

Towards Multimodal Speech and SVG Image Coordination for Input/Output Communication

Massimo Donini
massimo.donini@unito.it
Computer Science Dept.
University of Turin, Italy

Cristina Gena
cristina.gena@unito.it
Computer Science Dept.
University of Turin, Italy

Alessandro Mazzei
alessandro.mazzei@unito.it
Computer Science Dept.
University of Turin, Italy

Matteo Nazzario
matteo.nazzario@intesanpaolo.com
Intesa SanPaolo Innovation Center
Turin, Italy

Irene Borgini
irene.borgini@intesanpaolo.com
Intesa SanPaolo Innovation Center
Turin, Italy

Abstract

This work introduces a novel use of the humanoid robot Pepper in educational settings, emphasizing dialog-driven and multimodal interaction to support the teaching of abstract concepts. In contrast to conventional lecture-based approaches, the proposed system enables learner-centered activities in which students can engage in spontaneous conversations with the robot while interacting with visual material presented on its tablet. The robot addresses learners' questions by coordinating spoken responses with synchronized textual information and adaptive visual feedback, including real-time updates to vector-based images that visually emphasize key elements. By dynamically aligning instructional content with the learner's interests and questions, this approach seeks to promote a more engaging and interactive learning experience.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Natural language interfaces**; • **Computing methodologies** → **Cognitive robotics**; **Discourse, dialogue and pragmatics**.

Keywords

Multimodal interaction, Human-Robot Interaction, Educational robotics, SVG, Pepper robot, Dialogue systems

ACM Reference Format:

Massimo Donini, Cristina Gena, Alessandro Mazzei, Matteo Nazzario, and Irene Borgini. 2026. Towards Multimodal Speech and SVG Image Coordination for Input/Output Communication. In *Proceedings of HRI 2026 Workshop on Empowering Human-Robot Communication Through Non-Linguistic Pathways (HuRoCo @ HRI 2026)*. (HuRoCo @ HRI '26). ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HuRoCo @ HRI '26, Edinburgh, Scotland, UK

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

In recent years, social robots have attracted growing interest in educational contexts as a means to support more engaging, personalized, and interactive learning experiences [6]. This paper advocates the use of the humanoid robot Pepper in dialogic, learner-centered lessons for teaching abstract concepts, with a particular focus on the coordinated use of speech and adaptive visual representations.

We propose an exploratory learning paradigm in which students actively engage with visual content displayed on the robot's tablet while interacting with the robot through spontaneous dialogue. Rather than relying on static diagrams or pre-scripted presentations, our approach emphasizes real-time multimodal coordination, where the robot's spoken explanations are synchronized with dynamic visual modifications that highlight relevant elements of the displayed content.

A key aspect of our approach is the use of Scalable Vector Graphics (SVG) to enable fine-grained, structural manipulation of visual elements during interaction. By programmatically modifying specific components of vector-based diagrams in response to user questions, the robot can visually ground its explanations, for example by highlighting, resizing, or relabeling elements as they are discussed. We contend that this form of speech-SVG coordination represents a promising direction for improving clarity, engagement, and personalization in educational human-robot interaction.

The goal of this paper is twofold. First, we outline a conceptual framework for structured multimodal communication in educational robots, emphasizing the role of coordinated speech, text, and vector-based visuals. Second, we illustrate this vision through a proof-of-concept example in the domain of set theory, demonstrating how dynamic visual grounding can support the explanation of abstract relations. Together, these contributions aim to stimulate discussion on the design of learner-centered, multimodal interaction paradigms for educational robots.

The paper is structured as follows: Section 2 discusses related work, Section 3 describes the robot-centered software/hardware pipeline for answering/ the user's questions, Section 4 presents examples of interaction with the system and Section 5 concludes the paper.

2 Related Work

Interest in the integration of natural language with multimedia content has grown in Human-Robot Interaction (HRI), particularly

in assistive and educational robotics. Multimodal dialogue systems combining speech with visual cues have been shown to enhance engagement and interaction effectiveness. For example, Sun et al. demonstrated the use of images not only to support text but also as an alternative communication modality [13].

