7) participated in the experiment. The study was approved by the Institutional Review Board (KIST-202304-HR-003) and all participants gave consent for their participation.

The experiment took place in a university classroom, where all participants were sitting and watching a projector screen located at the front of the classroom simultaneously while the video stimuli were being presented at the projector screen (Fig. 5.3). A young female avatar was used to generate gesture videos, as young or female embodied agents have been preferred over old or male agents [87, 88].

5.4.4 Experimental Design and Procedure

Participants were asked to watch a total of 20 videos (4 gesture conditions $\times$ 5 speech contents) in random order. Five speech contents were retrieved from some of the most popular TED talks of all time [89]. Each speech content consisted of 49–83 words and was around 3–5 sentences long. Four gesture conditions were selected to answer following three research questions:

- Is the digital human with co-speech gestures generated by the proposed method perceived better than the digital human without co-speech gestures?
- Is the digital human with co-speech gestures generated by the proposed method

Figure 5.3: The experimental settings

perceived better than the digital human with co-speech gestures generated by the previous counterpart [12]?

- Does the language affect how the digital human with co-speech gestures generated by the proposed method is perceived?

Therefore, gesture conditions considered in this study were composed of combinations between three gesture generation methods (None, Ali et al. (2020) [12], and Ours) and two speech languages (Korean and English). Fig. 5.4 shows the four gesture conditions evaluated in this study: *NoneKor* (None & Korean), *AliKor* (Ali et al. (2020) [12] & Korean), *OursKor* (Ours & Korean), and *OursEng* (Ours & English). All 20 videos can be seen from the following link: https://bit.ly/3Ma6vyR

	NoneKor	***AliKor***	***OursKor***	***OursEng***
Gesture	None	Ali et al., 2020	Ours	Ours
Language	Korean	Korean	Korean	English

Figure 5.4: Four gesture conditions evaluated in this study

After watching each video, participants were asked to give ratings on a 7-point Likert scale (1=strongly disagree, 7=strongly agree) for five evaluation items to subjectively assess the user experience of each gesture condition. The five subjective evaluation items were naturalness ("Digital human was natural"), human-likeness ("Digital human was human-like"), time consistency ("Gesture timing was matched to speech"), semantic consistency ("Gesture was matched to speech content"), and social presence

("I felt like digital human was 'real' and 'there' "). These items have been frequently used in evaluating co-speech gestures of embodied agents [90, 91].

5.4.5 Data Analysis

The first speech content was considered as practice thus excluded, so a total of 4 gesture conditions $\times$ 4 speech contents $\times$ 51 participants = 816 trials was arranged for repeated-measures analysis of variance (RM-ANOVA) and Bonferroni post-hoc tests. The Shapiro-Wilk test for normality and inspection of Q-Q plots found no measures were severely deviated from the normal distribution. The degree of freedom was corrected with Greenhouse-Geisser correction if the p-value of Mauchly's test was equal to or less than 0.05. R 4.2.2 and "*rstatix*" R package were used to conduct all data processing and statistical analyses at a significance level of 0.05, and the effect size in terms of generalized eta-squared (η^2_G) [92] was further calculated to check practical significance.

5.4.6 Results and Discussion

Fig. 5.5 shows the subjective evaluation ratings for each of five investigated items. The significant main effect of gesture condition was found in all evaluation items: naturalness ($F_{1.73,86.55} = 39.42, p < 0.001, \eta^2_G = 0.146$), human-likeness ($F_{1.69,84.25} = 24.68, p < 0.001, \eta^2_G = 0.088$), time consistency ($F_{1.88,93.97} = 56.45, p < 0.001, \eta^2_G = 0.274$), semantic consistency ($F_{2.24,112.22} = 47.71, p < 0.001, \eta^2_G = 0.292$), and social presence ($F_{1.89,94.26} = 23.24, p < 0.001, \eta^2_G = 0.097$).

Regarding the pairwise comparison, we only highlight the results of comparisons between *NoneKor* and *AliKor*, *AliKor* and *OursKor*, and *OursKor* and *OursEng* here for better clarity. According to the post-hoc tests, the same pattern was found across all evaluation items. *OursKor* ($M \pm SD$; Naturalness: 3.98 ± 1.13, Human-likeness: 3.79 ± 1.27, Time consistency: 4.14 ± 1.12, Semantic consistency: 3.94 ± 1.16, So-

cial presence: 3.34 ± 1.17) had significantly higher ratings than *AliKor* (Naturalness: 3.56 ± 1.15, Human-likeness: 3.48 ± 1.26, Time consistency: 3.54 ± 1.16, Semantic consistency: 3.17 ± 0.99, Social presence: 2.97 ± 1.07), which had significantly higher ratings compared to *NoneKor* (Naturalness: 2.83 ± 1.34, Human-likeness: 2.87 ± 1.36, Time consistency: 2.50 ± 1.29, Semantic consistency: 2.26 ± 1.12, Social presence: 2.46 ± 1.10), whereas no significant differences (Naturalness: $p = 0.392$, Human-likeness: $p = 0.612$, Time consistency: $p = 0.170$, Semantic consistency: $p = 1.000$, Social presence: $p = 1.000$) were found between *OursKor* and *OursEng* (Naturalness: 4.10 ± 1.24, Human-likeness: 3.91 ± 1.35, Time consistency: 4.35 ± 1.15, Semantic consistency: 3.98 ± 1.14, Social presence: 3.40 ± 1.24).

Table 5.1: Subjective evaluation ratings

Evaluation Item	**NoneKor**	**AliKor**	**OursKor**	**OursEng**
Naturalness	2.83 ± 1.34	3.56 ± 1.15	$\mathbf{3.98 \pm 1.13}$	4.10 ± 1.24
Human-likeness	2.87 ± 1.36	3.48 ± 1.26	3.79 ± 1.27	$\mathbf{3.91 \pm 1.35}$
Time consistency	2.50 ± 1.29	3.54 ± 1.16	4.14 ± 1.12	$\mathbf{4.35 \pm 1.15}$
Semantic consistency	2.26 ± 1.12	3.17 ± 0.99	3.94 ± 1.16	$\mathbf{3.98 \pm 1.14}$
Social presence	2.46 ± 1.10	2.97 ± 1.07	3.34 ± 1.17	$\mathbf{3.40 \pm 1.24}$

The results indicate that the digital human with co-speech gestures generated by the proposed method was perceived more natural, more human-like, more temporally and semantically consistent, and more socially present than the digital human without co-speech gestures or with co-speech gestures generated by the previous counterpart [12]. The fact that the existence of co-speech gestures helps digital human to look more alive and engaging has been supported by numerous studies [93, 94, 95, 96]. Our results align with existing studies, emphasizing the significance of naturalness and human-likeness in perceptions of digital humans. Studies reported that lifelike co-speech gestures that closely mimic human gestures could enhance perception of the digital human's human-like characteristics [97, 98]. Our proposed methods could gen-

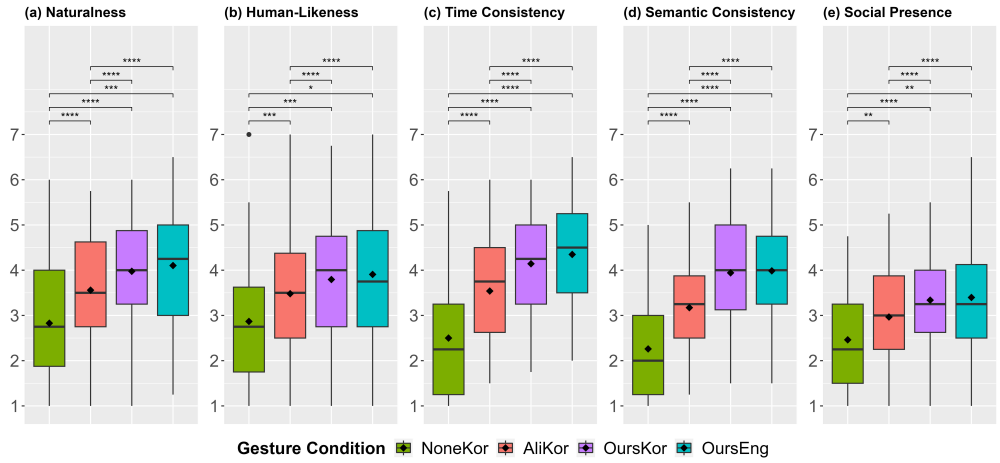


Figure 5.5: Ratings for subjective evaluation by gesture condition for each item. *Note*: Diamond symbols indicate mean ratings, and circular symbols indicate outliers outside of the interquartile range. Asterisks indicate the significance of Bonferroni post-hoc tests ($p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}, p < 0.0001^{****}$)

erate gestures with greater time and semantic consistency, which might have affected other evaluation aspects. Coherence between gestures and verbal timing and content play a significant role in improving the perceived naturalness and effectiveness of digital humans [99, 100, 101]. The perception of social presence, or the degree to which a virtual agent is perceived as a social entity, was also found to be influenced by co-speech gestures. Our results are in agreement with previous studies suggesting that gestures contribute to the perception of social presence in digital humans [102, 103]. The proposed method (*OursKor*) could generate more suitable co-speech gestures than others (*NoneKor* & *AliKor*), thus led to an increased sense of social presence.

In addition, the results revealed no significant differences between the evaluations when the speech language was Korean compared to when it was English. Our results suggest that the perception of co-speech gestures in digital humans can be relatively consistent across different languages, which is in line with some previous research [104, 105]. These studies argued that gestures are a universal aspect of human com-

munication and convey meaning irrespective of the specific language spoken, which could explain the lack of significant differences in our study. However, it is important to note that cultural factors can influence the perception of gestures. Different cultures may have different expectations and preferences for co-speech gestures [78, 79], which could affect the evaluations of digital humans. Our findings might indicate that the cultural background of the participants was not significantly different or diverse enough to reveal any potential effects of language on the evaluation of co-speech gestures. Another factor that might have influenced the result is the language proficiency of participants. People with higher language proficiency may have a better understanding of the nuances of the spoken language, enabling them to more accurately assess the appropriateness and naturalness of the accompanying gestures [106]. Nevertheless, it is unclear whether the relatively lower language proficiency in English compared to Korean of participants in this study positively or negatively affected the evaluation results.

