

Enhancing Social Robot Engagement through Real-Time Backchannel Generation

Stef Laurent

Student number: 02105477

Supervisor: Prof. dr. Tony Belpaeme

Counsellors: Maria Jose Pinto Bernal, Ir. Eva Verhelst

Master's dissertation submitted in order to obtain the degree of Master of Science in Computer Science Engineering

Academic year 2025-2026

Verbetering van Betrokkenheid van Sociale Robots door het Genereren van Real-Time Backchannels

Stef Laurent

Studentennummer: 02105477

Promotor: prof. dr. Tony Belpaeme

Begeleiders: Maria Jose Pinto Bernal, ir. Eva Verhelst

Masterproef voorgelegd voor het behalen van de graad Master of Science in Computer Science Engineering

Academiejaar 2025-2026

Acknowledgements

First off, I want to say a huge thank you to Maria José Pinto Bernal and Eva Verhelst for their incredible support and dedication throughout this whole journey. The amount of hours you both spent in meetings, answering my endless questions, and troubleshooting the system along the way is truly amazing. I couldn't have asked for better guidance.

I would also like to thank Prof. Dr. Tony Belpaeme for giving me the unique opportunity to work on this fascinating topic, and for always checking in with genuine curiosity and insightful questions whenever we crossed paths.

A massive thank you goes out to the entire social robotics team. The fun talks, laughs, and constant interest in my work over the semester really kept me motivated to push through and get the best possible result. I genuinely loved hanging out on the 8th floor, whether I was deep in thesis work or just taking a break to chat about related (and completely unrelated) things.

I am also incredibly grateful to all the participants who volunteered to sit down and chat with a robot for a whole hour! Your time and feedback gave me the valuable insights needed to evaluate this system, and this project wouldn't exist without you.

And with that, the stage is set! I really hope you enjoy diving into the pages ahead, have some fun reading it, and maybe even learn a few new things along the way.

Stef Laurent, May 2026

Admission to loan

“The author(s) gives (give) permission to make this master’s dissertation available for consultation and to copy parts of this master’s dissertation for personal use. In the case of any other use, the copyright terms have to be respected, in particular with regard to the obligation to state expressly the source when quoting results from this master’s dissertation.”

Toegang tot bruikleen

“De auteur(s) geeft (geven) de toelating dit afstudeerwerk voor consultatie beschikbaar te stellen en delen van het afstudeerwerk te kopiëren voor persoonlijk gebruik. Elk ander gebruik valt onder de beperkingen van het auteursrecht, in het bijzonder met betrekking tot de verplichting de bron uitdrukkelijk te vermelden bij het aanhalen van resultaten uit dit afstudeerwerk.”

Stef Laurent, May 2026

Thesis as part of an exam

“This master’s dissertation is part of an exam. Any comments formulated by the assessment committee during the oral presentation of the master’s dissertation are not included in this text.”

Stef Laurent, May 2026

Enhancing Social Robot Engagement through Real-Time Backchannel Generation

by
STEF LAURENT

Master's dissertation submitted in order to obtain the academic degree of Master of
Science in Computerscience Engineering

Supervisor: Prof. Dr. Tony Belpaeme
Counsellor: Maria Jose Pinto Bernal, Ir. Eva Verhelst

Faculty of Engineering and Architecture
Ghent University

Department of AIRO

Abstract - Contemporary paradigms in Human-Robot Interaction (HRI) frequently lack perceived naturalness due to the reactive, half-duplex nature of traditional dialogue systems. While advancements in Large Language Models (LLMs) have refined content generation, the structural execution of interactions remains artificial, particularly regarding the absence of backchannels and the inability to handle natural interruptions. This research addresses these interactional gaps by proposing a fully functional voice-agent architecture capable of concurrent timing and typology prediction for backchannels in real-time environments. Unlike previous studies that focus primarily on temporal prediction, this study develops a dual prediction pipeline to determine both the optimal moment for a backchannel and its specific type, whether context-dependent or not. The methodology involves an experimental approach where participants ($N = 20$) had to evaluate the system's efficacy on the perceived naturalness of timing algorithms and the impact of autonomous backchanneling on user engagement and rapport. By minimizing generation latency and integrating sophisticated listener feedback, this work seeks to transition robotic agents from passive listeners to proactive conversational partners.

Keywords - social robots, long-term engagement, voice-agent design, backchannels

Enhancing Social Robot Engagement through Real-Time Backchannel Generation

Stef Laurent

Prof. Dr. Tony Belpaeme

Maria Jose Pinto Bernal, Ir. Eva Verhelst

Abstract—Contemporary paradigms in Human-Robot Interaction (HRI) frequently lack perceived naturalness due to the reactive, half-duplex nature of traditional dialogue systems. While advancements in Large Language Models (LLMs) have refined content generation, the structural execution of interactions remains artificial, particularly regarding the absence of backchannels and the inability to handle natural interruptions. This research addresses these interactional gaps by proposing a fully functional voice-agent architecture capable of concurrent timing and typology prediction for backchannels in real-time environments. Unlike previous studies that focus primarily on temporal prediction, this study develops a dual prediction pipeline to determine both the optimal moment for a backchannel and its specific type, whether context-dependent or not. The methodology involves an experimental approach where participants (N=20) had to evaluate the system’s efficacy on the perceived naturalness of timing algorithms and the impact of autonomous backchanneling on user engagement and rapport. By minimizing generation latency and integrating sophisticated listener feedback, this work seeks to transition robotic agents from passive listeners to proactive conversational partners.

Index Terms—social robots, long-term engagement, voice-agent design, backchannels

I. INTRODUCTION

Current voice agents in Human-Robot Interaction (HRI) frequently lack conversational naturalness due to traditional “half-duplex” architectures. These systems rely on rigid turn-taking and require distinct periods of silence before responding, which introduces latency and disrupts fluid communication. While the integration of LLMs has improved content generation, the structural execution of these interactions remains flawed. Notably, agents struggle to process interruptions and fail to produce backchannels, the brief, non-lexical vocalizations essential for signalling active listening and user validation [3, 36].

Although low-latency predictive models are shifting systems toward full-duplex interactions (enabling simultaneous listening and speaking) [10, 19, 25], current backchannel generation models focus almost exclusively on temporal prediction [22], neglecting the context-dependent type of backchannel required.

This research addresses this gap by proposing a real-time voice-agent architecture capable of concurrent prediction of backchannel timing and typology. By optimizing this dual-prediction pipeline to minimize generation latency, we aim to facilitate highly fluid conversations. We evaluate the efficacy of this approach by addressing the following research questions:

- RQ1: To what extent does the autonomous generation of backchannels by a social robot enhance perceived naturalness and user engagement?

- RQ2: Which algorithmic approach for timing prediction is perceived as most natural by human interlocutors?

To answer these questions, this study outlines the development of the autonomous backchannel module and its validation through an experimental setup, evaluating the perceived naturalness of the backchannels during human-robot interactions.

II. FUNCTIONAL ROLE OF BACKCHANNELS IN A CONVERSATION

Natural conversations rely on a highly synchronized repertoire of behaviours, including complex structural mechanisms like turn-taking and overlapping speech [27]. While automated turn-taking has evolved from relying on rigid acoustic-prosodic cues [14] to employing advanced semantic [21, 31], transformer-based [10, 19, 25] and multimodal predictive models [20, 29], the strategic management of overlapping speech, remains a persistent challenge in Human-Robot Interaction (HRI) [9, 13].

Speech overlaps are traditionally categorized as either competitive (interruptions attempting to take the floor) or cooperative (backchannels) [7]. Backchannels are ubiquitous, non-obligatory signals, ranging from non-lexical vocalizations (“mhm”) and context-dependent lexical responses (“exactly”) [6] to non-verbal gestures like head nodding, that indicate active listenership without disrupting the speaker [17, 24, 32, 36].

For robotic agents, generating appropriate backchannels is essential for maintaining communicative efficiency and natural conversational flow [23]. They provide constant validation of the robot’s attention and comprehension [27]. Crucially, the autonomous integration of backchannels increases user engagement and encourages longer conversational contributions [18]. It also allows robotic systems to replace artificial hardware status indicators (e.g., UI prompts or LEDs) with naturalistic social feedback, taking a vital step toward seamless, human-like interaction behaviour. Conversely, the absence or temporal misalignment of these cues can severely disrupt a user’s fluency and degrade the agent’s perceived social presence [12].

III. GENERATION OF BACKCHANNELS

The autonomous generation of backchannels requires solving two primary challenges: predicting the optimal timing (Backchannel Opportunity Prediction, or BOP) and determining the appropriate form (Backchannel Category Prediction, or BCP). Furthermore, both components must dynamically adapt to the conversational context.

A. Temporal Prediction

Identifying the precise moment for a backchannel involves analysing prosodic, acoustic, and lexical features. Rule-Based Models synthesize acoustic parameters (pitch, pause duration) and visual cues (gaze) using deterministic thresholds [26, 35]. While being highly efficient, they lack adaptability and require manual tuning for different contexts [6, 23]. On the other hand, Machine Learning Architectures capture complex audio patterns that rules overlook. Approaches include multi-task frameworks integrating Automatic Speech Recognition (ASR) [8], pre-trained Transformer models like VAP [11, 17], and regression trees focused on prosodic variables [22]. Additionally, fusing acoustic data with linguistic text and visual cues (e.g., facial expressions) can improve the accuracy of the timing prediction [20, 33].

B. Category Prediction

Backchannels are functionally categorized into continuers (non-lexical vocalizations) and assessments (context-dependent) [5, 17]. Selecting the correct type requires understanding conversational context, including the agent’s persona and thematic progression [23]. While earlier systems relied on deterministic keyword triggers for computational efficiency [5], modern dual-pipeline frameworks process context in parallel to inform typology without sacrificing speed [2].

C. Context Dependencies

Backchannel behaviour is intrinsically linked to sociocultural and situational variables [23]. Frequencies vary drastically, for example, Japanese speakers use nearly three times as many backchannels as English speakers [34]. Age and gender also significantly modulate interaction dynamics [4, 6, 32]. Additionally, the syntactic placement of a BOP plays a role, placing it at a definitive sentence ending requires a different backchannel type than one occurring mid-sentence [6]. And finally, the intensity and purpose of the exchange dictate backchannel rates. Interaction-focused conversations naturally yield higher backchannel frequencies compared to purely informational exchanges where one speaker dominates the floor [4].

IV. SYSTEM DESIGN AND BACKCHANNEL MODULE

To evaluate real-time backchanneling, we developed an autonomous AI voice-agent architecture integrated with the Pepper humanoid robotic platform. The design prioritizes conversational fluency by processing real-time speech and backchannel generation in parallel.