In educational contexts, video-based modeling techniques, such as self-modeling, peer modeling, and video prompting—improve learners’ ability to follow sequential instructions and acquire skills [2, 5, 7, 14]. Beyond learning, aligning spoken explanations with visual content has been applied in cultural and therapeutic settings, showing improved user experience and engagement [5, 12].

Rule-based dialogue systems, like AIML, provide predictable and controllable responses, which are important in sensitive domains such as education or therapy [8]. While Large Language Models (LLMs) achieve broad NLP capabilities but their tendency to generate inaccurate or hallucinated outputs limits their applicability where correctness is critical [11].

Multimodal signaling in embodied agents further improves interaction quality. Zhang et al. found that agents providing real-time visual feedback—such as highlighting referenced objects—reduced errors and increased user confidence during task execution [15]. These results highlight the value of integrating verbal and visual cues for effective human-agent collaboration.

Our previous work explored multimodal strategies for HRI in educational settings [3], investigating aspects such as the number of visual elements, timing relative to speech, pauses and element sizing. A subsequent study addressed dynamic personalization of multimedia content, filtering images based on user profiles to improve engagement [4]. Both approaches relied on raster images, which allow holistic modifications but cannot alter sub-elements for detailed coordination.

The present work builds on these foundations by using Scalable Vector Graphics (SVG) to enable fine-grained, dynamic modifications of visual elements in sync with the robot’s speech. Vector-based images allow internal manipulation of fonts, colors and styles, supporting both multimodal coordination and user-adaptive personalization. Prior exploration of SVG integration with AIML demonstrated the feasibility of this approach [10], which we extend here in the context of educational HRI.

3 Interaction workflow and Architecture

To enable dialog-based lessons with the Pepper robot, we developed a workflow pipeline illustrated in Figure 1. This pipeline comprises several steps: recording the user’s voice and saving it as an audio file, sending the audio to a speech-to-text service, receiving the transcription in the user’s native language and textual format, sending the transcript to a dialog server, retrieving a JSON file containing both the robot’s verbal response and the modifications to apply to the image during the explanation and finally delivering a multimodal response to the user via Pepper. In addition to interacting with the robot through speech, the user can also request an explanation by directly tapping on an element of interest within the SVG image displayed on the robot’s tablet.

This workflow is supported by three main components:

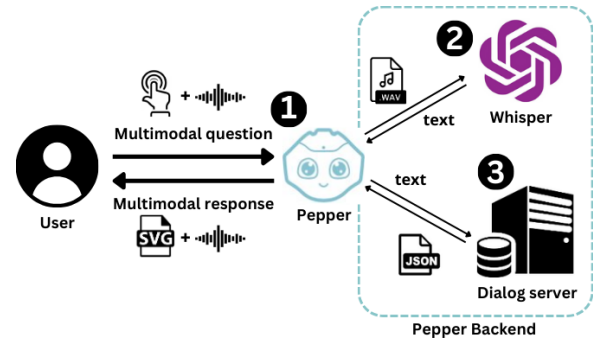


Figure 1: The workflow pipeline’s information flow.

- (1) **Pepper Robot:** We used Pepper’s built-in microphone to capture and store the user’s speech. To ensure accurate segmentation of the audio, we implemented a custom algorithm based on a volume threshold that detects the start and end of speech activity. This allowed for precise audio trimming, ensuring that only relevant speech segments were sent for transcription.
- (2) **Whisper (Speech-to-Text Service):** We developed a Python-based application to manage audio capture and communication with the transcription system. We deployed OpenAI’s Whisper¹ model (medium version) on a virtual machine hosted at the High-Performance Computing Center of the Computer Science Department at the University of Turin. A dedicated REST API server was set up to receive audio files and return high-quality transcriptions in text format.
- (3) **Dialog Server:** Running on a separate machine within the Computer Science Department and accessible via REST API, the dialog server processes user queries and returns a structured JSON file. This file contains the robot’s verbal response as well as instructions for dynamically modifying the image shown on Pepper’s tablet during the response.

Once the JSON file is received, Pepper provides a multimodal explanation using the communication strategies identified in our previous work [3]. During speech synthesis, the robot briefly pauses on key terms and applies the corresponding visual modifications (e.g., changing the color of a set element), followed by a short delay to allow users to process the updated visual context. Spoken utterances are also displayed as synchronized captions on the tablet, reinforcing the explanation through multiple modalities.