Some limitations should be acknowledged. First, the distance and angle between each participant and the video were not consistent since all participants watched the video in a single room at once, although we confirmed all participants including the ones sitting at the corner of the room could clearly perceive the details of the entire gestures. Second, a single question was used to evaluate each item in this study, but a more comprehensive evaluation can be performed in future works by adding multiple questions with slightly different perspectives.

5.5 Conclusion

In conclusion, our approach highlights the importance of utilizing text-based input to generate more contextually relevant and meaningful co-speech gestures for digital humans in virtual environments. By employing cross-lingual data, expert systems, and deep contrastive learning, we have developed an automatic rule-map that can generate

more natural, human-like gestures in multiple languages. The proposed method has been shown to outperform existing methods in terms of naturalness, human-likeness, temporal and semantic consistency, and social presence. Furthermore, our user study indicates that the perception of co-speech gestures remains relatively consistent across different languages, demonstrating the potential for creating a more engaging and immersive experience for users.

Future research should continue to explore the use of diverse text data and clean motion data from various sources to improve text and motion matching in gesture generation models. Additionally, expanding the supported languages and investigating the effects of cultural differences on gesture production will further enhance the applicability and effectiveness of digital humans in diverse applications. Overall, our approach takes a significant step towards creating more engaging, natural, and immersive communication experiences with digital humans in virtual environments.

Chapter 6

Rapid and Memory-Efficient Gesture Generation with Zero-Shot Speaker Style Adaptation

6.1 Introduction

Co-speech gestures, fundamental to human communication, play a pivotal role in the realm of Human-Computer Interaction (HCI). They are particularly indispensable in establishing engaging and authentic interactions between humans and virtual agents [74]. However, real-time generation of these gestures, especially for virtual or digital humans, presents an intricate task laden with computational and personalization challenges. Our study introduces ConGRets, a novel framework for co-speech gesture generation, which addresses these constraints and offers an efficient, scalable, and personalized solution.

6.1.1 Background and Importance of the Study

With the evolution of Virtual Reality (VR), Augmented Reality (AR), and pre-visualization (previs) technologies, the demand for digital humans in interactive experiences has

surged. These digital entities are not merely static models, they are expected to exhibit human-like behavior, including co-speech gestures, the synchronized movement of hands, arms, and body during speech [107, 108, 109]. These gestures significantly contribute to the overall communication effectiveness, augmenting the perceived naturalness of digital humans [43]. Despite its importance, the practical application of gesture generation methods remains challenging due to their large memory requirements, slow query speeds, and the intricacy of personalization.

6.1.2 Speaker Style Adaption Intuition

The crux of this research lies in the concept of speaker style adoption. To grasp this notion, let's revisit the definition of 'speaker style'. Our methodologies are founded upon two distinct sets of gestures: 2D poses and 3D gesture units.

Our system, by design, formulates a distinctive rule map for each individual. Text data is affiliated with the 2D pose derived from the speaker's video content. On the other hand, the text associated with the 3D gesture units is excised, leaving only the animation elements intact. The intriguing part unfolds when 3D gestures are transposed onto the 2D pose, and final text-gesture pairs are constructed.

The reference video influences the decision of which gesture to enact, while the 3D gesture units inform the execution technique. The visual style can be traced back to the 3D gesture units, and conceptually, it originates from the 2D video. Comprehending this conceptual origin can be challenging; thus, the visual style is a more relatable reference point.

The gesture encoder, a key system component, is intrinsically style-invariant. It has been trained specifically to match 2D poses with 3D animations. It has learned to encode the animations effectively and without bias to any particular style.

Having established the role of the gesture encoder, it's essential to understand how it interacts with the language model in our system. The alignment of the language

model with the gesture encoder is pivotal for generating embeddings from the language model because gestures are encoded using the gesture encoder. The alignment is supposed to enable the language model to generate similar embeddings as the gesture encoder. This alignment enables a search for the relevant gesture by measuring the similarity between the embedding produced by the language model and the set of gesture embeddings from a particular speaker.

The model learns the specific style of the speaker in focus because it's been trained on the rule map of that person. Mixing all the rule bases into one would confuse the model and produce averaged results. Given this design, each new introduction of speaker gesture data requires training a new model. A model trained on one speaker's data cannot generate dependable embeddings for another speaker due to the absence of a reference for the latter's idiosyncrasies. However, to create a more generalized model, it must be capable of differentiating between speakers. An alternative proposition to aid this differentiation is the employment of a speaker style encoder, which serves as a summary of the speaker's style.

The speaker-style encoder can assimilate all possible gestures of a person simultaneously, subsequently producing a single latent vector. This vector can be supplied to the language model, which utilizes it to align its output latent space with the gesture encoder, eliminating the necessity to retrain models when new gesture data is added.

An essential point to note is the difference in the operational functionality of the gesture and speaker encoders. While the gesture encoder is designed to encode only one gesture at a time, the speaker encoder can assimilate all the gestures and output a single latent vector.

6.1.3 Objectives and Contributions

Our study aims to tackle these challenges by introducing ConGRets, a lightweight, low-latency, and scalable co-speech gesture generation system. The framework is de-

signed to generate gestures that are not only in sync with the spoken text but are also personalized according to each speaker's unique motion style, allowing for zero-shot speaker adaptation. Our contributions to the field of HCI and gesture generation are three-fold:

1. We propose a novel contrastive learning and transformer-based model for co-speech gesture generation that significantly improves query speeds and reduces memory requirements.

2. We present a unique solution for zero-shot speaker adaptation, allowing the system to generate gestures for new speakers without additional training.

3. We demonstrate the practicality and efficiency of ConGRets through rigorous evaluation and comparison with existing methods, highlighting its potential for broad applications in VR, AR, and previs generation.

Drawing inspiration from previous studies in this field, our study takes a significant stride in addressing the challenge of co-speech gesture generation. By presenting ConGRets, we hope to advance the field of HCI, opening new doors for the development of expressive and interactive digital humans.

In this paper, we first provide an overview of related work in Section 2, discussing the strengths and limitations of existing rule-based gesture generation systems and the potential of contrastive learning in the field. Section 3 details our methodology, including an overview of the GestureCLR model, the introduction of our fine-tuned BERT model, and our approach to zero-shot speaker adaptation. We describe our experimental setup and evaluation metrics in Section 4. The results and discussions, including a comparison of ConGRets with the baseline method, are presented in Section 5. Finally, we discuss our work's potential applications and future directions in Section 6 and conclude in Section 7.

6.2 Methodology

6.2.1 GestureCLR: Overview and Adaptation

Ali et al. [15] developed GestureCLR, a deep contrastive learning model to recognize gestures from noisy and diverse input data. Built on the SimCLR architecture [81], GestureCLR was trained using motion capture data extracted from the Talking with hands dataset [82]. This data, segmented into 2-3 second units, was processed and subjected to Gaussian noise and temporal shifts to imitate the variability found in real-world 2D pose data.

The GestureCLR model employs a Transformer encoder architecture [83] to map 3D and 2D gesture data into a common latent space. Corresponding 3D and 2D gestures are closely positioned in this space, while non-corresponding pairs are far apart. The model consists of a Positional Encoding module, a Transformer Encoder, and a fully connected layer with batch normalization and a leaky ReLU activation function. The model was found to perform optimally with a latent space dimension of 10.

The model's training process involves a contrastive learning approach utilizing Normalized Temperature-Scaled Cross Entropy (NT-Xent) loss. AdamW optimizer [84] was employed for optimization, with a Cosine Annealing learning rate scheduler. The training pipeline was implemented using PyTorch Lightning [85], a high-level library that simplifies the training process.

The GestureCLR model has two primary applications: it can match 2D poses with 3D poses from motion capture data and can also serve as a motion encoder to cluster gesture units using the obtained latent representations. The model was further used to create (Text, Gesture) pairs from pose data.

During ConGRets training, GestureCLR was kept frozen.

6.2.2 Speaker Style Encoder Model Architecture and Training

The Speaker Style Encoder model Fig6.1 is architected to encode motion data with particular emphasis on individual speaker styles. The backbone of the architecture is the Transformer, a model that leverages self-attention mechanisms to encapsulate dependencies between elements in a sequence, irrespective of their positional distance.

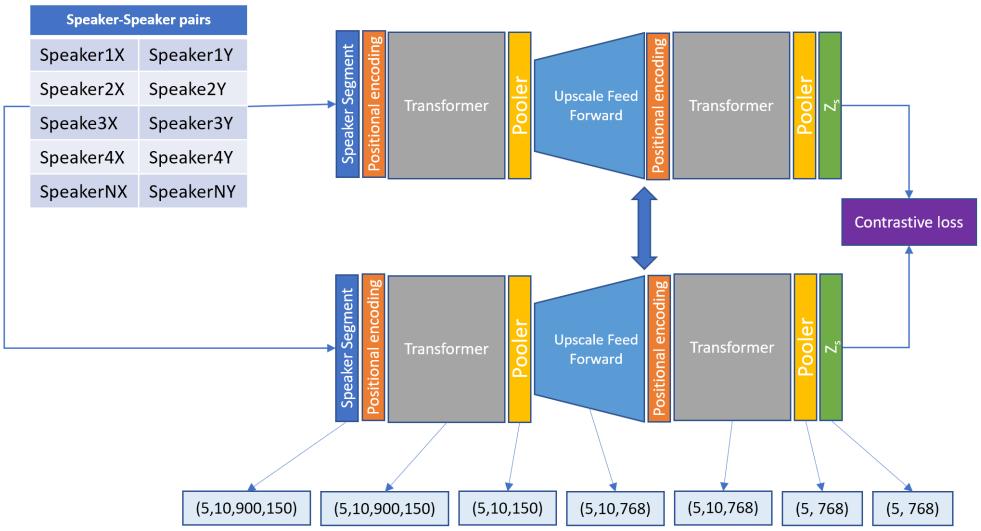


Figure 6.1: Speaker Style encoder model

6.2.2.1 Data Preparation and Preprocessing

The model employs a custom dataset derived from preprocessed motion data. Each data batch is dynamically sampled, ensuring a diverse representation of the data distribution. The process involves selecting random non-overlapping sets of motion samples, subsequently converted into tensors, thus making them consumable for the model.