Pepper’s humanoid morphology significantly increases user engagement and social presence compared to stationary smart speakers [1]. To circumvent the robot’s legacy hardware constraints, a hybrid integration strategy was deployed: a tablet mounted on the robot handled audio I/O, while Pepper’s physical gestures (such as head nods) mimicked the actual robot’s behaviour.

A. Backchannel module

The specialized backchannel module operates as an asynchronous parallel processing layer decoupled from the main dialogue pipeline, minimizing latency. Both the backchannel timing and type are being predicted by this module.

1) *Backchannel Timing Prediction (BOP)*: Two primary methodology paradigms were evaluated to identify the ideal Backchannel Opportunity Points (BOP): A Rule-Based Heuristic, modelled after the algorithm from Ward et al. [35]. This approach extracts low-dimensional prosodic features (pitch contours, energy dynamics, silence durations) to track state triggers [26]. It features minimal computational overhead and minimal latency. Secondary, a Machine-Learning (ML) approach leverages the Voice Activity Projection (VAP) model, a transformer-based framework that predicts fine-grained voice activity probabilities over a future 2-second window [11]. By utilizing the MaAI project’s¹ fine-tuned VAP model on online conversation data [15, 17], the system shifts from reactive feature tracking to proactive prediction.

2) *Backchannel Type Prediction (BCP)*: To classify whether a backchannel functions as a continuer or an assessment, a lightweight DistilBERT² model was fine-tuned on textual data. The model was trained using conversational text pairs extracted from the CANDOR Corpus, a dataset containing 850 hours of unscripted dyadic video interactions [28].

V. SYSTEM EVALUATION

The evaluation phase investigates user perceptions of backchanneling in Human-Robot Interaction (HRI). Our main goal for this user study is to find clear answers to both of our research questions. The study evaluated 20 participants (22.05 ± 1.16 years). To account for the significant socio-cultural variance in listener behaviour [12], we recorded each participant’s native language (*L1*) alongside a self-rated talkativeness scale. Interactions took place in a controlled meeting room using a Pepper robot. To bypass legacy hardware limitations of the Pepper robot and ensure precise Voice Activity Detection (VAD) and Automatic Speech Recognition (ASR), participants wore an external microphone.

After a brief familiarization phase, participants engaged in three 5-minute conversation blocks across the three conditions: Baseline (No backchannels), Rule-based backchannel timing and ML-based backchannel timing. Following each block, participants completed targeted questionnaires, completed with an end-of-session comparative evaluation and a brief semi-structured interview.

The efficacy of the backchannel modules is evaluated with subjective evaluations across four core constructs: Naturalness, Attentiveness, Engagement and Social Presence. All selected scales meet the $\geq 77\%$ threshold of the GMU HRI psychometric guidelines [30].

VI. RESULTS

By synthesizing subjective psychometric scales, objective system logs, and qualitative feedback, we address the study’s two

¹<https://github.com/MaAI-Kyoto/MaAI>

²<https://huggingface.co/distilbert/distilbert-base-uncased>

primary research questions regarding the utility of conversational backchannels across the four constructs. While our small sample size ($N = 20$) precluded formal statistical significance across several metrics, consistent descriptive trends emerged.

A. Do Backchannels Enhance the Perceived Naturalness?

Descriptive statistics consistently indicate that active backchannel models enhance the interaction compared to the baseline model. Per-block subjective ratings on the perceived naturalness initially contradicted our hypothesis, with active backchanneling scoring lower than the baseline. Qualitative interviews revealed an expectation violation, where participants simply did not expect a robot to produce backchannels. Furthermore, inherent system latency depressed the naturalness ceiling across all conditions. However, once participants acclimated to the backchannels, retrospective rankings showed a clear preference for the backchannel models over the baseline. Perceived attentiveness scores were uniform across all conditions. This lack of variance is likely due to a hardware-driven ceiling effect, where Pepper’s baseline physical embodiment (continuous eye contact, head tracking) was already sufficient to signal active listening. While subjective Likert scales for engagement showed flat results, objective behavioural metrics revealed a significant divergence. As established by Inoue et al., user turn duration is a reliable proxy for engagement [16]. System logs demonstrated that the backchannel models (specifically ML-based) extended user speech by approximately 2 seconds per turn, proving that responsive feedback successfully increases user engagement. Although, the Human-Likeness subscale yielded expectedly low scores due to Pepper’s intentionally non-anthropomorphic design. However, the baseline was ranked as the “least engaging”, highlighting that the absence of backchannels reduces the image of the robot as a social entity.

B. Which Timing Algorithm is Best?

When looking at the effect between both backchannel models, the Machine Learning (ML) architecture consistently scored better over the rule-based heuristic. The most important differentiator between the two models was objective engagement. The ML-based model’s ability to extend participant speech by nearly 2 seconds per turn compared to the rule-based heuristic indicates successful interactional engagement. The precise, nuanced timing of the neural network prompted users to elaborate on their thoughts. While per-block naturalness and attentiveness differences were subtle, end-of-session comparative rankings favoured the ML model. As a proxy for Social Presence, participants were asked which version of the robot they would prefer to converse with again. The ML-based architecture received the highest number of positive selections.

VII. CONCLUSION

Although limited statistical power prevents definitive inferential claims, the convergence of objective logs and retrospective subjective rankings clearly points toward the ML-based pipeline as the better architecture for generating natural, engaging, and socially present human-robot dialogue.

REFERENCES

- [1] Łukasik A. and A. Gut. “From robots to chatbots: unveiling the dynamics of human-AI interaction.” In: (2025). URL: <https://doi.org/10.3389/fpsyg.2025.1569277>.
- [2] Amalia Istiqlali Adiba, Takeshi Homma, and Toshi-nori Miyoshi. “Towards Immediate Backchannel Generation Using Attention-Based Early Prediction Model”. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2021, pp. 7408–7412. DOI: 10.1109/ICASSP39728.2021.9414193. URL: <https://ieeexplore.ieee.org/document/9414193> (visited on 11/13/2025).
- [3] Kyle Arnold. “Humming Along”. In: *Contemporary Psychoanalysis* 48.1 (2012), pp. 100–117. DOI: 10.1080/00107530.2012.10746491. eprint: <https://doi.org/10.1080/00107530.2012.10746491>. URL: <https://doi.org/10.1080/00107530.2012.10746491>.
- [4] Stefan Benus, Agustín Gravano, and Julia Hirschberg. “THE PROSODY OF BACKCHANNELS IN AMERICAN ENGLISH”. en. In: (2007).
- [5] Philippe Blache et al. “An integrated model for predicting backchannel feedbacks”. en. In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. Virtual Event Scotland UK: ACM, Oct. 2020, pp. 1–3. ISBN: 978-1-4503-7586-3. DOI: 10.1145/3383652.3423948. URL: <https://dl.acm.org/doi/10.1145/3383652.3423948> (visited on 04/21/2026).
- [6] Peter Blomsma et al. “Backchannel behavior is idiosyncratic”. en. In: *Language and Cognition* 16.4 (Dec. 2024), pp. 1158–1181. ISSN: 1866-9808, 1866-9859. DOI: 10.1017/langcog.2024.1. URL: <https://www.cambridge.org/core/journals/language-and-cognition/article/backchannel-behavior-is-idiosyncratic/F75D0AEEAF258399166A58E3DDCC7D7E> (visited on 01/19/2026).
- [7] Shiye Cao et al. *Interruption Handling for Conversational Robots*. arXiv:2501.01568 [cs]. Apr. 2025. DOI: 10.48550/arXiv.2501.01568. URL: <http://arxiv.org/abs/2501.01568> (visited on 11/04/2025).
- [8] Yong-Seok Choi, Jeong-Uk Bang, and Seung Hi Kim. “Joint streaming model for backchannel prediction and automatic speech recognition”. en. In: *ETRI Journal* 46.1 (2024). eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.4218/etrij.2023-0358>, pp. 118–126. ISSN: 2233-7326. DOI: 10.4218/etrij.2023-0358. URL: <https://onlinelibrary.wiley.com/doi/abs/10.4218/etrij.2023-0358> (visited on 01/19/2026).
- [9] Ronald Cumbal et al. “Let Me Finish First - The Effect of Interruption-Handling Strategy on the Perceived Personality of a Social Agent”. In: *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. IVA ’24. New York, NY, USA: Association for Computing Machinery, Dec. 2024, pp. 1–10. ISBN: 979-8-4007-0625-7. DOI: 10.1145/3652988.3673916. URL: <https://dl.acm.org/doi/10.1145/3652988.3673916> (visited on 11/04/2025).

