A Bilingual Social Robot with Sign Language and Natural Language

Xiaoxuan Hei
Autonomous Systems and Robotics Lab, U2IS, ENSTA Paris, Institut Polytechnique de Paris
Palaiseau, France
xiaoxuan.hei@ensta-paris.fr

Chuang Yu
UCL Interaction Centre, Computer Science Department, University College London
London, United Kingdom
chuang.yu@ucl.ac.uk

Heng Zhang
Autonomous Systems and Robotics Lab, U2IS, ENSTA Paris, Institut Polytechnique de Paris
Palaiseau, France
heng.zhang@ensta-paris.fr

Adriana Tapus
Autonomous Systems and Robotics Lab, U2IS, ENSTA Paris Institut Polytechnique de Paris
Palaiseau, France
adriana.tapus@ensta-paris.fr

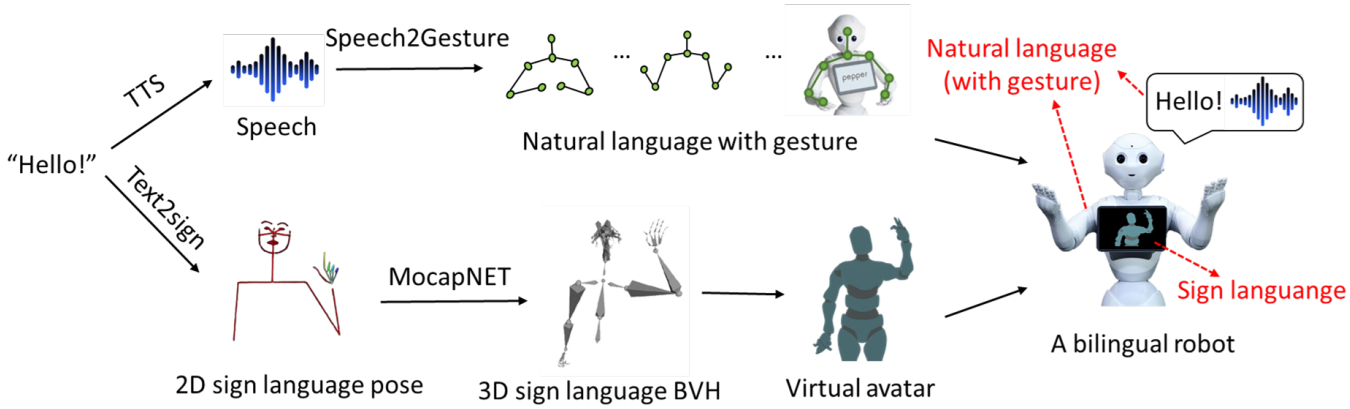

Figure 1: Framework Pipeline. The upper section outlines the generation process for natural language speech and gestures. The lower part shows the generation process for sign language animations of 3D virtual agent. Both outputs are simultaneously presented with the Pepper robot.

ABSTRACT

In situations where both deaf and non-deaf individuals are present in a public setting, it would be advantageous for a robot to communicate using both sign and natural languages simultaneously. This would not only address the needs for diverse users but also pave the way for a richer and more inclusive spectrum of human-robot interactions. To achieve this, a framework for a bilingual robot has been proposed in this paper. The robot exhibits the ability to articulate messages in spoken language, complemented by non-verbal cues such as expressive gestures, all while concurrently conveying information through sign language. The system can generate natural language expressions with speech audio, spontaneous prosody-based gestures, and sign language displayed on a virtual avatar on a robot's screen. The preliminary findings from this research showcase the robot's capacity to seamlessly blend natural language expressions with synchronized gestures and sign language, underlining its potential to revolutionize communication dynamics in diverse settings.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design**; • **Computer systems organization** → **Robotics**.

KEYWORDS

Human-robot interaction, sign language, gesture generation, virtual agent

ACM Reference Format:
Xiaoxuan Hei, Chuang Yu, Heng Zhang, and Adriana Tapus. 2024. A Bilingual Social Robot with Sign Language and Natural Language. In *Companion*

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03…$15.00
https://doi.org/10.1145/3610978.3640549

of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA. ACM, 4 pages. https://doi.org/10.1145/3610978.3640549

1 INTRODUCTION

As technology advances, robots are becoming increasingly integrated into our society, finding use in education [1][6], therapy [4][3], entertainment [21], and other fields. Both in human-human interaction and human-robot interaction (HRI) field, speech is the most friendly and natural mode of communication [20], which facilitates clear and conversational exchanges, while the expressiveness of tone and emotion contributes to a more engaging interaction [2]. Simultaneously, gestures are an integral part of human communication, and when robots can engage in gesture-based communication, it enhances the naturalness and rapport between humans and robots [24][22]. Gestures offer a visual dimension, complementing speech to convey information [9], express emotions [8], and guide actions [10], making robots more human-like and engaging [19]. Together, these modalities create a more intuitive and relatable HRI experience.