6.2.2.2 Positional Encoding

The model incorporates positional encoding to imbibe information about the position of each data point in the sequence. Given the Transformer's native lack of positional

awareness, this encoding serves to provide an implicit sense of sequence order.

6.2.2.3 Model Architecture

The model is designed with a dual Transformer Encoder setup, one dedicated for processing spatial data and the other for temporal data. Each channel of the input data is processed independently. The spatial Transformer Encoder, consisting of 4 layers , is applied to the positional encoded data. The output is subsequently pooled and upscaled using a feed-forward layer with a hidden size of 768. The resulting representations are concatenated and passed through the temporal Transformer Encoder. The final output is derived through mean pooling and dropout (with a rate of 0.1) applied to the temporal Transformer Encoder's output.

6.2.2.4 Model Training

We trained the model with a contrastive learning approach using the Normalized Temperature-Scaled Cross Entropy (NT-Xent) loss. The loss function is defined as:

$$L_{i,j} = -\log\left(\frac{\exp(\text{cos\_sim}_{i,j})}{\sum_k \exp(\text{cos\_sim}_{i,k})}\right) \tag{6.1}$$

where $\text{cos\_sim}_{i,j}$ is the cosine similarity between the i-th motion sequence and the j-th motion sequence, computed as:

$$\text{cos\_sim}_{i,j} = \frac{u_i \cdot v_j}{\|u_i\|_2 \|v_j\|_2}/T \tag{6.2}$$

In these equations:

- u_i and v_j are the latent representations of the i-th motion sequence and the j-th motion sequence, respectively. - $\cdot$ denotes the dot product. - $\|\cdot\|_2$ denotes the L2 norm. - T is the temperature hyperparameter, controlling the concentration of the distribution. - The index k runs over all possible negative samples.

The final loss is the mean over all pairs (i, j):

$$L = \frac{1}{N^2} \sum_{i,j} L_{i,j} \tag{6.3}$$

where N is the total number of samples in the batch.

This function aims to maximize the similarity of positive pairs and minimize the similarity of negative pairs, making it particularly apt for learning distinctive speaker styles. The model's forward pass, training, and validation steps are defined in accordance with this loss function.

6.2.2.5 Optimization and Learning Rate Scheduling

The model is optimized using the AdamW optimizer, which incorporates weight decay for regularization. The learning rate follows a cosine annealing schedule, mathematically defined as:

$$\eta_t = \eta_{\text{min}} + \frac{1}{2}(\eta_{\text{max}} - \eta_{\text{min}}) \left(1 + \cos\left(\frac{T_{\text{cur}}}{T_{\text{max}}}\pi\right)\right) \tag{6.4}$$

where η_t is the learning rate at epoch t, η_{min} and η_{max} are the minimum and maximum learning rates respectively, and T_{cur} and T_{max} are the current epoch number and maximum epoch number. This allows the learning rate to gradually decrease over epochs, ensuring an effective learning process.

6.2.2.6 Data Module and Training

Training and validation data are loaded using a custom data module, MotionDataModule, which receives the preprocessed data and respective batch sizes for training and validation. The data module employs the PyTorch's DataLoader for batch loading and shuffling of training data.

For model training, the training process leverages the PyTorch Lightning's Trainer object. The training methodology uses the AdamW optimizer with an initial learning rate (lr) of 0.001 and a weight decay of 0.01 for regularization. The learning rate is

scheduled to follow a cosine annealing schedule, providing a smooth, gradient-based optimization.

The training proceeds for a maximum number of epochs, defined by the $max - epochs$ hyperparameter in the trainer. During each epoch, the model parameters are updated to minimize the Normalized Temperature-Scaled Cross Entropy (NT-Xent) loss calculated from the output of the model.

Furthermore, the top-1 and top-3 accuracy metrics are computed and logged at each epoch for both training and validation data. These metrics measure the model's performance and its ability to correctly identify the speaker style from the given motion data. The learning rate is also monitored at every epoch to track its decrement along with the training process.

The trainer is configured to utilize GPU acceleration if available, enhancing the computational efficiency of the model training.

6.2.2.7 Training Monitoring and Model Saving

The training process is designed to monitor the validation loss and the top-3 accuracy during training. The training state, model parameters, optimizer state, and learning rate scheduler state are periodically saved as checkpoints when the validation loss is at a minimum. These checkpoints are vital as they provide the means to resume training from the saved state, enabling long-duration, fault-tolerant training. They are also used for model evaluation and inference post-training.

The model-saving strategy is designed to retain the top three models as per their performance on the validation data. This is governed by the 'save-top-k' parameter of the ModelCheckpoint object, set to 3 in this case.

6.2.2.8 Hyperparameters

The model is defined and trained with the following notable hyperparameters:

- **Spatial Dimension of Model** : 150. It defines the size of the input layer for the spatial Transformer Encoder.
- **Number of Heads in Spatial Transformer Encoder**: 3. It defines the number of parallel attention layers, or ”heads”, used in the spatial Transformer Encoder.
- **Dimension of Model** : 768. It is the size of the input layer for the temporal Transformer Encoder.
- **Number of Heads in Temporal Transformer Encoder**: 8. It defines the number of parallel attention layers, or ”heads”, used in the temporal Transformer Encoder.
- **Number of Transformer Encoder Layers** : 4. It specifies the number of sub-encoder-layers in the Transformer Encoder.
- **Dropout rate** : 0.1. It determines the probability at which elements of the input tensor are zeroed.

The described Speaker Style Encoder Model architecture and training methodology presents a robust mechanism to encode and distinguish between individual speaker styles in motion data. The model encapsulates spatial and temporal information by effectively leveraging Transformer Encoders, ensuring a comprehensive representation of speaker style.

6.2.3 ConGRets Model Architecture and Training

6.2.3.1 Model Architecture

The ConGRets model is an ingenious fusion of transformer-based models, the Gesture-CLR, and the Speaker Style Encoder. The architecture shown in Fig6.2 and training process is shown in Fig6.3 is designed as follows:

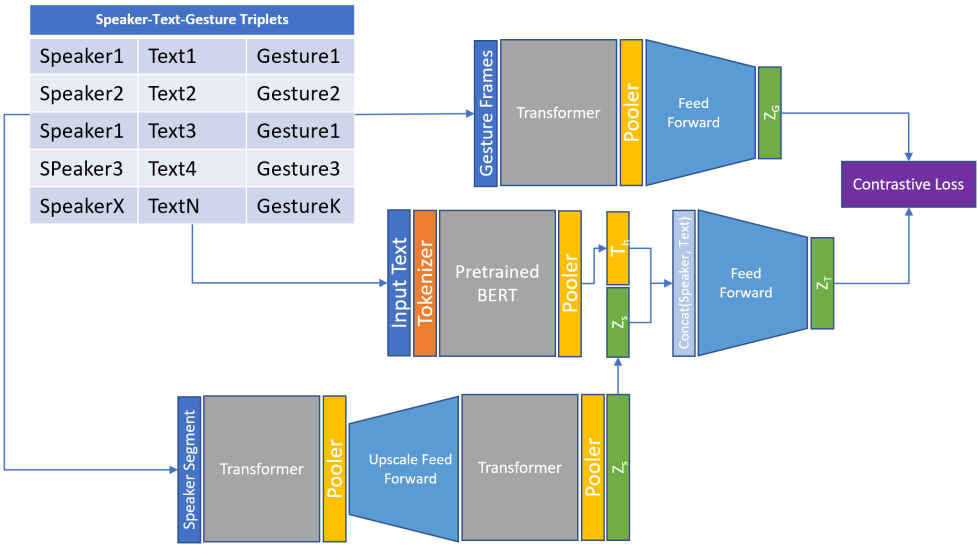


Figure 6.2: ConGRets with Speaker Style encoder

6.2.3.2 Transformer Encoder

The central component of the ConGRets model is a BERT model, specifically a compact variant of the BERT architecture pre-trained by google. This transformer-based model is utilized to capture both syntactic and semantic information within the text.

6.2.3.3 Mean Pooling

The output from the BERT model is passed through a mean-pooling layer. If we denote the token embeddings from the BERT model as X and the corresponding attention mask as M, the mean-pooled output Y is computed as:

$$Y = \frac{\sum_{i=1}^{n} X_i \cdot M_i}{\sum_{i=1}^{n} M_i}$$

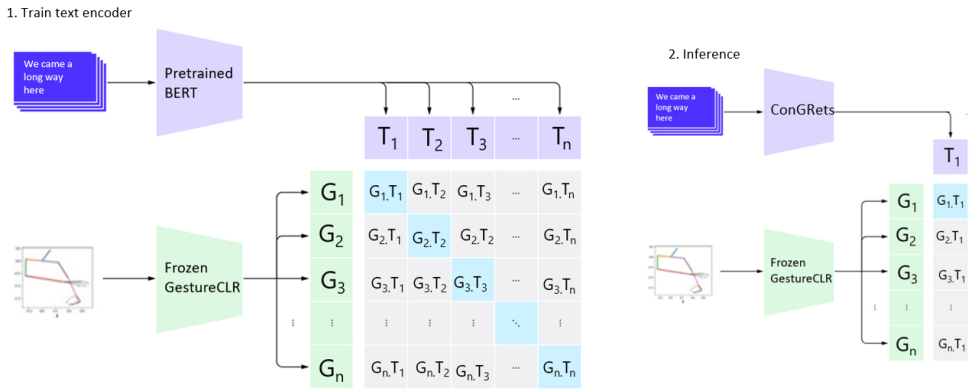


Figure 6.3: ConGRets Training and inference

6.2.3.4 GestureCLR and Speaker Style Encoder

The GestureCLR provides an initial clustering of gestures captured as latent representations. These latent gesture representations serve as target outputs during the training of the ConGRets model. On the other hand, the Speaker Style Encoder contributes by providing pre-computed speaker-style representations.