- [10] Erik Ekstedt and Gabriel Skantze. “TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. arXiv:2010.10874 [cs]. 2020, pp. 2981–2990. DOI: 10.18653/v1/2020.findings-emnlp.268. URL: <http://arxiv.org/abs/2010.10874> (visited on 11/04/2025).
- [11] Erik Ekstedt and Gabriel Skantze. *Voice Activity Projection: Self-supervised Learning of Turn-taking Events*. 2022. arXiv: 2205.09812 [eess.AS]. URL: <https://arxiv.org/abs/2205.09812>.
- [12] Olov Engwall, Ronald Cumbal, and Ali Reza Majlesi. “Socio-cultural perception of robot backchannels”. In: *Frontiers in Robotics and AI* 10 (Jan. 2023), p. 988042. ISSN: 2296-9144. DOI: 10.3389/frobt.2023.988042. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9909394/> (visited on 01/19/2026).
- [13] Kushal Garg, Lincy Meera Mathews, and Prashanth Kambl. “InterConv: Conversational Agent That Handles Interruptions Like Humans”. In: *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*. Vol. 3. Mar. 2025, pp. 1–6. DOI: 10.1109/IATMSI64286.2025.10985683. URL: <https://ieeexplore.ieee.org/abstract/document/10985683> (visited on 11/04/2025).
- [14] M. Grimm and K. Kroschel. *Robust Speech: Recognition and Understanding*. IntechOpen, 2007. ISBN: 978-3-902613-08-0. URL: <https://books.google.be/books?id=OuSgDwAAQBAJ>.
- [15] Koji Inoue et al. *Prompt-Guided Turn-Taking Prediction*. 2025. arXiv: 2506.21191 [cs.CL]. URL: <https://arxiv.org/abs/2506.21191>.
- [16] Koji Inoue et al. “Towards Objective Evaluation of Socially-Situated Conversational Robots: Assessing Human-Likeness through Multimodal User Behaviors”. In: *International Conference on Multimodal Interaction*. ICMI ’23. ACM, Oct. 2023, pp. 86–90. DOI: 10.1145/3610661.3617151. URL: <http://dx.doi.org/10.1145/3610661.3617151>.
- [17] Koji Inoue et al. *Yeah, Un, Oh: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection*. arXiv:2410.15929 [cs]. Feb. 2025. DOI: 10.48550/arXiv.2410.15929. URL: <http://arxiv.org/abs/2410.15929> (visited on 11/04/2025).
- [18] Divesh Lala et al. “Attentive listening system with backchanneling, response generation and flexible turn-taking”. In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Ed. by Kristiina Jokinen et al. Saarbrücken, Germany: Association for Computational Linguistics, Aug. 2017, pp. 127–136. DOI: 10.18653/v1/W17-5516. URL: <https://aclanthology.org/W17-5516/> (visited on 01/20/2026).
- [19] Borui Liao et al. *FlexDuo: A Pluggable System for Enabling Full-Duplex Capabilities in Speech Dialogue Systems*. arXiv:2502.13472 [cs] version: 2. May 2025. DOI: 10.48550/arXiv.2502.13472. URL: <http://arxiv.org/abs/2502.13472> (visited on 10/21/2025).
- [20] Yuxin Lin et al. *Predicting Turn-Taking and Backchannel in Human-Machine Conversations Using Linguistic, Acoustic, and Visual Signals*. en. arXiv:2505.12654 [cs]. May 2025. DOI: 10.48550/arXiv.2505.12654. URL: <http://arxiv.org/abs/2505.12654> (visited on 11/02/2025).
- [21] Openai. *Voice activity detection (VAD) - OpenAI API*. nl-NL. URL: <https://platform.openai.com> (visited on 11/04/2025).
- [22] Michael Paierl, Martin Hagmüller, and Barbara Schuppler. “Continuous prediction of backchannel timing for human-robot interaction”. en. In: *Interspeech 2025*. ISCA, Aug. 2025, pp. 3020–3024. DOI: 10.21437/Interspeech.2025-770. URL: https://www.isca-archive.org/interspeech_2025/paierl25_interspeech.html (visited on 11/04/2025).
- [23] Yo-Han Park et al. “Backchannel prediction, based on who, when and what”. en. In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 3570–3574. DOI: 10.21437/Interspeech.2024-2523. URL: https://www.isca-archive.org/interspeech_2024/park24b_interspeech.html (visited on 01/19/2026).
- [24] Yoonvin Park, Doha Kim, and Hayeon Song. “I Know You’re Listening: Designing Visual Backchannels for Voice User Interfaces”. In: *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*. IUI ’25 Companion. New York, NY, USA: Association for Computing Machinery, Mar. 2025, pp. 56–59. ISBN: 979-8-4007-1409-2. DOI: 10.1145/3708557.3716343. URL: <https://dl.acm.org/doi/10.1145/3708557.3716343> (visited on 11/04/2025).
- [25] Maria J. Pinto and Tony Belpaeme. “Predictive Turn-Taking: Leveraging Language Models to Anticipate Turn Transitions in Human-Robot Dialogue”. In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. Aug. 2024, pp. 1733–1738. DOI: 10.1109/RO-MAN60168.2024.10731379. URL: <https://ieeexplore.ieee.org/document/10731379/> (visited on 10/21/2025).
- [26] Ronald Poppe et al. “Backchannel Strategies for Artificial Listeners”. en. In: *Intelligent Virtual Agents*. Ed. by Jan Allbeck et al. Vol. 6356. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 146–158. ISBN: 978-3-642-15891-9 978-3-642-15892-6. DOI: 10.1007/978-3-642-15892-6_16. URL: http://link.springer.com/10.1007/978-3-642-15892-6_16 (visited on 01/19/2026).
- [27] Nikhil R. *Mastering Turn Detection and Interruption Handling in Voice AI Applications*. en. Oct. 2025. URL: <https://comparevoiceai.com> (visited on 10/21/2025).
- [28] Andrew Reece et al. “The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation”. en. In: *Science Advances* 9.13 (Mar. 2023), eadf3197. ISSN: 2375-2548. DOI: 10.1126/sciadv.adf3197. URL: <https://www.science.org/doi/10.1126/sciadv.adf3197> (visited on 02/13/2026).
- [29] Sam O’Connor Russell and Naomi Harte. *Visual Cues Enhance Predictive Turn-Taking for Two-Party Human Interaction*. Version Number: 2. 2025. DOI: 10.48550/

- ARXIV.2505.21043. URL: <https://arxiv.org/abs/2505.21043> (visited on 11/01/2025).
- [30] Laura Saad et al. “Choosing the “Perfect” Scale: A Primer to Evaluate Existing Scales in HRI”. In: *J. Hum.-Robot Interact.* 15.2 (Jan. 2026), 42:1–42:30. DOI: 10.1145/3772066. URL: <https://dl.acm.org/doi/10.1145/3772066> (visited on 05/14/2026).
- [31] Mohan Shi et al. *Semantic VAD: Low-Latency Voice Activity Detection for Speech Interaction*. arXiv:2305.12450 [eess]. May 2023. DOI: 10.48550/arXiv.2305.12450. URL: <http://arxiv.org/abs/2305.12450> (visited on 10/21/2025).
- [32] Thi Hoang Mai Vu and Thi Bich Ngan Tran. “Backchannel Across Cultures: A Review and Implications”. en. In: *VietTESOL International Convention Proceedings 1* (2021). URL: <https://proceedings.viettesol.org.vn/index.php/vic/article/view/189> (visited on 01/19/2026).
- [33] Jinhan Wang et al. *Turn-taking and Backchannel Prediction with Acoustic and Large Language Model Fusion*. arXiv:2401.14717 [cs]. Jan. 2024. DOI: 10.48550/arXiv.2401.14717. URL: <http://arxiv.org/abs/2401.14717> (visited on 11/15/2025).
- [34] Nigel Ward. *Backchannel Facts*. May 2017. URL: <https://www.cs.utep.edu/nigel/bc/> (visited on 01/19/2026).
- [35] Nigel Ward and Wataru Tsukahara. “Prosodic features which cue back-channel responses in English and Japanese”. In: *Journal of Pragmatics* 32.8 (2000), pp. 1177–1207. ISSN: 0378-2166. DOI: [https://doi.org/10.1016/S0378-2166\(99\)00109-5](https://doi.org/10.1016/S0378-2166(99)00109-5). URL: <https://www.sciencedirect.com/science/article/pii/S0378216699001095>.
- [36] Victor H. Yngve. “ON GETTING A WORD IN EDGEWISE”. In: *The Sixth Regional Meeting [of the] Chicago Linguistic Society*. 6th ed. 1970, pp. 567–578.

Table of Contents

Acknowledgements	iii
Admission to loan	iv
Thesis as part of an exam	v
Abstract	vi
Extended Abstract	vii
Table of Contents	xii
List of Figures	xiv
List of Tables	xiv
List of Acronyms	xv
1 Introduction	1
2 Literature Review	3
2.1 The Functional Role of Backchannels in Conversations	3
2.1.1 What makes a good conversation?	3
2.1.2 Why would a robot need to use backchannels	4
2.2 Generation of Backchannels	5
2.2.1 Determining the timing of a backchannel	5
2.2.2 Different types of backchannels	6
2.2.3 Differences in Context	6
3 System Design	7
3.1 Conversational Pipeline	7
3.1.1 Design Choices	7
3.1.2 System Implementation	8
3.1.3 Integration with Pepper	9
3.2 Backchannel module	11
3.2.1 Backchannel Timing	11
3.2.2 Backchannel Type Prediction	13
3.2.3 Reproducing the backchannel	15
4 Evaluation	16
4.1 Participants	16
4.2 Setting	17
4.3 Protocol	17
4.4 Methodology	18
4.4.1 Four Constructs	18
4.4.2 Validated Scales	19
5 Results	22
5.1 Subjective Measures	22
5.1.1 CCR Coordination	22
5.1.2 CCR Connection	22
5.1.3 Standalone Attentiveness Item	23

5.1.4	MOS-style Naturalness Item	24
5.1.5	End-of-session: PMQ	24
5.1.6	End-of-session: comparative ranking	25
5.1.7	Manipulation Check	26
5.2	Objective Measures	26
5.2.1	Turn Duration	26
5.2.2	Inter-turn latency	27
6	Discussion	28
6.1	Do Backchannels Enhance the Perceived Naturalness?	28
6.1.1	Naturalness	28
6.1.2	Attentiveness	29
6.1.3	Engagement	29
6.1.4	Social Presence	30
6.1.5	Conclusion	30
6.2	Which Timing Algorithm is Best?	30
6.2.1	Naturalness	30
6.2.2	Engagement	31
6.2.3	Attentiveness	31
6.2.4	Social Presence	31
6.2.5	Conclusion	31
7	Conclusion	32
7.1	Limitations	32
7.2	Future works	32
7.3	Conclusion	33
8	Societal Reflections	34
A	Item-by-Item Mappings	39
B	Use of Generative Artificial Intelligence	41

List of Figures

3.1	Full pipeline for backchannel implementation	8
3.3	Ten Framework - Ten Agent frontend	9
3.2	Ten Framework - Graph overview of complete setup	10
3.4	Rule-based approach: State-machine	12
4.1	Experimental setup: showing (a) a front view of the robot and (b) a participant interacting with it.	17
5.1	CCR-subscale score per Version	23
5.2	Attentiveness and Engagement Item from the CCR scale	23
5.3	Mean Attentiveness per Version	24
5.4	Mean Naturalness per Version	24
5.5	End-of-Session PMQ mean scores	25
5.6	Frequencies of Questions from Comparative Ranking	26
5.7	Mean Turn Durations for Robot and Human. Darker colour indicates robot turn, lighter colour indicates speaker turn.	27
5.8	Mean Turn Taking Duration between Robot and Human. Darker colour indicates robot turn, lighter colour indicates speaker turn.	27

List of Tables

3.1	Ten most common backchannels in CANDOR corpus	14
4.1	Structure of the study	18

List of Acronyms

VAD	Voice Activity Detection
VAP	Voice Activity Projection
LLM	Large Language Model
HRI	Human-Robot Interaction
BOP	Backchannel Opportunity Point
BCP	Backchannel Category Prediction
ML	Machine Learning
ASR	Automatic Speech Recognition
STT	Speech-to-Text
TTS	Text-to-Speech
S2S	Speech-to-Speech

Introduction

Current voice agents in HRI frequently struggle with a lack of perceived naturalness. Personalization remains superficial, turn-taking lacks the fluid synchronization inherent in human conversations, and systems often lack capabilities to get interrupted, but also the capability to reproduce interruptions by itself. These issues are primarily rooted in the “half-duplex” architecture of traditional dialogue systems, which operate on a rigid, alternative basis. Such models require a distinct period of silence from the user before initiating a response, introducing a latency that fundamentally disrupts natural conversational flow. However, the integration of low-latency predictive models facilitates a transition toward more sophisticated, full-duplex interactions, where simultaneous listening and speaking are supported by both speaking parties [15, 25, 37]. By leveraging these technologies, robotic agents can evolve from reactive, passive listeners into proactive conversational partners, capable of anticipating turn transitions and providing real-time feedback that mirrors human social interactions.