Nevertheless, when deaf individuals are present, speech and gestures might prove ineffective as means of communication. Gibson [5] illustrated that individuals experiencing impairment in one sensory system often develop increased proficiency in alternative sensory channels as a compensatory mechanism. This phenomenon, referred to as sensory compensation, implies that individuals with deafness may display heightened sensitivity in visual and tactile modalities. In many countries, sign language is the first language for people with hearing loss [11]. Therefore, in situations where deaf or hearing-impaired and non-deaf individuals are involved at the same time, a robot that can communicate in both natural and sign languages is of great significance. It can help bridge the communication gap between those who primarily use spoken language and those who rely on sign language, allowing everyone to effectively interact and participate in various social, educational, and professional contexts.

This paper presents a preliminary study aimed at displaying an avatar with sign language and subtitles on the tablet of Pepper robot, while simultaneously enabling Pepper to speak natural language with accompanying gestures, since the degree of freedom of the robot limits our direct use of its hands for sign language. The multimodal nature of the robot's communication system ensures that it can cater to diverse audiences, accommodating both sign language users and individuals proficient in spoken language.

2 RELATED WORK

In recent years, the application of robots in sign language has attracted considerable attention. Meghdari et al. [12] designed a humanoid robot which has a upper-body of 29 actuated degrees of freedom for teaching Persian sign language to hearing-impaired children, while Nandy et al. [14] proposed a new method for recognizing Indian sign language with HOAP-2 robot. Additionally, Homburg et al. [7] investigated the capabilities of humanoid robots in sign language translation. They experimented by 3D printing two arms for the InMoov robot, enabling it to execute German sign language. Thinh et al. [18] designed a robot system which could translate Vietnamese sign language into Vietnamese spoken language and recognize Vietnamese speech to sign language. In addition, Saunders et al. [17] introduced Progressive Transformers, an autoregressive model designed to translate discrete spoken language sentences into continuous 3D sign pose sequences. However, as far as we know, no previous work has been worked on simultaneously displaying spoken language and sign language.

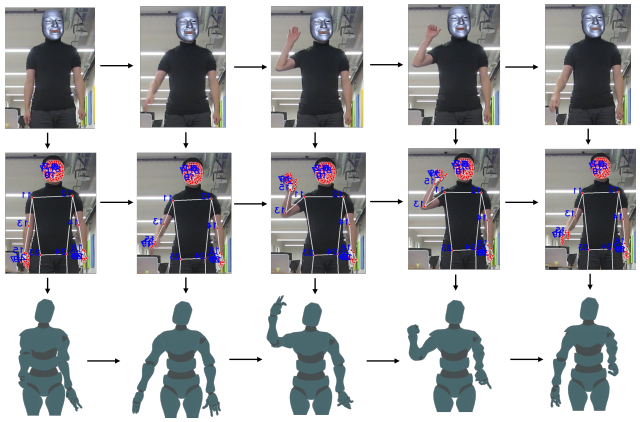

Figure 2: A sequence of gestures corresponding to a sequence of 2D poses

3 METHODOLOGY

The diagram in Figure 1 illustrates the pipeline used in our framework. The upper segment delineates the procedure for generating natural language speech and gestures, while the lower segment depicts the process for creating sign language animations for a 3D virtual agent. Ultimately, both outputs are concurrently displayed using the Pepper robot. In summary, when provided with a text input, the framework generates robot speech, gestures, and virtual avatar sign language, aligning them seamlessly.