To support the coordination between SVG images and dialogue, we adopted NoVABOT, a dialogue system extending AIML with dedicated support for SVG manipulation [10]. Unlike standard AIML-based systems [1, 9], NoVABOT enables the selection and modification of specific SVG elements through additional tags, allowing visual changes to be directly synchronized with the robot’s responses.

As illustrated in Figure 2, NoVABOT introduces dedicated tags to associate dialogue rules with SVG images and to highlight specific visual components, enabling fine-grained coordination between speech and vector-based visuals.

¹<https://openai.com/index/whisper/>

```

<category>
  <pattern>
    * ELEMENTS * SECOND CIRCLE * CONNECTED * ARROW *
  </pattern>
  <template>
    Each element of the second circle (in the
    codomain) has at least one incoming arrow
    <svgElement style-name="fill"
    style-value="#04ed00">set-domain</svgElement>
  </template>
</category>

```

Figure 2: Example of an extended AIML rule for handling SVG images.

```

{
  "query": "Are there elements of the second circle
  connected by any arrows?",
  "answer": "Each element of the second circle (in the
  codomain) receives at least one arrow.",
  "id_elements": ["#set-domain"],
  "style_names": ["fill"],
  "style_values": ["#04ed00"]
}

```

Figure 3: Example of a JSON response from NoVABOT.

This AIML category matches input such as *elements of the second circle connected by arrows.*, which are specified in the pattern tag. The template then enables the system to respond with *each element of the second circle (in the codomain) has at least one incoming arrow*, while visually highlighting the codomain set.

In our architecture, for each user input, we send a POST request to NoVABOT, using the output from the speech recognizer as input. In response, the API returns a JSON object containing the system’s answer and the elements to highlight in the SVG image. An example of the API output is shown in Figure 3.

The presence of `id_elements`, `style_names`, and `style_values` in the API response enables us to highlight specific elements in the SVG image during the conversation, as shown in Figure 4.

4 Examples of Multimodal Interaction

This section presents examples of interactions with the developed system, focusing on a scenario involving mathematical sets. We describe in detail three multimodal responses generated by the robot in reply to typical student questions about the sets. In all the cases, the interaction begins after the robot has recorded the user’s question, transcribed the audio using the speech-to-text service and obtained a response from the dialog server.

- (1) *User question: “What are the elements of circle A?”*. In the case illustrated in Figure 4a, the dialog server returns the following response: *“In the first circle, the domain, there are 4 elements: Anna, Mario, Luisa and Fabio”*. Since the focal element is mentioned at the beginning of the sentence, the robot’s response is delivered in two stages. First, it pronounces the initial segment: *“In the first circle, the domain ...”*, then it

highlights the relevant graphical element within the SVG image to visually emphasize it (Figure 4b). After a brief pause, the robot completes the sentence by pronouncing the rest of the message. Simultaneously, the words being spoken are displayed as captions on Pepper’s tablet, synchronized with the speech.

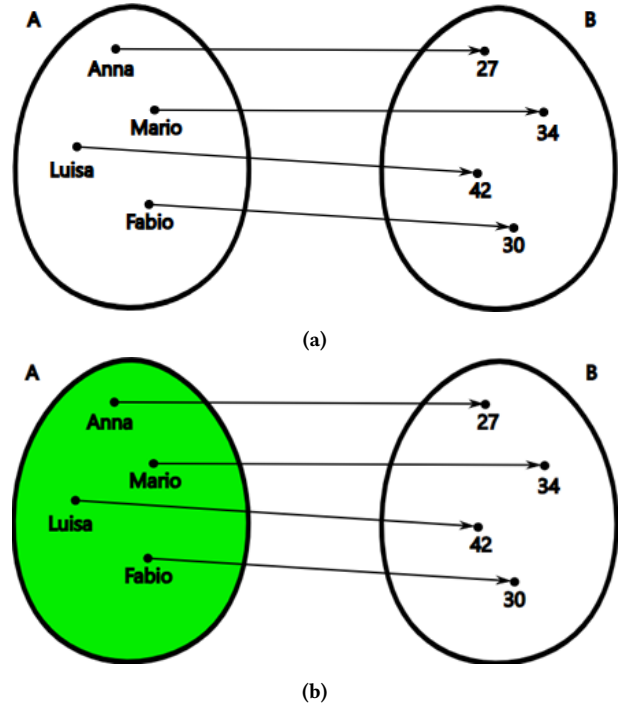


Figure 4: Depiction of the Pepper robot’s tablet at key moments of multimodal communication for the sentence “In the first circle, the domain, there are 4 elements: Anna, Mario, Luisa and Fabio.”