While the computational prediction of turn-taking has been extensively addressed through various predictive methodologies [15, 25, 37, 46], achieving human-like fluidity in HRI remains an unresolved challenge. Beyond the timing of turns, the content of interactions often lacks coherence, personalization is frequently hindered by constrained memory and the emotional aspect of the conversation often get lost (due to the conversion from speech to text). Furthermore, the objective evaluation of conversational quality remains a significant methodological hurdle [30].

Even as the emergence of LLMs promises to refine content generation, the structural execution of the interaction often remains flawed. Historically, robotic systems have struggled with interruption detection, where the hardware or software effectively ignores user input while the robot is speaking, thereby preventing natural interruptions. Moreover, contemporary agents largely fail to produce backchannels, the brief, often non-lexical vocalizations issued by a listener to signal attentiveness [4, 52]. These signals are fundamental in the communicative process, as they provide the speaker with constant validation of the listener’s engagement and comprehension.

A possible application for these social robots, could be the deployment of these robots in care homes. Social isolation and loneliness among elderly people increases globally, approximately one out of ten elderly people experiences loneliness [1]. By building social relationships between humans and robots, these robots can stimulate a sense of connection and reduce feelings of loneliness among elderly people [38].

This research aims to address especially the reproduction of backchannels in interactional dynamics by a robot. While recent studies have begun exploring the real-time generation of backchannels [23, 34], these models often focus exclusively on the temporal prediction of backchannel opportunities, frequently neglecting the specific type of backchannel, such as whether a backchannel should be context-dependent.

The present study proposes the development of a fully functional voice-agent architecture capable of concurrent timing and type prediction of backchannels within a real-time environment. By optimizing this dual-prediction pipeline, the research seeks to minimize generation latency, thereby facilitating more fluid HRI. To evaluate the efficacy of this approach, the following research questions are addressed:

- **RQ1:** To what extent does the autonomous generation of backchannels by a social robot enhance perceived naturalness and user engagement?
- **RQ2:** Which algorithmic approach for timing prediction is perceived as most natural by human interlocutors?

This dissertation outlines the development and validation of an autonomous backchannel generation

module designed to predict both backchannel opportunities and specific backchannel typologies. The subsequent chapters are organized as follows:

The first chapter provides a critical analysis of the current state of the art in conversational AI voice agents, identifying the technical and social benchmarks of modern dialogue systems. Building upon this theoretical foundation, the second section details the technical architecture and implementation of the proposed system. The research then transitions to a comprehensive evaluation with $n = 20$ participants who evaluated the system. A dual approach was taken to both distinguish between a baseline model and the backchannel generation models as well as between two backchannel temporal prediction models. We conclude this book with the findings from the experimental evaluation period, where we try to ground our two research questions against four major constructs: Naturalness, Engagement, Attentiveness and Social Presence.

Literature Review

Authentic human interactions are multifaceted, requiring the integration of numerous behaviours to be perceived as socially competent. For robotic systems, managing these behaviours simultaneously represents a significant challenge, often cited as an application of Moravec’s paradox [32]. The paradox observes that high-level intelligent tasks (like playing grandmaster-level chess or solving complex stock market equations) require very little computational power, whereas low-level sensorimotor skills (like walking through a crowded room or perceiving the subtle social cues of a conversation) require enormous resources. Within a dialogue, turn-taking involves more than a simple chronological sequence of speakers, it functions as a fluid exchange where transitions frequently occur before the previous speaker has concluded [47]. Furthermore, the role of the listener is critical. Non-verbal signals and backchanneling provide essential feedback, such as confirmation or dissent, without requiring the listener to claim the floor [52]. Despite the ubiquity of backchannels in human-to-human dialogue, mainstream voice assistants frequently lack these mechanisms in real-time interactions [28]. This absence is often attributed to the half-duplex nature of traditional speech processing pipelines, which require a user to finish their utterance before the system begins processing a response. Consequently, a substantial body of research is currently dedicated to determining the optimal timing and specific typology of backchannels required for various conversational contexts. This includes efforts to categorize “when” a listener should intervene and “what” form the response should take to maintain the flow of interaction.

The following chapter covers the fundamental mechanics of human conversation while examining recent advancements in HRI, specifically regarding the evolution of voice-agent architectures and autonomous backchannel generation. By situating these developments within a broader communicative framework, the section provides a critical overview of current methodologies and identifies the technical benchmarks required to achieve naturalistic human-robot interaction.

2.1 The Functional Role of Backchannels in Conversations

Backchannels are ubiquitous in human conversations and play an important role in active listening [23, 48], showing interest to the speaker without taking the floor. First, the general role of backchannels in a conversation is explored. It is necessary to cover the fundamental components and structural aspects that shape a conversation. By situating backchannels within this broader communicative framework, the subsequent sections of this chapter will define their specific characteristics, categorize their various forms, and examine the computational methods required for reproducing backchannels in robotic systems.

2.1.1 What makes a good conversation?

A natural conversational exchange is not simply a sequence of transmitted information but relies on a complex repertoire of communicative behaviours, including verbal, paralinguistic, and non-verbal signals [41]. These behaviours operate in a highly synchronized manner to manage the flow of information and maintain social presence. In human conversations, this complexity is managed through a variety of structural mechanisms, ranging from turn-taking and topic management to backchanneling and gaze synchronization. Understanding these dynamics is essential for translating the nuances of human interaction into robotic architectures capable of fluid, socially-aware conversations.

The first important aspect of this process is turn-taking: systematic coordination of speaker transitions that ensures a fluid dialogue [14]. Early computational approaches primarily relied on acoustic-prosodic features such as pitch or intensity variations, and pause durations to identify relevant transition places

[20]. While these methods established a foundation for automated interaction, their performance often degraded in noisy environments and introduced additional latencies, which led to long waiting times in between turns.

With the advancement of more resilient, machine learning-based turn detection systems, research shifted toward semantic turn-taking detection. These frameworks utilize linguistic information, such as syntactic completeness or punctuation cues, to determine whether a speaker has concluded their turn [33, 46]. More recently, the field has moved toward context-aware architectures that leverage transformer-based models to predict turn-taking dynamics with higher precision [15, 25, 37].

Recent studies have further expanded this domain by integrating multimodal inputs. Incorporating visual data such as gaze direction and head pose together with the traditional prosodic and lexical features, improving prediction accuracy [27, 44].

Conversations are further characterized by the occurrences of overlapping speech, where participants intermittently interject while another speaker holds the floor. These interruptions are traditionally divided into cooperative and competitive categories [9]. Cooperative overlaps, such as backchannels, serve as brief non-intrusive utterances that provide validation or indicate continued engagement without attempting to take the turn. In contrast, competitive interruptions represent changing the conversational floor, necessitating a distinct communicative response or disagreement with the conversational partner.

A significant challenge in HRI lies in the strategic management of these interjections. Specifically, a system must determine whether to disregard the input, provide a direct response, or offer a brief acknowledgment before resuming the primary conversation [12]. The correct management of these interruptions based on the context introduces an additional challenge. Current architectural approaches often address this through finite-state machines that approximate conversational context [9] or a compact LLM to perform intent recognition on interruptions, allowing the agent to dynamically adapt its behaviour based on the perceived function of the user’s input [19].

Backchannels represent a specialized category of conversational interruptions that have a distinct functional meaning. Unlike competitive interruptions, backchannels are non-obligatory vocal or gestural signals produced by a listener while another participant maintains the floor [52]. These signals serve as indicators of active listenership, providing the speaker with continuous feedback regarding the listener’s attentiveness without taking the turn.

Backchannels manifest in several distinct morphological forms. Non-lexical Vocalizations consist of utterances without any semantic content, such as “mhm”, “uh-huh”, or “hmm”, which function purely as indicators of active listening. Lexical Backchannels involve brief words that may be context-independent (e.g., “yes”, “right”) or context-dependent (e.g., “oh wow”, “exactly”) [8]. Non-verbal Backchannels are embodied signals, e.g. head nodding, which fulfil the same feedback function through visual rather than auditory channels [36].

The following sections of this chapter focus on a systematic exploration of backchanneling as an important aspect of conversational fluidity. First, the functional significance of these signals in human conversation is examined to establish why they are indispensable for perceived naturalness. Subsequently, the technical challenges of backchannel generation are addressed, specifically focusing on the dual requirements of temporal precision (identifying the optimal timing) and typological appropriateness (selecting the correct type of backchannel). Finally, the chapter analyzes how these variables are intrinsically linked to the broader conversational context, including cultural, situational, and interlocutor-specific factors and how these dependencies shape the development of context-aware robotic agents.

2.1.2 Why would a robot need to use backchannels

Backchannels function as critical interactional elements employed by a listener to facilitate communicative efficiency and maintain conversational flow [35]. These subtle cues, ranging from rhythmic head nods to non-lexical vocalizations and brief verbal interjections, serve as indicators of active engagement. In the context of HRI, these signals demonstrate that the robotic agent is maintaining attention and agrees or disagrees with the speaker’s statements. By clarifying the listener’s role, backchannels enhance both the enjoyment and the efficiency of the exchange [41]. Conversely, the absence or temporal misalignment of these cues can negatively impact the perceived social presence of the agent and disrupt the speaker’s fluency [17].

Furthermore, the maintenance of active listenership has been shown to increase user engagement

and encourage longer conversational contributions, as the speaker receives constant validation of the robot’s comprehension [24]. Critically, the autonomous integration of backchannels reduces the reliance on artificial status indicators, such as LED signals or UI prompts. By providing naturalistic social feedback, the robot signals its state of active listenership through behaviour rather than hardware, thereby moving closer to a seamless, human-like interaction behaviour.