3.1 Natural language and gesture generation

In this study, first we use Microsoft Azure TTS (Text to Speech) [1] to generate speech, which enables the generation of natural language gestures through the gesture generation model proposed by Yu [23]. This model consists of a generator and a discriminator. The generator comprises a temporal encoder and a temporal decoder. The encoder takes the speech audio as input and produces the final hidden state, which is then used as input for the subsequent decoder. The decoder combines the encoder output with random noise to map it into a corresponding gesture. On the other hand, the discriminator takes as input either the generated gesture or the ground-truth gesture along with spontaneous speech audio. Its role is to determine whether the speech and the gesture are coherent and match each other. By training the discriminator, the model learns to generate more realistic and accurate gestures that align with the provided speech.

[1] https://azure.microsoft.com/en-us/services/cognitive-services/text-tospeech/

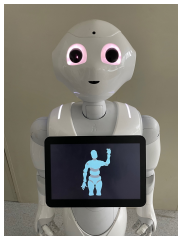

Figure 3: A sample of Pepper robot with sign language avatar

3.2 Sign language generation

In our framework, the Text2sign [13] tool is used to obtain 2D sign language poses from the text. MocapNet [16] is a method to estimate 3D human pose based on 2D human joints positions. With the 2D sign language poses originated before as input, MocapNet will generate 3D sign language in BVH format, which is then applied to a 3D human mesh obtained from Mixamo2 [2], an open database of animated 3D characters.

For the preliminary study, we utilized images captured by a 2D RGB camera to evaluate MocapNet's capability to generate 3D poses from 2D poses. The sequence of gestures produced is depicted in Figure 2, where the first row exhibits the 2D RGB images, the second row illustrates the 3D joint estimation, and the third row showcases the 3D pose with mesh. The avatar is then displayed on the tablet of the Pepper robot [15], as shown in Figure 3.

3.3 Natural language and sign language alignment

After generating each component separately, namely speech, gesture, and sign language videos, it is crucial to synchronize them effectively. Although our preliminary study has not yet encompassed this aspect, we acknowledge its significance and plan to address it in our future work. The synchronization process entails aligning the timing and cohesiveness of the generated speech, gesture, and sign language videos to ensure a seamless and coherent multimodal output.

Synchronizing the components involves careful coordination and integration to ensure that the gestures accurately correspond to the speech and align with the intended meaning. This synchronization is essential for facilitating effective communication and comprehension between users. It enables a more natural and seamless experience, promoting inclusivity and accessibility for users, especially for individuals who rely on sign language as their primary mode of communication.

4 DISCUSSION AND FUTURE WORKS

Our initial findings indicate that the approach we proposed is feasible, although more extensive research such as user study should be required to validate the effectiveness of the system.

In our future research endeavors, we will dedicate efforts to develop methodologies and algorithms for robust synchronization. We will explore techniques such as temporal alignment, motion mapping, and linguistic cues to enhance the synchronization accuracy and overall quality of the generated multimodal output. A sign language recognition algorithm will also be helpful to build a real conversation with hearing impaired individuals.

Additionally, we are excited to explore the integration of Large Language Models (LLMs) into our system. LLMs have emerged as powerful tools in natural language processing, enabling machines to comprehend and generate human-like text with remarkable accuracy. By leveraging LLMs, we aspire to enhance our robot's ability to recognize and respond effectively to a diverse range of questions posed by human users. This integration would enable our system to handle nuanced inquiries, adapt to different linguistic styles, and generate more contextually relevant and coherent responses.

Overall, our future work will focus on achieving seamless alignment between spoken language and sign language, while leveraging the power of LLMs to enhance our robot's language comprehension and response generation. By embracing these advancements, we strive to create a more inclusive and interactive communication platform, bridging the gap between different modalities and facilitating meaningful interactions between humans and our robot.

[2]https://www.mixamo.com/

REFERENCES

[1] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science robotics* 3, 21 (2018), eaat5954.

[2] Cynthia Breazeal. 2003. Emotion and sociable humanoid robots. *International journal of human-computer studies* 59, 1-2 (2003), 119–155.

[3] Pauline Chevalier, Jean-Claude Martin, Brice Isableu, Christophe Bazile, David-Octavian Iacob, and Adriana Tapus. 2016. Joint attention using human-robot interaction: Impact of sensory preferences of children with autism. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 849–854.

[4] Daniel Hernández García, Pablo G Esteban, Hee Rin Lee, Marta Romeo, Emmanuel Senft, and Erik Billing. 2019. Social robots in therapy and care. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 669–670.

[5] Eleanor Jack Gibson. 1969. Principles of perceptual learning and development. (1969).