- (2) *User question: “Who is connected to element 42?”*. In the case illustrated in Figure 5a, the dialog server returns the following response: *“The elements Luigina and Remo of set A are connected to element 42 of set B”*. Here, the robot proceeds step-by-step through the sentence. It first says *“The elements Luigina”*, then highlights the corresponding element in the SVG image (Figure 5b), followed by a short pause. It continues with *“and Remo,”* again applying a visual emphasis to the element before pausing briefly (Figure 5c). Finally, it delivers the remainder of the sentence. As in the previous example, all spoken words are simultaneously displayed on the tablet to provide a synchronized textual channel alongside the verbal and visual ones.

These examples demonstrate the system’s ability to coordinate verbal, textual and visual modalities in real time, dynamically adapting the explanation structure based on the position and number of visual elements to be emphasized as analyzed in our previous work [3].

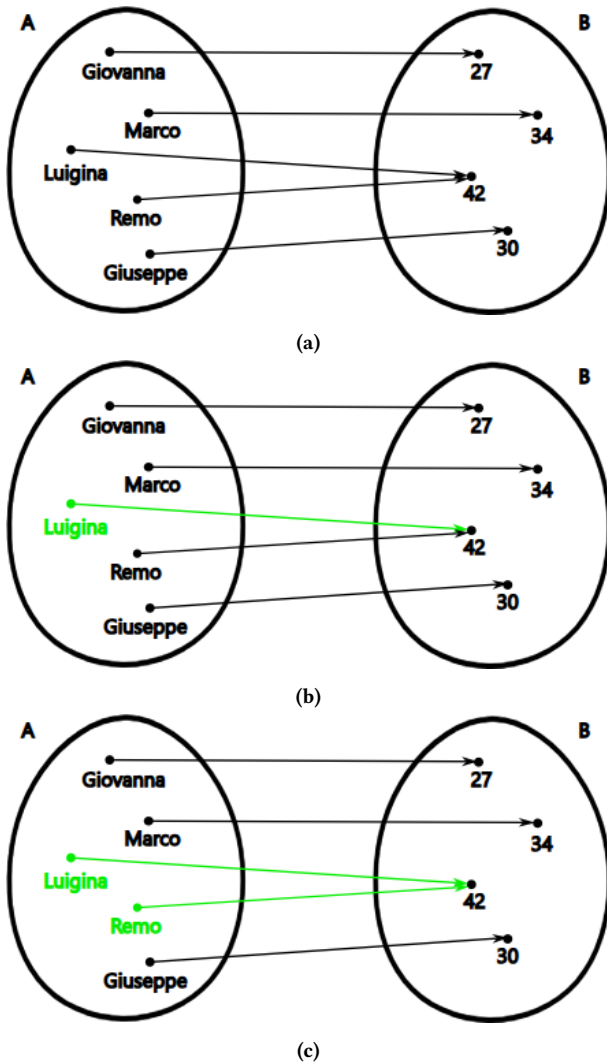


Figure 5: Depiction of the Pepper robot’s tablet at key moments of multimodal communication for the sentence “The elements Luigina and Remo of set A are connected to element 42 of set B.”

5 Conclusions and Ongoing Work

This paper presented initial results toward a multimodal dialogical educational robot system based on the coordinated use of speech, touch input and vector-based visual content. The use of SVG images enables the modification of internal visual components and the coordination of these changes with Pepper’s speech, thereby supporting structured multimodal communication.

The approach was illustrated through examples in which Pepper explains functional relations between sets, highlighting the potential of vector-based visuals for grounding abstract explanations.

Future work will include user studies to assess usability and learning impact, the integration of large language models to enhance dialogue capabilities and the exploration of alternative visual

highlighting strategies to better support personalization and accessibility. The system will also be extended to additional educational domains to evaluate the generalizability of the proposed multimodal framework.