The speaker-style representations are concatenated with the mean-pooled output from the BERT model, aligning the speaker-style information with the textual information. This combined representation is then passed through a fully connected layer.

6.2.4 Training and Hyperparameters

We trained the model with a contrastive learning approach using the Normalized Temperature-Scaled Cross Entropy (NT-Xent) loss. The loss function is defined as:

$$L_{i,j} = -\log\left(\frac{\exp(\cos\_\text{sim}_{i,j})}{\sum_k \exp(\cos\_\text{sim}_{i,k})}\right) \tag{6.5}$$

where $\cos\_\text{sim}_{i,j}$ is the cosine similarity between the i-th u motion sequence from gestureCLR and the j-th v motion sequence from speaker style encoder, computed as:

$$\text{cos\_sim}_{i,j} = \frac{u_i \cdot v_j}{\|u_i\|_2 \|v_j\|_2} / T \tag{6.6}$$

In these equations:

- u_i and v_j are the latent representations of the i-th motion sequence from gesture-CLR and the j-th motion sequence from speaker style encoder, respectively. - $\cdot$ denotes the dot product. - $\| \cdot \|_2$ denotes the L2 norm. - T is the temperature hyperparameter, controlling the concentration of the distribution. - The index k runs over all possible negative samples.

The final loss is the mean over all pairs (i, j):

$$L = \frac{1}{N^2} \sum_{i,j} L_{i,j} \tag{6.7}$$

where N is the total number of samples in the batch.

The training process aims to maximize the similarity between the output of the ConGRets model and the target latent gesture representations from the GestureCLR.

AdamW optimizer has a learning rate of 5e-4 and a weight decay of 1e-4. A Cosine Annealing learning rate scheduler is employed, gradually reducing the learning rate as training progresses.

The model is trained for 1000 epochs, with batch sizes of 128 for training and 32 for validation. The training process leverages the PyTorch Lightning framework for model checkpointing and learning rate monitoring.

6.3 Dataset and Experimental Setup

In this section, we present an overview of the dataset and experimental setup used to assess the performance of our ConGRets method. We delineate the rulemaps and speaker data used in training and evaluation, and detail the process we followed to train our models and assess their effectiveness.

6.3.1 Rulemaps and Speaker Data

In training and evaluating our ConGRets method, we employed a diverse dataset composed of both rulemaps and speaker data. We utilized the BEAT dataset[33], a motion-captured compilation of 30 speakers. This dataset was divided into three subsets: 20 speakers for training, 5 for validation, and the remaining 5 for testing.

Rulemaps are made up of pairs of text segments and corresponding gesture units. They serve as a tool to instill the relationship between speech and gestures into the model. For each speaker, gesture units and rulemaps were generated using the method proposed by Ali et al.[15]. The rulemaps, 30 in total and identical in size, contained the same text data; the only difference between them being the gesture units.

The speaker data comprises animations of various speakers captured through motion capture. The motion-captured data from the BEAT dataset[33] was processed independently for the speaker style encoder, as explained in the Speaker Style Encoder section. This process provides a thorough representation of the diverse speaking styles and mannerisms.

6.3.2 Training and Evaluation

Our training process is comprised of two primary steps: utilizing a pretrained small BERT model and performing fine-tuning. We employed a publicly available pretrained small BERT model, provided by Google on their GitHub repository. This model, already trained on a vast corpus of text data to learn contextualized word embeddings, enables us to forego the pretraining phase and focus on fine-tuning.

We subsequently fine-tuned the pretrained small BERT model on our rulemaps dataset to prepare it for gesture generation. During this process, the model learns to map the contextualized word embeddings to the gesture latent space, which facilitates efficient gesture generation.

We trained the speaker encoding model on speaker data using contrastive learning

for the zero-shot speaker adaptation component. This method helps the model learn a latent space of speaker-specific animations, enabling the generated gestures to adapt to the target speaker's style.

In assessing the effectiveness of our method, we focused on two crucial parameters: execution time and speaker style adaptability. Execution time was evaluated by comparing the generated gestures with baseline and other methods, while speaker style adaptability was measured by how well the system adopted the style of new speakers.

During the evaluation, we benchmarked the scaling performance of ConGRets against a previous rule-based system and three other systems[75, 34, 35]. These systems were chosen because their authors provided the training code or models, allowing for fair evaluations. To evaluate the effect of the speaker style encoder, we trained ConGRets both with as in Fig6.2 and without it as in Fig6.4. Without the speaker style encoder, models for each speaker were trained separately; with the encoder, only a single model needed to be trained. Additionally, we measured the minimum data level with which speaker encoder can effectively differentiate the speaker's gesture and produce embeddings. These evaluations helped us discern our approach's relative strengths and weaknesses in speed, memory efficiency, and speaker alignment.

The baseline rule-based system used for comparison was constructed using the methods and text-pose data from Ali et al. [12].

In the subsequent section, we introduce the results of our experiments and explore their implications for the field of co-speech gesture generation.

6.4 Results

In this section, we present the results of our experiments comparing ConGRets to the rule-based system and other state-of-the-art methods. Our evaluation focuses on performance comparison, speed and memory efficiency, and zero-shot adaptation. We measured execution speed on both CPU and GPU for ConGRets, and implemented a

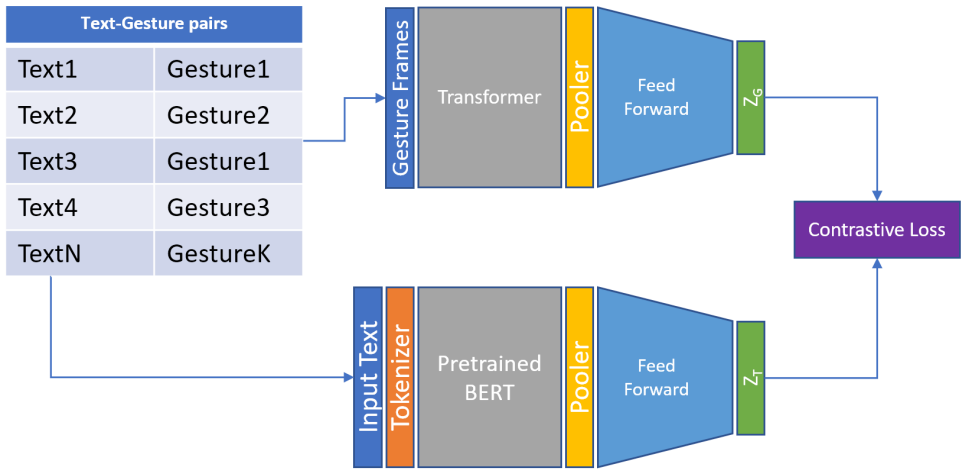


Figure 6.4: ConGRets without Speaker Style encoder

parallel pipeline for the rule-based system to speed up the matching process.

6.4.1 Performance Comparison

Our quantitative results demonstrate that ConGRets significantly outperforms the rule-based system in terms of speed. When processing 1000 text segments, ConGRets was able to complete the task in under 500 milliseconds on CPU. Moreover, it achieved even faster processing times on GPU, completing the task in under 200 milliseconds as shown in Fig6.5(d). In contrast, the rule-based system took up to 16 seconds to process only 50 queries as shown in Fig6.5(c). The slower performance of the rule-based system can be attributed to the use of a larger BERT model for improved matching accuracy and the linear search in the rule-based approach. Fig6.5(a,b) shows the execution time compared with the other three methods. Our method significantly improves over the previous methods in terms of scaling performance