[6] Xiaoxuan Hei, Valentine Denis, Pierre-Henri Oréfice, Alia Afyouni, Paul Laborde, Damien Legois, Ioana Ocnarescu, Margarita Anastassova, and Adriana Tapus. 2023. Evaluating Students' Experiences in Hybrid Learning Environments: A Comparative Analysis of Kubi and Double Telepresence Robots. In *International Conference on Social Robotics*. Springer, 148–159.

[7] Daniel Homburg, Mirja Sophie Thieme, Johannes Völker, and Ruth Stock. 2019. Robotalk-prototyping a humanoid robot as speech-to-sign language translator. (2019).

[8] Michelle Karg, Ali-Akbar Samadani, Rob Gorbet, Kolja Kühnlenz, Jesse Hoey, and Dana Kulić. 2013. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing* 4, 4 (2013), 341–359.

[9] Sotaro Kita. 2000. How representational gestures help speaking. *Language and gesture* 1 (2000), 162–185.

[10] Yoshinori Kuno, Kazuhisa Sadazuka, Michie Kawashima, Keiichi Yamazaki, Akiko Yamazaki, and Hideaki Kuzuoka. 2007. Museum guide robot based on sociological interaction analysis. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1191–1194.

[11] Penelope Markellou, Maria Rigou, Spiros Sirmakessis, and Athanasios Tsakalidis. 2000. A Web adaptive educational system for people with hearing difficulties. *Education and Information Technologies* 5 (2000), 189–200.

[12] Ali Meghdari, Minoo Alemi, Mohammad Zakipour, and Seyed Amir Kashanian. 2019. Design and realization of a sign language educational humanoid robot. *Journal of Intelligent & Robotic Systems* 95 (2019), 3–17.

[13] Amit Moryossef and Yoav Goldberg. 2021. Sign Language Processing. https://sign-language-processing.github.io/.

[14] Anup Nandy, Soumik Mondal, Jay Shankar Prasad, Pavan Chakraborty, and GC Nandi. 2010. Recognizing & interpreting indian sign language gesture for human robot interaction. In *2010 international conference on computer and communication technology (ICCCT)*. IEEE, 712–717.

[15] Amit Kumar Pandey and Rodolphe Gelin. 2018. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine* 25, 3 (2018), 40–48.

[16] Ammar Qammaz and Antonis A Argyros. 2019. MocapNET: Ensemble of SNN Encoders for 3D Human Pose Estimation in RGB Images. In *British Machine Vision Conference (BMVC 2019)*. BMVA, Cardiff, UK. http://users.ics.forth.gr/argyros/res_mocapnet.html

[17] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 687–705.

[18] Nguyen Truong Thinh, Tuong Phuoc Tho, Truong Cong Toai, and Le Thanh Ben. 2019. Robot supporting for deaf and less hearing people. In *Robot Intelligence Technology and Applications 5: Results from the 5th International Conference on Robot Intelligence Technology and Applications 5*. Springer, 283–289.

[19] Nguyen Tan Viet Tuyen, Armagan Elibol, and Nak Young Chong. 2020. Learning bodily expression of emotion for social robots through human interaction. *IEEE Transactions on Cognitive and Developmental Systems* 13, 1 (2020), 16–30.

[20] Jorge Wuth, Pedro Correa, Tomás Núñez, Matías Saavedra, and Néstor Becerra Yoma. 2021. The role of speech technology in user perception and context acquisition in hri. *International Journal of Social Robotics* 13 (2021), 949–968.

[21] Shigeo Yoshida, Takumi Shirokura, Yuta Sugiura, Daisuke Sakamoto, Tetsuo Ono, Masahiko Inami, and Takeo Igarashi. 2015. RoboJockey: designing an entertainment experience with robots. *IEEE computer graphics and applications* 36, 1 (2015), 62–69.

[22] Chuang Yu and Adriana Tapus. 2019. Interactive robot learning for multimodal emotion recognition. In *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*. Springer, 633–642.

[23] Chuang Yu and Adriana Tapus. 2020. SRG 3: Speech-driven Robot Gesture Generation with GAN. In *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 759–766.

[24] Heng Zhang, Chuang Yu, and Adriana Tapus. 2022. Towards a Framework for Social Robot Co-speech Gesture Generation with Semantic Expression. In *International Conference on Social Robotics*. Springer, 110–119.