2.2 Generation of Backchannels

Given the functional importance of backchannels, their autonomous generation requires a robust computational framework capable of resolving two primary challenges. First, the system must predict the correct timing by identifying the optimal moments for a response during active listenership, a task typically categorized as Backchannel Opportunity Point (BOP) prediction. Second, the agent must determine the specific lexical or non-lexical form that fits the conversation, also called Backchannel Category Prediction (BCP). Finally, the research examines how these timing and typology predictions are influenced by the conversational context. By integrating situational and linguistic variables, the system can adapt its feedback to the unique dynamics of the exchange, ensuring that both the timing and the typology of the backchannel remain coherent.

2.2.1 Determining the timing of a backchannel

A foundational component of the backchannel generation process is the precise identification of correct timings for backchannels, a task formally defined as BOP prediction. Various methodologies have been developed to predict these optimal moments, utilizing a combination of prosodic, acoustic and lexical features. Among the most fundamental approaches are rule-based models, which leverage predetermined heuristic thresholds to identify potential opportunities. These models typically synthesize acoustic parameters, specifically pitch contours and pause durations, sometimes combined with visual cues such as gaze direction to determine the moment of a BOP [39, 51]. A primary advantage of the deterministic, rule-based approach is its minimal computational overhead, which results in significantly lower processing latency compared to recent Machine Learning (ML) architectures. This efficiency makes such models particularly suitable for real-time applications where immediate responsiveness is critical. Conversely, a significant limitation lies in their lack of inherent adaptability. These models often require extensive manual fine-tuning for specific linguistic contexts and individual user profiles, as conversational expectations vary across different languages and idiosyncratic speaking styles [8, 35].

An alternative to deterministic models involves ML architectures that process the acoustic signals of the primary speaker to identify optimal backchannel opportunities. These methods analyse sequential audio frames to capture complex patterns that rule-based systems may overlook. For instance, Choi et al. implemented a multi-task learning framework that integrates backchannel prediction with Automatic Speech Recognition (ASR), utilizing a dual-pipeline architecture to leverage shared representations between tasks [11]. Other researchers have adapted large-scale pre-trained models. Inoue et al. utilized the Voice Activity Projection (VAP) framework, a Transformer-based model pre-trained on extensive conversational datasets, and fine-tuned it specifically for the BOP task [16, 23]. Conversely, some ML approaches focus exclusively on prosodic features rather than raw audio or high-dimensional embeddings. Paierl et al. trained a regression-based decision tree model on a discrete set of prosodic variables to determine the precise temporal placement of backchannels [34].

Beyond acoustic and prosodic signals, recent research has increasingly focused on integrating textual and visual modalities to enhance the accuracy of BOP prediction. Wang et al. implemented a late fusion architecture that synthesizes acoustic embeddings with linguistic data derived from textual representations. By merging these distinct feature spaces before the final classification stage, the model leverages both the paralinguistic cues of the audio and the semantic context of the conversation to identify BOPs [49]. However, a significant limitation of this approach is the inherent architectural latency introduced by the ASR conversion process. To mitigate this delay, Adiba et al. developed a dual-pipeline that decouples the BOP classifier from the ASR system. By separating these components, the architecture reduces the temporal overhead associated with sequential processing, thereby facilitating the near-instantaneous responses required for naturalistic backchanneling. Furthermore, visual signals, such as gaze, facial expressions, and head movements, provide critical non-verbal markers that can be incorporated into the final prediction classifier. The video data is typically embedded into unimodal feature representations. These

embeddings are subsequently integrated into a classifier, allowing the system to capture the multimodal aspect of the BOP prediction task [27].

2.2.2 Different types of backchannels

A critical secondary dimension of backchannel synthesis is the selection of the appropriate lexical or non-lexical form, or BCP. In linguistic literature, backchannels are frequently divided into two functional categories: continuers and assessments [7, 23]. Continuers are typically non-lexical vocalizations whose primary objective is to signal ongoing attention and encourage the speaker to maintain the floor. Conversely, assessments are context-dependent lexical interjections that convey a specific emotional reaction to the propositional content of the conversation.

Determining the appropriate functional class is computationally demanding, as it requires a nuanced understanding of the conversational context. To address this, Park et al. integrated contextual variables into the prediction classifier by utilizing embeddings that represent the agent’s “persona”, including historical backchannel frequency and average turn duration, alongside thematic topics and conversational progression [35]. Recognizing that the latency inherent in ASR systems can delay the availability of text for such modelling, Adiba et al. developed a dual-pipeline architecture. This framework ensures that conversational context is processed in parallel, allowing for informed typology prediction without compromising the real-time responsiveness of the system [3]. Earlier backchannel systems employed deterministic classifiers that utilized specific lexical markers to categorize the appropriate response type [7]. These models rely on the identification of predetermined linguistic triggers, such as specific keywords or syntactic completion points to select a corresponding backchannel from a fixed library. While these rule-based systems provide a computationally efficient method for ensuring semantic consistency, they are often limited by their inability to account for the broader context of the conversation.

2.2.3 Differences in Context

A third critical dimension in backchannel synthesis is the broader conversational context. Both the temporal placement and the selection of these signals are deeply dependent upon the sociocultural and situational environment [35]. Backchannel behaviour varies in the frequency of occurrence and the distribution of specific typologies. Cross-cultural linguistics reveals substantial variance in these patterns. For instance, Japanese users exhibit nearly triple the density of backchannels during a conversation compared to English or Dutch users [50]. Beyond cultural factors, the demographic profile of the interlocutor further modulates interactional dynamics. Research indicates that backchanneling behaviour is significantly adapted based on the age of the participant, such as in interactions with children versus adults, to better align with their specific communicative and cognitive needs [8]. Furthermore, gender-based differences have been shown to influence the frequency and prosodic characteristics of backchannels, suggesting that a truly social robot must account for a diverse array of speaker identity markers to maintain interactional rapport [6, 48].

Furthermore, the placement of the BOP within the syntactic structure directly influences the appropriate functional typology. There is a marked distinction in the type of backchannel selected depending on whether the opportunity occurs at a definitive sentence ending or during a syntactically incomplete segment [8]. Finally, paralinguistic elements, such as the intensity of the exchange, further dictate the appropriate choice and rate of response [6]. Meaning that during interaction-focused conversations, the rate of backchannels naturally increases compared to more informative conversations, where only one person is taking the floor [48].

System Design

To conduct the study, a complete autonomous AI voice-agent architecture was developed and integrated with the Pepper humanoid robotic platform¹. The system design prioritizes conversational fluency by orchestrating multiple concurrent processes, spanning from real-time ASR to natural response generation. A core component of this architecture is the specialized backchannel generation module, which functions as a parallel processing layer. This module is responsible for the continuous monitoring of cues to predict both the optimal temporal placement (BOP) and the functional category (BCP) of backchannels. By executing these predictions within a low-latency pipeline, the system enables the robot to provide responsive, real-time feedback that maintains the rhythm of natural dialogue. In the following section, both the conversational pipeline as well as the backchannel module are being covered.

3.1 Conversational Pipeline

The implementation of a conversational pipeline requires a strategic selection between several architectures, each offering distinct advantages depending on the specific requirements of the interaction. To establish a robust foundation for this study, the following sections cover the available system configurations and provide a technical justification for the selected framework. This is followed by an explanation of the autonomous backchannel generation module and the integration of the system with the Pepper robotic platform.

3.1.1 Design Choices

The architectural design of a conversational voice agent is decided by the desired capabilities and the required degree of system modularity. Generally, voice agent pipelines can be categorized into three primary configurations, each offering varying levels of flexibility.

The cascaded architecture functions by integrating three distinct modules: Speech-to-Text (STT), an LLM and Text-to-Speech (TTS) [10, 55]. This setup offers the highest degree of modularity, as each component can be independently selected, optimized or replaced. Due to this inherent flexibility, in this research a cascaded configuration is utilized as its foundational framework.

Alternatively, the semi-cascaded configuration integrates STT and LLM processing within a single model. This approach preserves the paralinguistic and emotional nuances contained within the original audio signal information that is frequently lost during the transcription phase of a fully cascaded system [10].

Finally, the Speech-to-Speech (S2S) configuration consolidates all three stages into a singular end-to-end model. While this paradigm yields the most naturalistic results by processing audio to audio directly, it presents significant challenges regarding system flexibility and computational overhead due to the massive scale of the underlying models [29, 53, 54]. Furthermore, S2S models face substantial memory constraints, as the absence of an intermediate textual representation complicates the long-term storage and retrieval of conversational history. Although S2S systems are very difficult to create due to the large amount of needed conversational data, they naturally introduce the ability to output backchannels, due to their presence in the training set (which is just a set of conversations). While S2S architectures present significant implementation challenges, primarily due to the massive volumes of conversational data required for training, they offer a unique advantage in interactional realism. Because these models are trained on

¹<https://us.softbankrobotics.com/pepper>

raw, bidirectional human conversations, they inherently learn to model the backchannel behaviour of a real human. In this paradigm, backchannels are not treated as external modules but emerge naturally as part of the learned distribution of conversational speech. By training on datasets where backchannels are already embedded within the natural flow of dialogue, S2S systems can potentially replicate the timing and prosody of listener interjections without the need for separate heuristic or predictive pipelines [43].

A secondary architectural consideration is the communicative modality of the system, specifically the distinction between a half-duplex and full-duplex conversation. In a half-duplex framework, the system is incapable of simultaneous transmission and reception. Consequently, the agent cannot process incoming audio while generating speech, which effectively precludes user interruptions. While this paradigm is increasingly considered insufficient for naturalistic HRI, it used to be the legacy standard for many traditional voice interfaces.

In contrast, full-duplex communication enables continuous, bidirectional data flow, mirroring the simultaneous nature of human dialogue. To bridge this gap, modern research demonstrates that half-duplex architectures can be transitioned toward full-duplex functionality through the integration of pluggable, predictive modules. Systems such as TurnGPT and FlexDuo facilitate this by providing real-time coordination of turn-taking and interruption logic, allowing for a more responsive and asynchronous conversational flow [15, 25].

By synthesizing the aforementioned architectural paradigms, the comprehensive system design was established as illustrated in Figure 3.1. The framework utilizes a modular cascaded configuration, integrating STT, an LLM, and TTS components. To facilitate a full-duplex conversational modality, a Voice Activity Detection (VAD) module is integrated into the primary pipeline, allowing the system to continuously monitor acoustic input even during active speech synthesis.

The backchannel generation module is architecturally decoupled from the primary conversational pipeline, operating as an autonomous parallel processing layer rather than a sequential bottleneck. By asynchronously retrieving real-time audio features and textual data from the ongoing conversation, this module identifies optimal BOPs and determines the appropriate response category. This decoupled design ensures that listener feedback can be generated with minimal latency.