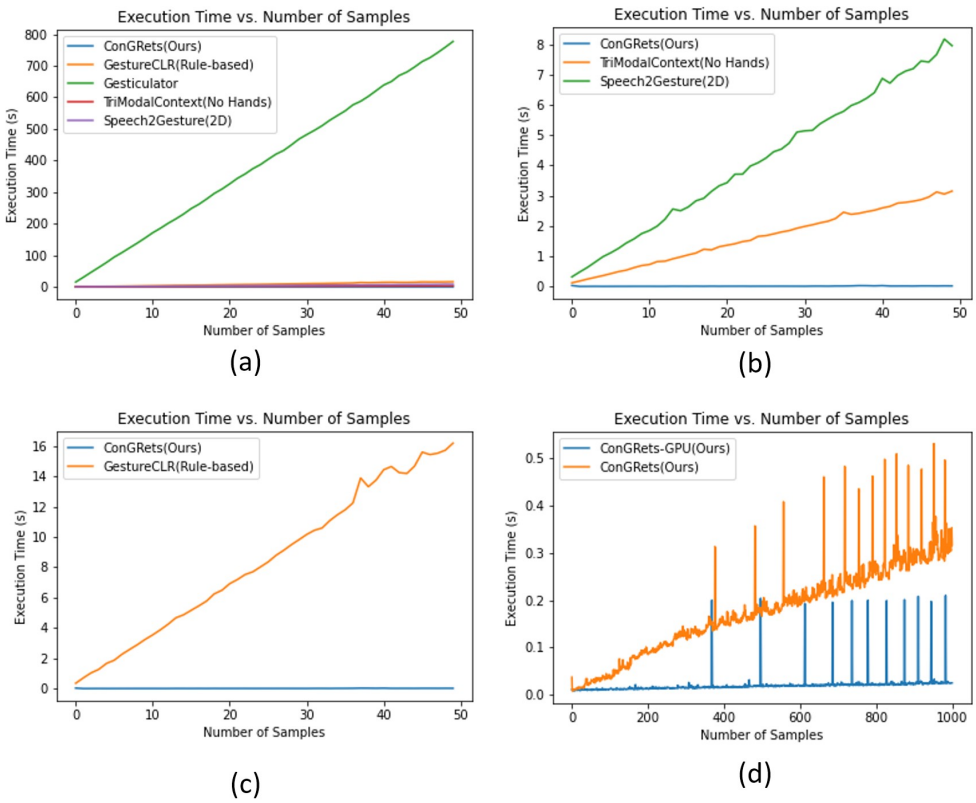


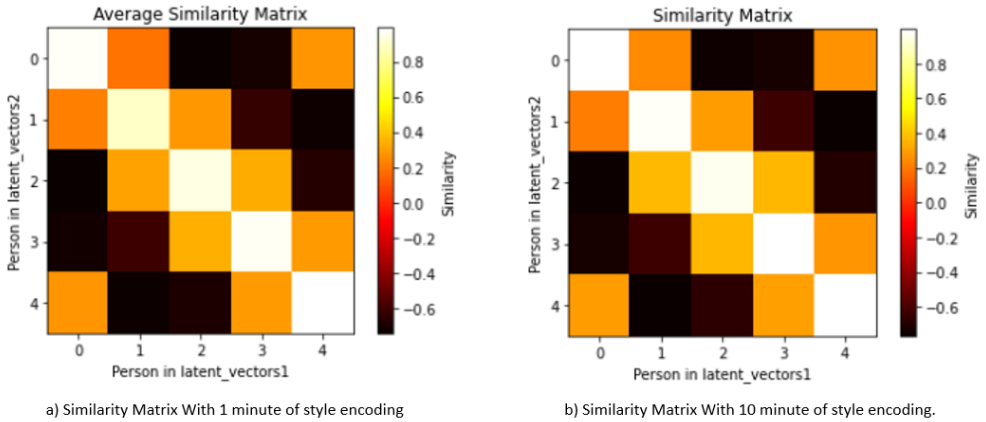
Figure 6.5: ConGRets execution time as compared to baseline

6.4.2 Speed and Memory Efficiency

ConGRets demonstrates remarkable speed and memory efficiency compared to the rule-based system. As a small language model, ConGRets can process text segments much faster, significantly improving the execution speed. Moreover, the memory usage of ConGRets is substantially lower (100 MB) compared to the rule-based system (2 GB), which further highlights the benefits of our proposed method.

6.4.3 Zero-Shot Adaptation

We evaluated the zero-shot speaker adaptation capabilities of ConGRets by testing it on a diverse dataset of previously unseen speakers. We aimed to determine whether a single model could effectively serve all speakers, conducting two evaluations: a minimum data requirement analysis and an ablation study to ascertain the effectiveness of the speaker style encoder in style adoption. Both were measured using cosine similarity, with the resulting cosine similarity matrices presented in FigureYa and FigureYb.



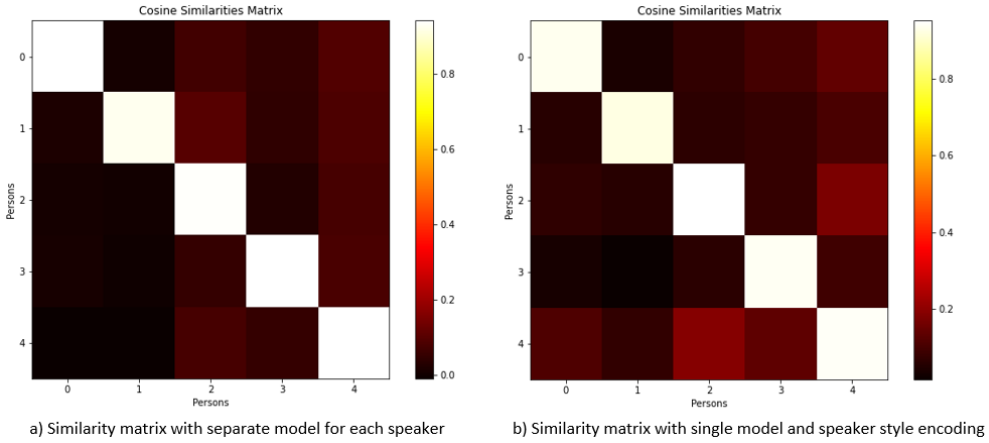
a) Similarity Matrix With 1 minute of style encoding

b) Similarity Matrix With 10 minute of style encoding.

Figure 6.6: Speaker Style encoder matrix

Fig6.6(a) presents the cosine similarity matrix derived from using only 1 minute of data, while Fig6.6(b) illustrates the matrix with 10 minutes of data. While the results suggest that our model can effectively differentiate between speakers even with 1 minute of data, using as much data as possible to generate speaker-style embeddings is optimal.

As outlined in the Training and Evaluation section, the zero-shot speaker style facilitates a single model's use for any speaker. Fig6.7(a) indicates that models trained on specific speakers tend to perform optimally for that speaker. However, as Fig6.7(b) depicts, a single model can learn to perform equivalently when provided with the



a) Similarity matrix with separate model for each speaker

b) Similarity matrix with single model and speaker style encoding

Figure 6.7: Speaker Style adaption matrix

speaker's style, thereby obviating the need for speaker-specific models.

Our findings underscore that the proposed method can adeptly adapt to the styles and mannerisms of new speakers, generating animations that accurately align with the target speaker. This demonstrates the flexibility of ConGRets in accommodating a wide array of speakers, hinting at its potential to enhance user experience across various applications.

In summary, our results demonstrate that ConGRets offers significant advantages over the rule-based system in terms of execution speed and memory efficiency while showcasing effective zero-shot adaptation capabilities. These findings suggest that ConGRets is a promising approach for co-speech gesture generation, with the potential to improve the user experience in various applications.

6.5 Discussion

In this section, we discuss the implications of our findings, the limitations of the current study, and potential avenues for future research.

6.5.1 Implications

The ConGRets method has significant implications for the field of human-computer interaction and the development of virtual characters and agents. Our results demonstrate that the new approach is considerably faster and more memory-efficient than the previous rule-based system, making it more feasible for real-time applications and deployment on devices with limited computational resources.

Furthermore, the zero-shot speaker adaptation capability of ConGRets offers a promising direction for personalizing virtual agents and characters to individual users, leading to a more immersive and engaging user experience. By leveraging contrastive learning, the method can generalize to new speakers without requiring extensive retraining or manual annotation, making it more scalable and adaptable to various scenarios.

6.5.2 Limitations and Future Work

Despite the promising results and contributions of this study, there are some limitations that warrant further investigation. First, the current implementation of ConGRets relies on a small BERT model, which may not capture all the nuances of human language and co-speech gestures. Future research could explore the use of alternative language models or hybrid architectures to improve gesture generation quality.

Second, while the zero-shot speaker adaptation is an important step towards personalized virtual agents, the current method only considers the generation of gestures aligned with the speaker. Additional research is needed to incorporate other aspects of individuality, such as facial expressions, prosody, and body language, for a more complete representation of a speaker's communication style.

Lastly, the evaluation of ConGRets primarily focused on speed, memory efficiency, and alignment with speakers. Future work could include a more comprehensive assessment of the system's performance, such as user studies to evaluate the perceived

naturalness and effectiveness of the generated gestures in different contexts and applications. Additionally, the exploration of other evaluation metrics, such as the semantic coherence between the generated gestures and the spoken content, would provide valuable insights into the method's effectiveness.

To summarize the discussion, our study presents the ConGRets method as an effective approach to improve co-speech gesture generation, offering significant advantages in terms of speed, memory efficiency, and zero-shot speaker adaptation. The findings and limitations of this research provide a strong foundation for future studies aimed at further enhancing the naturalness and personalization of virtual characters and agents in human-computer interaction.

6.6 Conclusion

In this paper, we presented ConGRets, a novel method for co-speech gesture generation that leverages contrastive learning to improve the speed, memory efficiency, and speaker alignment of virtual characters and agents. By shifting from a rule-based system that matches text with text to a contrastive learning approach that matches gestures, we demonstrated substantial improvements in both computational efficiency and speaker alignment compared to the previous method.

Our approach incorporates a small pretrained BERT model, a pooling head, and linear layers to generate latents similar to those of gestureCLR, a model previously used for matching poses to gesture units. We also introduced a zero-shot learning technique for adapting to new speakers' animation encodings, which allows for a more personalized and immersive user experience.

The results of our experiments show that ConGRets significantly outperforms the previous rule-based system in terms of speed, memory usage, and speaker alignment. However, we also acknowledge the limitations of our current method, including the reliance on a small BERT model and the focus on gesture generation without considering

other aspects of individual communication styles.

Future research can build upon our findings to explore alternative language models, hybrid architectures, and the incorporation of facial expressions, prosody, and body language for a more comprehensive representation of a speaker's communication style. Additionally, user studies and the evaluation of semantic coherence between generated gestures and spoken content will provide valuable insights into the effectiveness and naturalness of ConGRets in various contexts and applications.

In short, ConGRets offers a promising direction for the development of more efficient, personalized, and natural virtual characters and agents in the field of human-computer interaction.

Chapter 7

Discussion

This chapter delves into a comprehensive analysis of the research findings, breakthroughs, and implications in the evolution of co-speech text-to-gesture generation for digital humans. Furthermore, we engage in a thoughtful exploration of the challenges faced during the process, their implications for virtual human interaction, potential applications, and prospects for future development. Our reflection navigates between the pragmatic realities of our work and a broader philosophical perspective on the transformative potential of this research.