References

- [1] Pier Felice Balestrucci, Elisa Di Nuovo, Manuela Sanguinetti, Luca Anselma, Cristian Bernareggi, and Alessandro Mazzei. 2024. An Educational Dialogue System for Visually Impaired People. *IEEE Access* 12 (2024), 49800–49812. doi:10.1109/ACCESS.2024.3392509
- [2] David Cihak, Paul A Alberto, Teresa Taber-Doughty, and Robert I Gama. 2006. A comparison of static picture prompting and video prompting simulation strategies using group instructional procedures. *Focus on Autism and Other Developmental Disabilities* 21, 2 (2006), 89–99.
- [3] Massimo Donini, Cristina Gena, and Alessandro Mazzei. 2024. Multimodal Strategies for Robot-to-Human Communication. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, 417–421.
- [4] Massimo Donini, Cristina Gena, Alessandro Mazzei, Irene Borgini, and Matteo Nazzario. 2024. Dynamic Personalization of Multimedia Content Based on User Model. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. Association for Computing Machinery, New York, NY, USA, 211–214.
- [5] Cristina Gena, Claudio Mattutino, Andrea Maieli, Elisabetta Miraglio, Giulia Ricciardiello, Rossana Damiano, and Alessandro Mazzei. 2021. Autistic Children’s Mental Model of an Humanoid Robot. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 128–129.
- [6] Cristina Gena, Claudio Mattutino, Gianluca Perosino, Massimo Trainito, Chiara Vaudano, and Davide Cellie. 2020. Design and Development of a Social, Educational and Affective Robot. In *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems, EAIS 2020, Bari, Italy, May 27-29, 2020*. IEEE, 1–8. doi:10.1109/EAIS48028.2020.9122778
- [7] Kathleen McCoy and Emily Hermansen. 2007. Video modeling for individuals with autism: A review of model types and effects. *Education and Treatment of Children* 30, 4 (2007), 123–140. doi:10.1353/etc.2007.0029
- [8] Michael McTear. 2022. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*. Springer Nature, Cham, Switzerland.
- [9] Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface: Talking to Smart Devices* (1st ed.). Springer Publishing Company, Incorporated, Cham, Switzerland.
- [10] Michael Oliverio, Margherita Piroi, Daniele De Giorgi, Pier Felice Balestrucci, Carola Manolino, Alessandro Mazzei, Luca Anselma, Cristian Bernareggi, Marina Serio, Cristina Sabena, Tiziana Armano, Sandro Coriasco, and Anna Capietto. 2024. NoVAGraphS: Towards an Accessible Educational-Oriented Dialogue System. In *Proceedings of the Second International Workshop on Artificial Intelligent Systems in Education (AIxEDU 2024) co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024)*, Vol. 3879. CEUR-WS.org, Aachen, Germany, 1–9. https://ceur-ws.org/Vol-3879/AIxEDU2024_paper_34.pdf
- [11] Junaid Qadir. 2023. Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education. In *2023 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, IEEE, Piscataway, NJ, USA, 1–9.
- [12] Antonio Sorgente, Paolo Vanacore, Antonio Origlia, Enrico Leone, Francesco Cutugno, and Francesco Mele. 2016. Multimedia Responses in Natural Language Dialogues. In *Proceedings of the Workshop on Advanced Visual Interfaces: Smart Interactions (AVI*CH)*. CEUR-WS.org, Napoli, Italy, 15–18. https://www.researchgate.net/publication/311911354_Multimedia_Responses_in_Natural_Language_Dialogues
- [13] Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2021. Multimodal Dialogue Response Generation. *arXiv preprint* 2110.08515 (2021), 1–10. <https://arxiv.org/abs/2110.08515> arXiv:2110.08515
- [14] Connie Wong, Samuel L Odom, Kara A Hume, Ann W Cox, Angel Fetti, Suzanne Kucharczyk, Matthew E Brock, Joshua B Plavnick, Veronica P Fleury, and Tia R Schultz. 2015. Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of autism and developmental disorders* 45 (2015), 1951–1966.
- [15] Tianyi Zhang, Colin Au Yeung, Emily Aurelia, Yuki Onishi, Neil Chulpongatarn, Jiannan Li, and Anthony Tang. 2025. Prompting an Embodied AI Agent: How Embodiment and Multimodal Signaling Affects Prompting Behaviour. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–25.