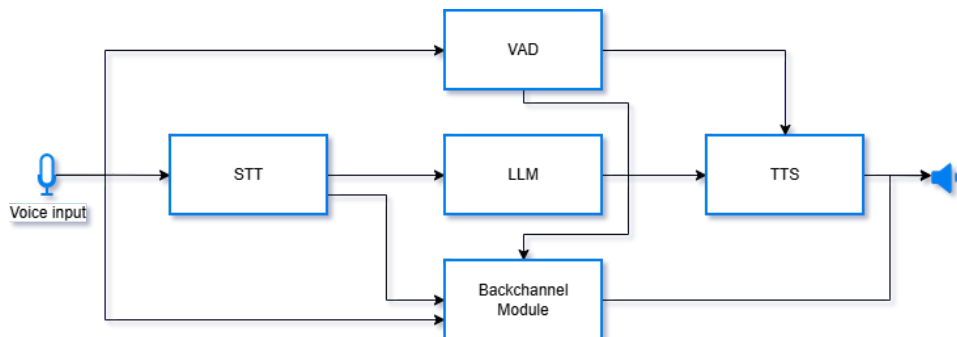


Figure 3.1: Full pipeline for backchannel implementation

3.1.2 System Implementation

Building upon the architectural requirements established in the previous section, the conversational pipeline was implemented utilizing the Ten Framework². This development environment facilitates the seamless, low-latency transfer of multimodal data streams, including audio, video, and control signals, between various modules. A defining characteristic of this framework is its strict modularity. It enables the independent development of functional modules in python without requiring knowledge of the broader system. This decoupled approach is particularly advantageous for integrating complex, parallel components such as the backchannel generation module. Due to its increasing popularity, the framework integrates many existing extensions for TTS, LLM, STT,... which makes development of a very basic AI voice agent relatively easy. Existing templates and projects already implement this pipeline, leaving only the development of the backchannel module as our responsibility.

²<https://github.com/TEN-framework/ten-framework>

Compared to alternative frameworks such as Pipecat³ or LiveKit⁴, the Ten Framework was selected for its architectural flexibility and integrated graph interface. This interface allows for the visualization of system extensions and their respective data-flow interconnections, as shown in fig. 3.2. By utilizing a graph-based representation, the complexity of the full-duplex communication pipeline becomes more manageable, allowing for precise control over the synchronization between speech recognition, linguistic processing, and real-time listener feedback.

Complementing the backend architecture, the Ten Framework provides a specialized frontend interface, as illustrated in fig. 3.3. Optimized specifically for conversational AI agents, this framework facilitates comprehensive multimodal interaction by supporting both audio streams and real-time video input. Added perceptual capabilities, by integrating them using the module-based interface, allow the voice agent to process visual data from the device’s camera, which can then be interpreted alongside linguistic input to enhance conversational context. Furthermore, the web-based interface is designed with a focus on accessibility and user experience, its streamlined, intuitive design ensures that interlocutors can engage in fluid dialogue without the cognitive overhead often associated with complex robotic control systems.

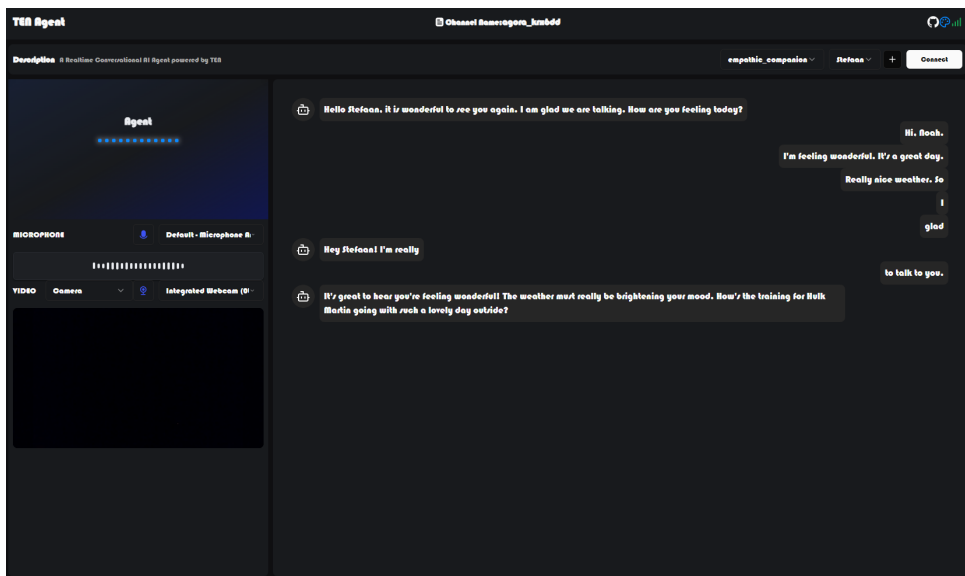


Figure 3.3: Ten Framework - Ten Agent frontend

3.1.3 Integration with Pepper

For the user study, the Ten Framework was integrated with the Pepper humanoid robotic platform, facilitating a more immersive interaction aligned with human conversational expectations. The humanoid morphology of Pepper serves as a significant social indication, unlike the transactional nature of interactions with stationary smart speakers [2]. Pepper’s physical presence encourages users to engage in more naturalistic, relational conversation.

Given the constraints of the robot’s legacy hardware, a hybrid integration strategy was adopted. A tablet was mounted on the robot’s torso to serve as the primary interaction hub, handling both the visual interface and the audio capture and output required by the Ten Framework. This configuration minimized architectural overhead while ensuring robust system performance. To maintain the illusion of an integrated agent, an external synchronization module was developed to map the robot’s movement behaviour to the speech. By utilizing a WebSocket connection for low-latency communication, the robot’s physical gestures were synchronized with textual cues from the primary pipeline. Crucially, this included the coordination of backchannels, where physical movements, such as head nods, were executed synchronized with audio cues to reinforce the visual impression of active comprehension.

³<https://www.pipeecat.ai>

⁴<https://livekit.com>

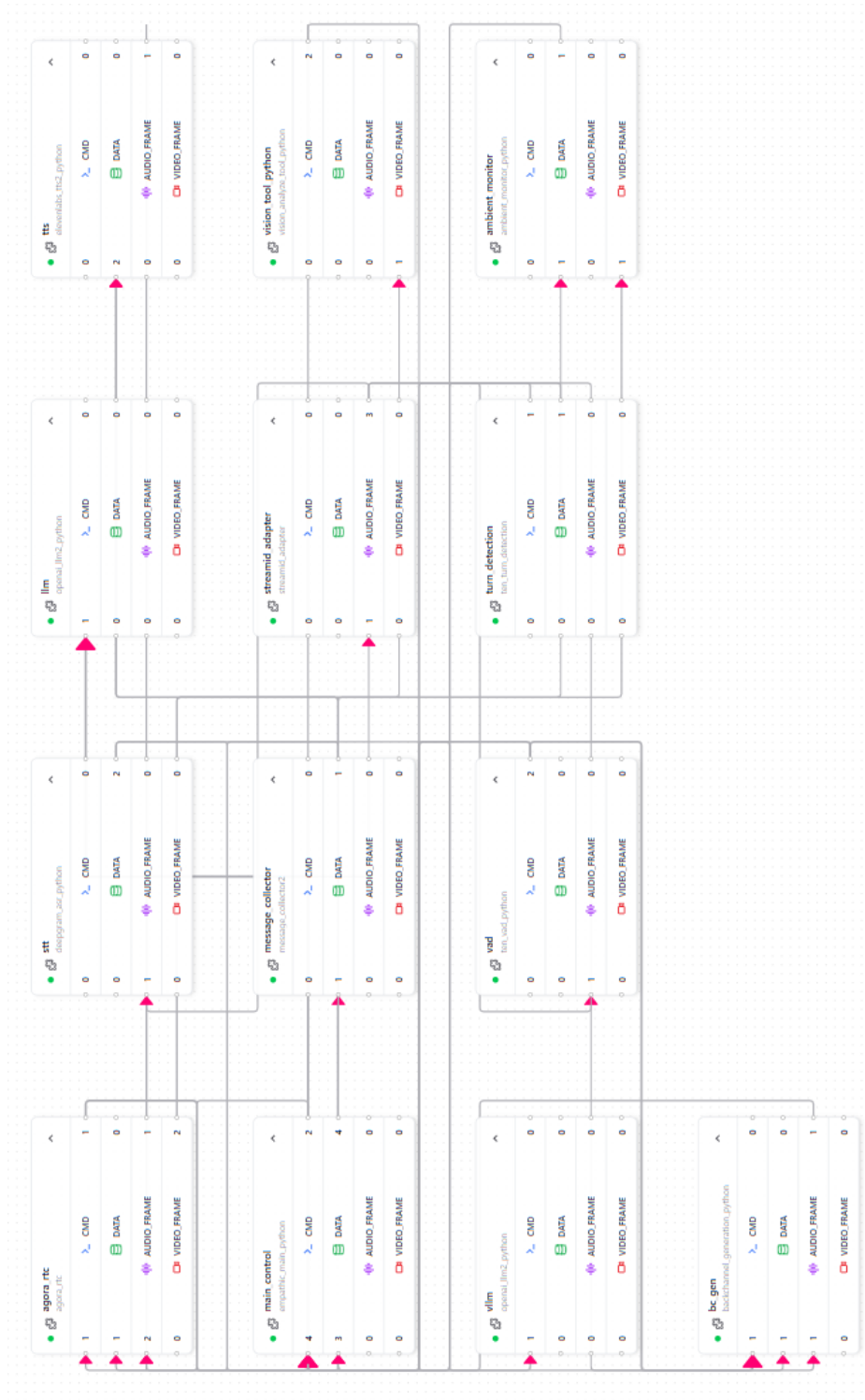


Figure 3.2: Ten Framework - Graph overview of complete setup

3.2 Backchannel module

To facilitate the autonomous production of backchannels, a specialized module was developed inside the Ten Framework, incorporating a dual-stream architecture for independent timing and typology prediction. This module utilizes a bimodal processing approach: audio frames are analysed for timing prediction, while textual data streams inform the type selection.

A critical feature of this implementation is its fully asynchronous execution. By decoupling the backchannel module from the primary conversational pipeline, the system ensures that listener responses are generated without influencing the processes of the main pipeline.

Because incorporating broader conversational context (e.g., demographics or persona) remains a significant computational challenge due to architectural overhead, the actual context of the conversation was left out for easier backchannel generation. Consequently, a parametric optimization study was conducted to identify a general configuration that maximizes perceived naturalness across diverse interactions (look at section evaluation). The following sections provide a detailed technical analysis of the underlying timing algorithms and the typological prediction frameworks utilized within this architecture.