7.1 Overview of the main findings

This research journey was initiated from the recognition of the need for a more efficient and dynamic system to generate authentic body gestures for digital humans in mixed-reality environments. We sought to create a harmonious integration of verbal and non-verbal communication, mirroring the complexity and depth of human interaction. Our initial development of an intelligent virtual agent system marked the first step in this journey, but it soon became evident that we needed a more automated approach to

gesture generation.

Motivated by this requirement, we proposed an automated rule-based co-speech gesture mapping system that transcended the limitations of traditional human-created mappings. It was a major leap, but the road to a truly immersive and human-like digital interaction was still far.

Our most significant breakthrough came with the development of GestureCLR, a deep contrastive learning model. This model fundamentally transformed the way we matched 2D poses from videos to 3D gesture units, offering a substantial improvement over the earlier mean cosine similarity method. GestureCLR paved the way for the creation of a cross-lingual method that maintained a consistent and high-quality user experience across different languages, an achievement that truly underscored the universal potential of our research.

The apex of our research was the development of ConGRets, a cutting-edge framework that balances low latency performance and resource usage. In a philosophical sense, ConGRets represents the culmination of our quest to bridge the gap between the digital and human world. Its zero-shot speaker style adaptation, a feature that allows the system to generate gestures that align with the text input and the speaker's unique motion style, brought us closer to creating a digital human that can mirror the individuality and uniqueness of human beings.

7.2 Evaluations and User Studies

Our research on co-speech gesture generation was guided by the principle that the style of gestures is a deeply personal and non-deterministic aspect of human communication which is a consistent position with others[53, 36, 75]. We thus concentrated on aspects that could be reliably measured and evaluated while acknowledging that objective measures may not accurately capture certain subjective elements of style.

Through the literature review and analysis of previous work, we noted several chal-

lenges associated with objective evaluation methods. For instance, the evaluation of gesture generation often relied heavily on measures such as acceleration, jerk, Probability of Correct Keypoints (PCK), and mean squared error (MSE). The main purpose of these methods is to measure how close the generated gesture is to the ground truth gesture. The problem is that even humans cannot repeat a gesture twice in the same way. which means a gesture maybe perfect but might get a bad score from these methods. While these measures can capture the technical aspects of the generated gestures, they fall short of reflecting the human perception of naturalness and appropriateness. Furthermore, a meta-study [76] revealed a poor correlation between objective and subjective evaluation methods, especially when assessing how closely the generated output aligns with the ground truth.

We also noted inconsistent Fréchet Inception Distance (FID) interpretations in previous works. FID is used primarly to a becnchmark the image generation models. The lower FID, the better the model is. While some researchers, like Yoon et al.[35] and Ahuja et al.[36], used high FID values to argue for the diversity of their gesture generations, others, such as Bhattacharya et al.[50] used lower FID values to claim their system is better because it is closer to the ground truth. This lack of consensus further underscores the limitations of relying solely on objective measures to evaluate the quality of gesture generation. So we used objective and empirical evidence only where we could measure it tangibly.

It is interesting to note that subjective evaluation of a method is often conducted without comparing it to other methods, as such studies can be quite costly. Instead, objective criteria are typically used to compare methods, with subjective evaluation used to further strengthen the argument that a particular method is superior to others. The GENEA challenge[76], for instance, utilized crowd-sourced evaluation to study various methods, which were evaluated independently based on two criteria: human-like and appropriate. In our own work, we also employed a similar approach, comparing

our method to others solely based on objective criteria. When conducting subjective evaluations, we used a study by Lee et al.[30] as a baseline in our first study, and our own previous method as a baseline in a subsequent study.

7.2.1 Objective Evaluations

As part of our work, we present six objective and empirical evaluations that assess different aspects of gesture generation. We prioritize selecting aspects that can be measured reliably. In an empirical appraisal of the distribution frequency of gestures culled from our rule-based repository, we implemented a threshold. This predefined limit was critical in ensuring an adequate variety in the gestures generated. The inherent value in this approach lay in our ability to circumvent the creation of excessively random sequences of gestures. This, in turn, contributed substantially to the overall naturalness and coherence inherent in the gesture sequences generated.

In our quest for superior results, we discovered that embeddings, which consider the nuances of semantics, significantly outperformed the more traditional approach of string-based text matching. This critical finding highlighted the pivotal role played by semantic congruity in bridging the gap between speech and gestures. This is a key consideration when the objective is to generate co-speech gestures that are convincingly human-like and credible.

We then turned our attention to GestureCLR, a model we trained to achieve high accuracy in matching 2D and 3D motions. While this model exhibits exceptional prowess in motion matching, it, unfortunately, falls short when tasked with evaluating the gestures that have been generated. This limitation arises from the intricate tie between the gestures generated and the words spoken.

In evaluating the processing time of ConGRets, our analysis uncovered its outstanding efficiency, particularly when dealing with extensive input sequences. This performance contrasts sharply with rule-based methods, which can process multiple

text segments for several seconds. ConGRets, on the other hand, manages to process a thousand segments in less than half a second. This feature is particularly critical in interactive situations, where protracted delays in gesture generation may adversely affect the user experience. A significant challenge encountered in our research was the frequent unavailability of original implementations from other researchers. This issue often made it difficult to conduct fair comparisons. Despite this constraint, our method demonstrated superior efficiency in execution time comparisons.

Further, our exploration of the functionality of the Speaker Style Encoder demonstrated its ability to effectively differentiate between speakers, even when provided with minimal data. While this discovery does not directly pertain to style-matching, it nonetheless provides clear evidence of the versatility and effectiveness of our proposed method.

Finally, we conducted an ablation study to gauge the utility of the Speaker Style Encoder in a zero-shot adaptation context. The study's results highlighted the capability of a single model to perform on par with separate models designed for each speaker.

7.2.2 Subjective Evaluations

In our user studies, our method consistently outperformed the baseline, providing strong subjective evidence for the effectiveness of our approach in generating human-like and appropriate gestures. This success further strengthens our position that subjective methods of evaluation are crucial in the field of co-speech gesture generation.

Our research corroborated the observed disconnect between objective measures of accuracy and human perception of gesture naturalness and appropriateness. Even gestures perfectly generated according to objective measures may significantly deviate from the ground truth, further emphasizing the importance of subjective evaluations for assessing the perceived quality of gesture generation.

Our research underscores the need for a nuanced and human-centric approach to

evaluating co-speech gesture generation. The evidence from our studies strongly suggests that subjective evaluations, despite their inherent challenges, provide a more accurate measure of the quality of gesture generation. Future research should endeavor to further refine these methods, as we continue to strive for more believable, natural, and appropriate co-speech gestures.

7.3 Implications for virtual human interaction and applications

The implications of our research are profound and far-reaching, dramatically transforming virtual human interaction and its myriad applications. The development of the automated gesture generation system and the deep contrastive learning model, GestureCLR, is not merely a technical advancement; it represents a philosophical shift in our interaction with digital humans. We have brought digital humans closer to mimicking the intricacy of human gestures, a step towards breaking the barrier between the real and digital realms.

The cross-lingual capabilities of our system fundamentally democratize digital human interaction, transcending language barriers and facilitating global communication. This development holds immense potential for sectors such as entertainment, education, and customer service, where the ability to communicate without language constraints can create more inclusive and diverse interactions. Some examples of the application of our method are shown in Fig7.1

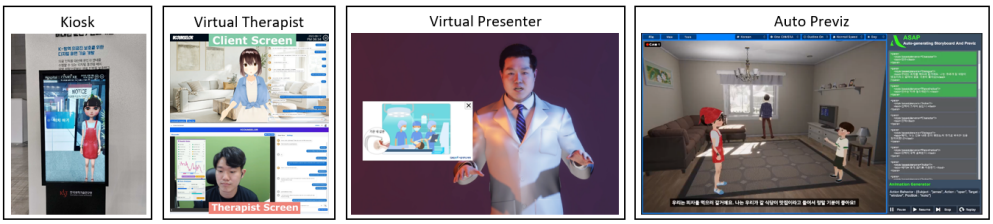


Figure 7.1: Applicaitons of our system

The introduction of ConGRets, optimized for low latency performance and resource usage, heralds a new era of dynamic, real-time interaction with digital humans. This enhancement can revolutionize sectors like virtual meetings and live presentations, making them more engaging and realistic. Furthermore, the potential applications in medical and therapy sectors underscore the system's capacity to create more human-like, intuitive, and empathetic user experiences.

7.4 Limitations and future work

While our research represents a significant stride in the right direction, it also illuminates areas that need further investigation. A key limitation is our system's inability to generate reliable iconic and deictic gestures, primarily due to the dependency on monologue video data. Future exploration should focus on capturing conversational video data, emphasizing interactive dialogues that mimic real-world communication. This would likely enrich the gesture generation system, enabling it to produce more contextually accurate and diverse gestures.

Another significant limitation lies in the system's incapacity to generate emotional gestures. This gap underscores the complexity of human communication, where emotions play a pivotal role in shaping interactions. From a philosophical standpoint, this limitation poses an intriguing question about the depth of human expressiveness that can be encoded into a digital entity. Future research should endeavor to develop a system that comprehends and reproduces gestures corresponding to different emotional states, enhancing the emotional intelligence of digital humans.

Beyond these specific limitations, the broader vision for future work could encompass the inclusion of other non-verbal cues, such as facial expressions and eye movements, into the digital human model. These elements are integral to human communication, often conveying subtleties that words fail to express. Incorporating these elements could significantly enrich the communicative abilities of digital humans, making

them more relatable and effective.

In closing, while our research marks a significant advancement in co-speech text-to-gesture generation for digital humans, it also poses profound philosophical questions about the nature of communication and interaction. We are on the cusp of a transformative era where digital entities might communicate as effectively as their human counterparts, blurring the lines between reality and the digital realm. This journey, while challenging, offers exciting possibilities that could redefine the future of human-computer interaction.

Chapter 8

Conclusion

This thesis has systematically investigated and optimized the process of co-speech text-to-gesture generation for digital entities. It commenced with the development of an intelligent virtual agent system, progressed to the creation of an automated, rule-based co-speech gesture mapping system, and culminated in the realization of ConGRets, a state-of-the-art framework with zero-shot speaker style adaptation capabilities.

Throughout this research, significant strides have been made in enhancing the non-verbal communication capabilities of digital humans. The development of Gesture-CLR, a deep contrastive learning model, facilitated a substantial improvement in matching 2D poses from videos to 3D gesture units. Furthermore, this model enabled the creation of a cross-lingual method for gesture generation, thereby amplifying the universal applicability of our research.

The apex of this research is represented by ConGRets, a system that successfully balances real-time performance and resource usage. By integrating the feature of zero-shot speaker style adaptation, ConGRets has managed to refine the alignment between the text input and the motion style of the individual speaker. The implications of this

development are profound, encompassing a wide array of applications and potentially revolutionizing sectors such as entertainment, education, customer service, and healthcare.

Despite these advancements, the research also highlighted limitations that warrant further exploration. The current system's inability to generate reliable iconic, deictic, and emotional gestures poses a challenge. Future research should, therefore, focus on capturing conversational video data, developing methods to comprehend and reproduce gestures associated with different emotional states, and incorporating additional non-verbal cues such as facial expressions and eye movements.

In sum, this research represents a significant contribution to the field of co-speech text-to-gesture generation for digital humans. It marks notable advancements in the field and unveils potential future directions. Doing so prompts philosophical reflections on the nature of communication and interaction and the increasingly blurred lines between reality and the digital realm. As such, it serves as both a landmark in the present landscape and a beacon guiding future development in the field of human-computer interaction.

Bibliography

[1] R. Xiao, J. Schwarz, N. Throm, A. Wilson, and H. Benko, "Mrtouch: Adding touch input to head-mounted mixed reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 4, p. 2018, 2018.

[2] D. Richards, "Agent-based museum and tour guides," 2012.

[3] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: The behavior expression animation toolkit," *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001*, pp. 477–486, 2001.

[4] M. Neff, M. Kipp, I. Albrecht, and H. P. Seidel, "Gesture modeling and animation based on a probabilistic re-creation of speaker style," *ACM Transactions on Graphics*, 2008.

[5] L. Cai, B. Liu, J. Yu, and J. Zhang, "Human behaviors modeling in multi-agent virtual environment," *Multimedia Tools and Applications*, vol. 76, no. 4, pp. 5851–5871, 2015.

[6] J. Arroyo-Palacios and R. Marks, "Poster believable virtual characters for mixed reality. 2017 ieee international symposium on mixed and augmented reality (ismar-adjunct)," *Nantes*, pp. 121–123, 2017.

[7] A. Bogdanovych, T. Trescak, and S. Simoff, “Formalising believability and building believable virtual agents,” *Lecture Notes in Computer Science*, pp. 142–156, 2015.

[8] M. Studdert-Kennedy, “Hand and Mind: What Gestures Reveal About Thought,” 1994.

[9] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, “BEAT: The Behavior Expression Animation Toolkit,” in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001*, 2001.

[10] J. Lee and S. Marsella, “Nonverbal behavior generator for embodied conversational agents,” in *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), 2006.

[11] M. Neff, M. Kipp, I. Albrecht, and H. P. Seidel, “Gesture modeling and animation based on a probabilistic re-creation of speaker style,” *ACM Transactions on Graphics*, 2008.

[12] G. Ali, M. Lee, and J. Hwang, “Automatic text-to-gesture rule generation for embodied conversational agents,” *Computer Animation and Virtual Worlds*, vol. 31, no. 4-5, p. e1944, 2020.

[13] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff, “A comprehensive review of data-driven co-speech gesture generation,” vol. 42, 1 2023.

[14] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2014.

[15] G. Ali and J.-I. Hwang, “Improving co-speech gesture rule-map generation via wild pose matching with gesture units,” *SIGGRAPH Asia 2022 Posters*, vol. PartF168147-3, pp. 1575–1585, 12 2022.

[16] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, “Analyzing input and output representations for speech-driven gesture generation,” 2019.

[17] H. Kim, G. Ali, and J.-I. Hwang, “Asap: Auto-generating storyboard and previz with virtual humans,” in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct* (*ISMAR-Adjunct*), pp. 316–320, 2021.

[18] W. Liu, O. Fernando, A. Cheok, J. Wijesena, and R. Tan, “Science museum mixed reality digital media exhibitions for children,” *Second Workshop on Digital Media and its Application in Museum Heritages* (*DMAMH*, vol. 2007, pp. 389–394, 2007.

[19] S. McArthur, E. Davidson, V. Catterson, and et al., “Multi-agent systems for power engineering applications – part i: concepts, approaches, and technical challenges,” *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 1743–1752, 2007.

[20] S. McArthur, E. Davidson, V. Catterson, and et al., “Multi-agent systems for power engineering applications – part ii: technologies, standards, and tools for building multi-agent systems,” *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 1753–1759, 2007.

[21] V. Catterson, E. Davidson, and S. McArthur, “Practical applications of multi-agent systems in electric power systems,” *European Transactions on Electrical Power*, vol. 22, no. 2, pp. 235–252, 2012.

[22] R. Castano, G. Dotto, R. Suma, A. Martina, and A. Bottino, “Virtual-me: A library for smart autonomous agents in multiple virtual environments,” *Communications in Computer and Information Science*, pp. 34–45, 2015.

[23] M. Alencar and J. Netto, *Improving cooperation in Virtual Learning Environments using multi-agent systems and AIML*. Frontiers in Education Conference (FIE). Rapid City, 2011.

[24] J. Xie and C. Liu, “Multi-agent systems and their applications,” *Journal of International Council on Electrical Engineering*, vol. 7, no. 1, pp. 188–197, 2017.

[25] M. Anabuki, H. Kakuta, H. Yamamoto, and H. Tamura, “Welbo: an embodied conversational agent living in mixed reality space,” *In CHI EA ’00 CHI ’00 Extended Abstracts on Human Factors in Computing Systems* (*CHI ’00*). *The Hague, The Netherlands*, pp. 10–11, 2000.

[26] S. Kopp and et al., “Max - a multimodal assistant in virtual reality construction,” *KI - K u nstliche Intelligenz*, vol. 4, pp. 11–17.

[27] O. M. Machidon, M. Duguleana, and M. Carrozzino, “Virtual humans in cultural heritage ict applications: A review,” *Journal of Cultural Heritage*, vol. 33, pp. 249–260, 2018.

[28] S. Vosinakis, N. Avradinis, and P. Koutsabasis, “Dissemination of intangible cultural heritage using a multi-agent virtual world,” *Lecture Notes in Computer Science*, pp. 197–207, 2017.

[29] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, “Animated conversation: Rule-based generation of facial expression, gesture spoken intonation for multiple conversational

agents," *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1994*, pp. 413–420, 7 1994.

[30] J. Lee and S. Marsella, "Nonverbal behavior generator for embodied conversational agents," *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), vol. 4133 LNAI, pp. 243–255, 2006.

[31] K. Takeuchi, D. Hasegawa, S. Shirakawa, N. Kaneko, H. Sakuta, and K. Sumi, "Speech-to-gesture generation," pp. 365–369, ACM Press, 2017.

[32] Y. Ferstl, M. Neff, and R. McDonnell, "Expressgesture: Expressive gesture generation from speech through database matching," *Computer Animation and Virtual Worlds*, vol. 32, 6 2021.

[33] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng, "Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), vol. 13667 LNCS, pp. 612–630, 2022.

[34] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning Individual Styles of Conversational Gesture," jun 2019.

[35] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics*, vol. 39, no. 6, 2020.

[36] C. Ahuja, D. W. Lee, Y. I. Nakano, and L.-P. Morency, "Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach," *arxiv.org*, pp. 248–265, 7 2020.

[37] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents," in *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pp. 413–420, 1994.

[38] J. Cassell, "Embodied conversational interface agents," *Communications of the ACM*, vol. 43, no. 4, pp. 70–78, 2000.

[39] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 477–486, 2001.

[40] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro, "Virtual character performance from speech," in *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 25–35, 2013.

[41] M. Lhommet, Y. Xu, and S. Marsella, "Cerebella: automatic generation of nonverbal behavior for virtual humans," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[42] B. Ravenet, C. Pelachaud, C. Clavel, and S. Marsella, "Automating the production of communicative gestures in embodied characters," *Frontiers in psychology*, vol. 9, p. 1144, 2018.

[43] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, *et al.*, "Towards a common framework for multimodal generation: The behavior markup language," in *International conference on intelligent virtual agents*, pp. 205–217, Springer, Berlin, Heidelberg, 2004.

[44] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel, “Gesture modeling and animation based on a probabilistic re-creation of speaker style,” *ACM Transactions on Graphics* (*TOG*), vol. 27, no. 1, pp. 1–24, 2008.

[45] M. Kipp, *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers, 2005.

[46] M. Kipp, M. Neff, K. H. Kipp, and I. Albrecht, “Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis,” in *IVA*, vol. 7, pp. 15–28, 2007.

[47] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, “Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots,” in *Proc. of The International Conference in Robotics and Automation* (*ICRA*), 2019.

[48] E. Asakawa, N. Kaneko, D. Hasegawa, and S. Shirakawa, “Evaluation of text-to-gesture generation model using convolutional neural network,” *Neural Networks*, vol. 151, pp. 365–375, 2022.

[49] C.-C. Chiu, L.-P. Morency, and S. Marsella, “Predicting co-verbal gestures: A deep and temporal modeling approach,” in *Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26-28, 2015, Proceedings 15*, pp. 152–166, Springer, 2015.