3.2.1 Backchannel Timing

The primary challenge for the backchannel module lies in the accurate identification of BOPs. To investigate the most effective solution, this research evaluates two distinct algorithms: a rule-based heuristic approach and a ML-based approach.

Both implementations process raw audio frames, through different analytical frameworks. The rule-based method utilizes signal processing to extract specific prosodic features, such as pitch contours, energy levels, and silence durations, to identify backchannel triggers. Conversely, the ML approach leverages a transformer-based architecture, feeding the raw audio data into a neural network designed to capture complex, non-linear temporal patterns that may elude traditional heuristic markers. By testing both methods, the study aims to determine whether explicit linguistic rules or implicit learned representations provide a more robust foundation for real-time interactional backchannel generation.

3.2.1.1 Rule-based approach

The first implementation utilizes a heuristic framework based on the research of Ward et al. This algorithm identifies BOPs by analysing specific prosodic features, primarily focusing on pitch contours (falls and rises) and energy dynamics, coupled with temporal markers such as silence duration [39, 51].

A significant advantage of this rule-based approach is its high degree of transparency and reproducibility. As delineated in listing 1, the heuristic is defined by a discrete set of conditional triggers, allowing for direct translation into efficient code. Furthermore, because the algorithm operates on low-dimensional prosodic features rather than high-dimensional neural representations, it offers better computational performance. This low processing overhead per audio frame ensures that the module can maintain the strict real-time requirements of a conversation.

```

1 def process_frame(audio_frame, time_since_last_bc, pitch_variation_detected):
2     audio_buffer = audio_buffer + audio_frame
3     if pitch_variation_detected:
4         if is_silent(audio_buffer, 400ms):
5             output_bc()
6             pitch_variation_detected = False
7             audio_buffer.empty()
8             time_since_last_bc = audio_frame.time
9     else:
10        if has_pitch_variation(audio_buffer, 100ms, 30hz):
11            if not is_silent(audio_buffer, 1000ms)
12                and audio_frame.time - time_since_last_bc > 1400ms:
13                pitch_variation_detected = True
14                audio_buffer.empty()

```

Listing 1: Pseudocode for the heuristic of Ward et al.

The implemented heuristic operates on a continuous stream of audio frames, utilizing a state-machine approach to monitor prosodic triggers. The algorithm maintains two primary state variables that regulate the frequency and sequence of the output:

time_since_last_bc, a temporal constraint that tracks the duration elapsed since the previous backchannel reproduction. This variable prevents “over-backchanneling” by enforcing a cool-down period, ensuring the robot does not produce repetitive or rapid-fire responses.

pitch_variation_detected, a boolean flag that manages the two-stage detection logic. It determines whether the system is in the primary detection phase, monitoring for significant pitch fluctuations, or the secondary confirmation phase, monitoring for a specific duration of silence following a detected pitch event.

To facilitate the execution of this logic, several signal processing helper functions are integrated into the pipeline:

is_silent(buffer, duration): This function performs an energy-based analysis of the most recent audio buffer. It compares the Root Mean Square (RMS) amplitude against a calibrated threshold to determine if the segment is silent. It returns True if the energy levels remain below the threshold for the specified duration.

has_pitch_variation(buffer, duration, threshold): This function analyses the pitch within the buffer to identify significant rises or falls. It evaluates whether the rises or falls exceed a predetermined threshold within the given temporal window, signalling a potential communicative transition point.

output_bc(): Once the conditions for a BOP are satisfied, this function triggers the transmission of the backchannel to the subsequent typological classification module (section 3.2.2), which determines the specific lexical or non-lexical form to be synthesized.

The full state-machine diagram can be found in fig. 3.4, which includes all the states the algorithm traverses before reproducing a backchannel.

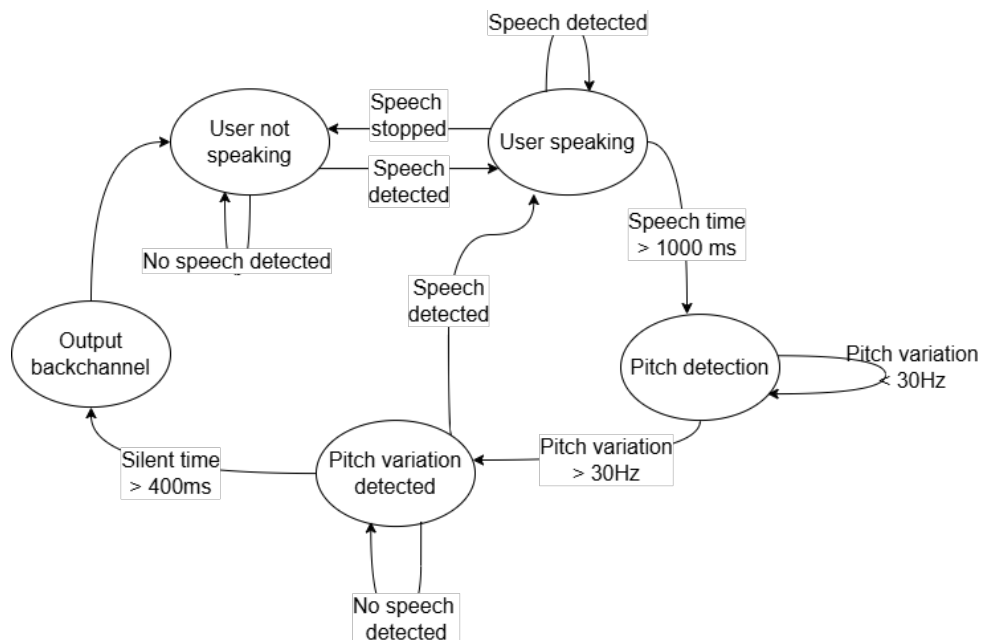


Figure 3.4: Rule-based approach: State-machine

3.2.1.2 Machine-Learning-based approach

The second approach to backchannel reproduction leverages current advancements in deep learning through the VAP model [16]. Originally developed for turn-taking coordination, VAP is a self-supervised transformer-based architecture that generates fine-grained projections of speaker activity over a 2-second

future window. By modelling the probability of voice activity for both interlocutors simultaneously, the system can anticipate conversational shifts and transitions before they occur. While the primary objective of VAP is turn-taking, its extensive pre-training on a diverse corpus of conversational audio provides a robust latent representation of interactional dynamics. This allows the model to be effectively adapted for more nuanced tasks through fine-tuning [23]. Building upon this foundation, the MaAI⁵ project developed a specialized backchannel prediction module by fine-tuning the VAP architecture on the online conversation dataset [21]. This fine-tuning process shifts the model’s predictive focus from general turn-transitions to the specific prosodic and temporal markers that characterize backchannel opportunities. By utilizing this neural approach, the system transitions from reactive detection (waiting for a pitch fall) to proactive prediction, potentially increasing the interactional synchrony between the robotic agent and the user.

3.2.2 Backchannel Type Prediction

The secondary objective of the backchannel generation module is the prediction of the appropriate BCP. This classification process is driven by textual data retrieved asynchronously from the ASR system. To maintain the strict latency requirements, the module processes partial transcripts rather than waiting for finalized utterances. This incremental approach ensures that semantic analysis does not compromise temporal synchrony.

Based on the available linguistic context, the system employs a predictive model to categorize the backchannel into one of two frameworks: a multi-class model (section 3.2.2.2), which predicts the backchannel types from a specific predefined set of the most common backchannels, or a dual-class model (section 3.2.2.3), which divides responses into the functional categories of continuers and assessments, predicts it and selects an accurate backchannel based on the predicted type. The following sections delineate the CANDOR Corpus, which served as the foundational dataset for model training, followed by a comparative analysis of both prediction architectures and their respective performance benchmarks.

3.2.2.1 CANDOR Corpus

To train and validate the backchannel type prediction models, this research utilizes the CANDOR Corpus, a massive collection of approximately 850 hours of dyadic conversational data [42]. The corpus is comprised of 1,656 semi-structured video-chat interactions between participants who were given no specific interactional constraints or prompts. This absence of formal instruction ensures a high degree of naturalness of the conversations, as the resulting recordings capture the unscripted, naturalistic nuances of human-to-human discourse. For the purposes of this study, the corpus is particularly valuable due to its high-resolution processing: all interactions have been transcribed and annotated for backchannel occurrences. These annotations include precise temporal indications of backchannels, providing a robust ground truth for training models to distinguish between varied listener responses.

The initial extraction from the CANDOR corpus yielded a total of 458,418 backchannel instances. In its raw state, the dataset exhibited high lexical diversity, with over 14,600 unique tokens. However, this diversity was largely a due to inconsistent transcription artifacts, such as varying punctuation (e.g., distinguishing between “yeah”, and “yeah”). To refine this dataset for effective model training, some preprocessing steps were executed. Redundant tokens were unified by removing non-alphabetic characters, such as punctuation markers and standardizing casing. Tokens with extremely low occurrence rates were removed to prevent the model from learning noise or speaker-specific utterances. Following these refinement steps, the vocabulary was condensed to 127 unique backchannel types.

3.2.2.2 Multiclass prediction

The first iteration of the typological prediction module utilizes a fine-tuned DistilBERT architecture⁶. DistilBERT is a distilled version of the BERT model that maintains significant linguistic performance while offering reduced latency and a smaller memory footprint. Leveraging its self-supervised pre-training on expansive textual corpora, the model possesses an inherent understanding of semantic context and syntactic structure.

To adapt this model for backchannel synthesis, it was fine-tuned using conversational pairs of backchannel with its preceding text of the conversational partner, extracted from the CANDOR Corpus. The classification task was constrained to the ten most frequent backchannel responses [18], table 3.1. The

⁵<https://github.com/MaAI-Kyoto/MaAI>

⁶<https://huggingface.co/distilbert/distilbert-base-uncased>

model was trained to analyse the preceding user utterance (the context window) and predict the most appropriate specific lexical response from this prioritized set.