[50] U. Bhattacharya, N. Rewkowski, A. Banerjee, P. Guhan, A. Bera, and D. Manocha, “Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents,” in *2021 IEEE virtual reality and 3D user interfaces* (*VR*), pp. 1–10, IEEE, 2021.

[51] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 97–104, 2019.

[52] Y. Yang, J. Yang, and J. Hodgins, "Statistics-based motion synthesis for social conversations," in *Computer Graphics Forum*, vol. 39, pp. 201–212, Wiley Online Library, 2020.

[53] Y. Ferstl, M. Neff, and R. McDonnell, "Multi-objective adversarial gesture generation," *Proceedings - MIG 2019: ACM Conference on Motion, Interaction, and Games*, 2019.

[54] M. Fares, C. Pelachaud, and N. Obin, "Zero-shot style transfer for multimodal data-driven gesture synthesis," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition* (*FG*), pp. 1–4, 2023.

[55] F. Freigang and S. Kopp, "This is what's important – using speech and gesture to create focus in multimodal utterance," *Intelligent Virtual Agents*, pp. 96–109, 2016.

[56] J. Kolkmeier, J. Vroon, and D. Heylen, "Interacting with virtual agents in shared space: Single and joint effects of gaze and proxemics," *Intelligent Virtual Agents*, pp. 1–14.

[57] B. Yumak, Z. van den Brink and A. Egges, "Autonomous social gaze model for an interactive virtual character in real-life settings.," *Computer Animation and Virtual Worlds*, vol. 28, pp. 3–4, 2017.

[58] S. Uchino, N. Abe, K. Tanaka, T. Yagi, H. Taki, and S. He, "Vr interaction in real-time between avatar with voice and gesture recognition system," *Ad-*

vanced Information Networking and Applications Workshops (*AINAW'0*, vol. 7, pp. 959–964, 2007.

[59] C. Rebman, M. Aiken, and C. Cegielski, "Speech recognition in the human–computer interface," *Information Management*, vol. 40, no. 6, pp. 509–519, 2003.

[60] V. Avramova, F. Yang, C. Li, C. Peters, and G. Skantze, "A virtual poster presenter using mixed reality," *Intelligent Virtual Agents*, pp. 25–28, 2017.

[61] M. L. Knapp and J. A. Hall, *Nonverbal communication in human interaction*. 1972.

[62] M. Anabuki, H. Kakuta, H. Yamamoto, and H. Tamura, "Welbo: An embodied conversational agent living in mixed reality space," in *Conference on Human Factors in Computing Systems - Proceedings*, 2000.

[63] J. Arroyo-Palacios and R. Marks, "Believable Virtual Characters for Mixed Reality," in *Adjunct Proceedings of the 2017 IEEE International Symposium on Mixed and Augmented Reality, ISMAR-Adjunct 2017*, 2017.

[64] V. Avramova, F. Yang, C. Li, C. Peters, and G. Skantze, "A virtual poster presenter using mixed reality," in *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), 2017.

[65] R. Castano, G. Dotto, R. Suma, A. Martina, and A. Bottino, "Virtual-me: A library for smart autonomous agents in multiple virtual environments," in *Communications in Computer and Information Science*, 2014.

[66] S. Kopp, B. Jung, N. Leßmann, and I. Wachsmuth, "Max – A Multimodal Assistant in Virtual Reality Construction," *Künstliche Intelligenz*, 2003.

[67] O. M. Machidon, M. Duguleana, and M. Carrozzino, "Virtual humans in cultural heritage ICT applications: A review," 2018.

[68] D. Richards, "Agent-based museum and tour guides," 2012.

[69] S. Vosinakis, N. Avradinis, and P. Koutsabasis, "Dissemination of Intangible Cultural Heritage Using a Multi-agent Virtual World," in *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), 2018.

[70] A. Bogdanovych, T. Trescak, and S. Simoff, "Formalising believability and building believable virtual agents," in *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), 2015.

[71] F. De Coninck, Z. Yumak, G. Sandino, and R. Veltkamp, "Non-verbal behavior generation for virtual characters in group conversations," in *Proceedings - 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality, AIVR 2019*, 2019.

[72] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[73] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing Input and Output Representations for Speech-Driven Gesture Generation," in *IVA 2019 - Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019.

[74] G. Ali, H.-Q. Le, J. Kim, S.-W. Hwang, and J.-I. Hwang, "Design of seamless multi-modal interaction framework for intelligent virtual agents in wearable

mixed reality environment," in *CASA '19*, (New York, NY, USA), p. 47–52, Association for Computing Machinery, 2019.

[75] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, and G. E. Henter, "A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020," *International Conference on Intelligent User Interfaces, Proceedings IUI*, pp. 11–21, 4 2021.

[76] Y. Yoon, P. Wolfert, T. Kucherenko, C. Viegas, T. Nikolov, M. Tsakov, and G. E. Henter, "The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation," *ACM International Conference Proceeding Series*, pp. 736–747, 11 2022.

[77] S. Kita, *Pointing: Where language, culture, and cognition meet*. Psychology Press, 2003.

[78] D. Archer, "Unspoken diversity: Cultural differences in gestures," *Qualitative sociology*, vol. 20, no. 1, p. 79, 1997.

[79] S. Kita, "Cross-cultural variation of speech-accompanying gesture: A review," *Language and cognitive processes*, vol. 24, no. 2, pp. 145–167, 2009.

[80] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[81] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[82] G. Lee, Z. Deng, S. Ma, T. Shiratori, S. Srinivasa, and Y. Sheikh, "Talking with hands 16.2m: A large-scale dataset of synchronized body-finger motion and

audio for conversational motion analysis and synthesis," in *2019 IEEE/CVF International Conference on Computer Vision* (*ICCV*), pp. 763–772, 2019.

[83] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[84] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the International Conference on Learning Representations*, 2018.

[85] W. Falcon and T. P. L. team, "Pytorch lightning," 2019.

[86] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 11 2019.

[87] Y. Shiban, I. Schelhorn, V. Jobst, A. Hörnlein, F. Puppe, P. Pauli, and A. Mühlberger, "The appearance effect: Influences of virtual agent features on performance and motivation," *Computers in Human Behavior*, vol. 49, pp. 5–11, 2015.

[88] S. ter Stal, M. Tabak, H. op den Akker, T. Beinema, and H. Hermens, "Who do you prefer? the effect of age, gender and role on users' first impressions of embodied conversational agents in ehealth," *International Journal of Human–Computer Interaction*, vol. 36, no. 9, pp. 881–892, 2020.

[89] "The most popular talks of all time." `https://www.ted.com/playlists/171/the_most_popular_talks_of_all`. Accessed: 2022-11-20.

[90] Y. Liu, G. Mohammadi, Y. Song, and W. Johal, "Speech-based gesture generation for robots and embodied agents: A scoping review," in *Proceedings of*

the 9th International Conference on Human-Agent Interaction, HAI '21, (New York, NY, USA), p. 31–38, Association for Computing Machinery, 2021.

[91] P. Wolfert, N. Robinson, and T. Belpaeme, "A review of evaluation practices of gesture generation in embodied conversational agents," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 3, pp. 379–389, 2022.

[92] S. Olejnik and J. Algina, "Generalized eta and omega squared statistics: Measures of effect size for some common research designs," *Psychological Methods*, vol. 8, pp. 434–447, 2003.

[93] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *International Journal of Social Robotics*, vol. 4, pp. 201–217, 2012.

[94] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, pp. 313–323, 2013.

[95] C. Breazeal, K. Dautenhahn, and T. Kanda, "Social robotics," *Springer handbook of robotics*, pp. 1935–1972, 2016.

[96] H. J. Smith and M. Neff, "Understanding the impact of animated gesture performance on personality perceptions," *ACM Transactions on Graphics* (*TOG*), vol. 36, no. 4, pp. 1–12, 2017.

[97] C. L. Sidner, C. Lee, L.-P. Morency, and C. Forlines, "The effect of head-nod recognition in human-robot conversation," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 290–296, 2006.

[98] N. C. Krämer, N. Simons, and S. Kopp, "The effects of an embodied conversational agent' s nonverbal behavior on user' s evaluation and behavioral

mimicry," in *Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, 2007 Proceedings 7*, pp. 238–251, Springer, 2007.

[99] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan, "Embodiment in conversational interfaces: Rea," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 520–527, 1999.

[100] K. Bergmann, S. Kopp, and F. Eyssel, "Individualized gesturing outperforms average gesturing–evaluating gesture production in virtual humans," in *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, pp. 104–117, Springer, 2010.

[101] C.-C. Chiu and S. Marsella, "Gesture generation with low-dimensional embeddings," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 781–788, 2014.

[102] C. N. Gunawardena and F. J. Zittle, "Social presence as a predictor of satisfaction within a computer-mediated conferencing environment," *American journal of distance education*, vol. 11, no. 3, pp. 8–26, 1997.

[103] J. N. Bailenson, K. Swinth, C. Hoyt, S. Persky, A. Dimov, and J. Blascovich, "The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments," *Presence*, vol. 14, no. 4, pp. 379–393, 2005.

[104] A. Kendon, *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[105] D. McNeill, *Gesture and thought*. University of Chicago press, 2005.

[106] N. Zielinski and E. M. Wakefield, “Language proficiency impacts the benefits of co-speech gesture for narrative understanding through a visual attention mechanism,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43, 2021.

[107] M. Thiebaux, S. Marsella, A. N. Marshall, and M. Kallmann, “Smartbody: Behavior realization for embodied conversational agents,” in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, pp. 151–158, 2008.

[108] Y. Yoon, K. Park, M. Jang, J. Kim, and G. Lee, “Sgtoolkit: An interactive gesture authoring toolkit for embodied conversational agents,” in *The 34th Annual ACM Symposium on User Interface Software and Technology*, pp. 826–840, 2021.

[109] M. Rebol, C. Gütl, and K. Pietroszek, “Real-time gesture animation generation from speech for virtual human interaction,” in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–4, 2021.