Backchannel	Frequency
yeah	73983
mhm	36559
okay	13136
oh	13095
right	10955
mm	7672
uh huh	7386
uh	5313
yes	3150
wow	2875

Table 3.1: Ten most common backchannels in CANDOR corpus

Despite the semantic depth of the DistilBERT architecture, the multi-class lexical model encountered significant performance limitations. First, the task of distinguishing between nearly identical non-lexical vocalizations, such as “uh-huh” or “hmm”, presents a challenge in inter-annotator agreement, as even human listeners frequently struggle to achieve a consensus on the specific lexical form based on textual context alone. Furthermore, the model was severely impacted by a pronounced class imbalance inherent in the English language and the CANDOR corpus. The backchannel “yeah” occurs more than 2 times as frequently as the next most common entry, causing the model to converge on a biased “majority-class” prediction. This lack of stochastic variety resulted in repetitive and predictable agent behaviour, failing to capture the nuanced dynamics of naturalistic feedback. To address these issues of low predictive accuracy and lexical bias, a revised classification framework was developed, shifting the focus from individual word selection to broader functional categories.

3.2.2.3 Dual class prediction

Subsequently, the type prediction was switched towards a dual class prediction model. Instead of predicting the ten most common backchannels, the DistilBert model was fine-tuned on predicting only between assessments and continuers based on the preceding text. The CANDOR dataset was preprocessed, assigning each backchannel a specific type. Some assignments of backchannels with its corresponding type: Continuers: ‘yeah’, ‘mhm’, ‘uhhuh’, ‘mm’ and ‘uh’. Assessments: ‘okay’, ‘oh’, ‘right’, ‘yes’ and ‘wow’.

To resolve the predictive challenges identified in the multi-class model, the typological framework was transitioned to a dual-class functional prediction architecture. In this refined paradigm, the DistilBERT model was fine-tuned specifically to distinguish between the two primary categories of backchannels: continuers and assessments [7, 23]. This binary classification approach mitigates the impact of class imbalance by aggregating infrequent lexical tokens into broader, functionally equivalent categories.

To facilitate this training, a secondary preprocessing step was conducted on the backchannels in the CANDOR dataset where individual backchannel instances were mapped to their corresponding class. The classification was operationalized based on the following taxonomic assignments. Continuers were identified as neutral vocalizations and lexical markers intended primarily to signal attention and encourage the speaker to maintain their turn (e.g., “yeah”, “mhm”, “uh-huh”, “mm”, and “uh”). Meanwhile, assessments indicate evaluative or reactive markers that signal a degree of emotional or cognitive processing of the speaker’s content (e.g., “okay”, “oh”, “right”, “yes” and “wow”).

By simplifying the output space to these two distinct communicative intents, the model achieves higher predictive reliability. This functional split allows the system to accurately determine the interactional intent required at a specific BOP, delegating the final lexical selection to the bifurcated execution logic described in the next section.

Once a backchannel type is classified, the system follows two distinct operational paths based on the functional role of the response. When the model predicts a continuer, the system performs a random selection from a predefined set of non-context dependent vocalizations (e.g., “yeah”, “mhm”, “uh-huh”, “mm”, and “uh”). Because these markers serve primarily to signal active listenership without providing evaluative feedback, their random rotation does not negatively impact perceived naturalness. It introduces a degree of “human-like” variability that prevents the agent from sounding repetitive.

In contrast, assessments require a significantly more sophisticated approach due to their inherent context-dependency. Unlike continuers, an assessment (e.g., “wow”, “right”, or “oh”) contains an evaluative or emotional weight that must align with the speaker’s propositional content. To ensure this semantic coherence, the system routes the user’s partial utterance to a high-performance, low-latency LLM (GPT-4o-mini). The model analyses the immediate linguistic context to select the most appropriate marker from a prioritized set of context-dependent backchannels (“okay”, “oh”, “right”, “yes”, and “wow”). This ensures that a user sharing surprising news is met with a “wow” rather than a misplaced “okay”, thereby maintaining the interactional integrity and emotional intelligence of the robotic agent.

3.2.3 Reproducing the backchannel

To satisfy the latency requirements of interactive dialogue, a static pre-caching mechanism was implemented for backchannel synthesis. Rather than generating audio dynamically upon each detection, which would introduce unacceptable network or processing delays, the audio fragments for the ten most frequent backchannels were generated prior to the interaction (and cached for later use). By pre-generating these high-probability assets, the system ensures that as soon as a BOP is detected and its typology is classified, the corresponding audio frames can be retrieved instantaneously from local memory and transmitted to the output buffer.

To ensure communicative coherence and maintain the “social persona” of the agent, all pre-cached backchannels were generated using the same TTS model and vocal profile utilized for the agent’s primary conversational voice. This consistency in timbre and prosody prevents the “uncanny valley” effect that can occur when listener feedback sounds perceptibly different from the agent’s main voice. This optimized pipeline effectively decouples the high-latency synthesis task from the time-sensitive interactional task, allowing the robot to provide immediate, vocally consistent feedback that reinforces the user’s perception of an attentive and socially present interlocutor.

Evaluation

The primary objective of this study was to validate user perceptions of backchanneling behaviour within the HRI loop. Specifically, the research evaluates (i) whether the presence of backchannels significantly enhances the perceived naturalness of the dialogue compared to a baseline condition (no backchannels) and (ii) seeks to identify which timing algorithm, rule-based or ML-based, best aligns with human conversational expectations.

This validation phase serves as a critical optimization step for future deployment in a care home environment. In such settings, where communicative clarity is paramount and the margin for interactional error is minimal, ensuring a robust and human-like feedback loop is essential for user trust and engagement. While the subsequent deployment and longitudinal evaluation within the care home remain outside the scope of this thesis, the current study provides the empirical foundation necessary to refine the system for high-stakes, real-world social conversations.

4.1 Participants

The evaluation was conducted with 20 participants (12 male, 8 female; 22.05 ± 1.16 years). Each participant engaged in approximately 20 minutes of total interaction time with the Pepper robot, followed by a psychometric questionnaire assessing the conversational experience. To maximize recruitment efficiency and ensure a diverse range of feedback, the study was conducted with related research, all participants took part in both user studies. A coordinated protocol was implemented wherein each participant first completed the other experimental study (taking 30 to 40 minutes) before transitioning to the current backchannel study. This sequential approach was designed to minimize inter-experimental interference and reduce participant fatigue, ensuring that the data collected for each study remained independent and robust. The total interaction time (about 1 hour) allowed participants to move beyond the initial novelty of the humanoid platform, providing more reliable insights into the long-term naturalness of the robotic backchanneling behaviour.

To ensure the internal validity of the findings, each participant completed a demographic and background assessment. This survey documented the participants' prior exposure to social robotics and voice assistant technologies, as these factors can significantly influence a user's baseline expectations and their threshold for natural interaction.

In addition to demographic and behavioural traits, the native language ($L1$) of each participant was recorded. This measure was necessitated by research highlighting the significant socio-cultural variance in listener behaviour, as the expected frequency and typology of backchannels are heavily influenced by a speaker's primary language [17]. Participants could choose to either talk in Dutch or English with the robot. Accounting for linguistic background is essential for evaluating the perceived naturalness of the robot's feedback. By documenting these $L1$ profiles, the study can better contextualize individual user ratings and identify whether the robot's performance aligns more effectively with certain linguistic groups than others.

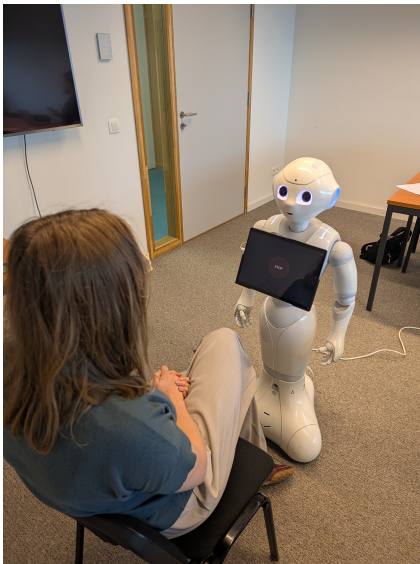
Furthermore, the study incorporated a self-rated talkativeness scale. By recording individual predispositions toward verbal engagement, the research accounted for potential confounding variables. For example, a highly talkative user may perceive backchanneling differently than a more reluctant participant. These measures allow for a more nuanced statistical analysis, ensuring that the perceived naturalness of the robotic behaviour is evaluated relative to the user's specific interactional profile and technological familiarity. All procedures were approved by the Departmental Ethics Committee, and participants provided written informed consent.

4.2 Setting

The physical setup (illustrated in fig. 4.1), took place in a meeting room within the research lab. The participant was asked to sit down looking at the robot, which was placed facing to a wall for better face tracking. The participant had to wear an external microphone to bypass the technical limitations of the robot’s onboard sensors and to allow for session recordings, which could be used for analysis later on. This configuration significantly reduced environmental noise interference, facilitating precise VAD and ASR.

To manage the interaction, a tablet was mounted to the robot’s chassis, serving as both an intuitive tactile interface for session control and the primary audio output node. Due to the legacy constraints of the robot’s native hardware (which was used due to its availability in the research lab), routing the synthesized speech and backchannels through the tablet minimized internal processing overhead, ensuring no additional overhead in the response times required for naturalistic feedback.

The Ten framework and the ML models for backchannel prediction, which served as the main processing needed for the robot, was hosted on a high-performance external workstation equipped with a dedicated GPU. This distributed approach decoupled high-level inference from the tablet, which wasn’t able to run all necessary software and allowed the computationally intensive modules to remain active between sessions, significantly reducing setup overhead and ensuring consistent performance across all twenty participants.



(a) Front view of the Pepper robot



(b) Participant interaction

Figure 4.1: Experimental setup: showing (a) a front view of the robot and (b) a participant interacting with it.

4.3 Protocol

The overall structure of the study is shown in table 4.1. First, informed consent is obtained from all participants. Once this was obtained, the participant completes a short demographic questionnaire and undergoes a brief, unrecorded familiarisation phase (~ 3 minutes) with the robot, in which they had a small conversation with the baseline model.

The core of the experiment consists of three conversation blocks, each approximately 5 minutes long, with the participant talking with Pepper about a topic chosen from a small set of standardised topic cards (“Something interesting recently”, “A special place” and “A hobby or activity”). Each conversation block is conducted under a different backchannel condition: Baseline model, Rule-based model, or ML-based model. The order of conditions is fully counterbalanced across participants using a Latin square design (six possible orders), and topic cards are independently counterbalanced across blocks. Participants were not informed of the specific condition they were talking to.

Following each test condition, participants will complete post-test questionnaires to capture their perception of the system’s performance and the technology itself. After all three blocks are complete,

an end-of-session questionnaire will compare their experiences across the three conditions, followed by a brief semi-structured interview, debriefing, and compensation. Total session duration is approximately 30 minutes per participant.

Procedure	Duration
Introduction	3 min
Informed Consent	1 min
Demographics Survey	2 min
Familiarisation	3 min
First Interaction	5 min
Questionnaire	2 min
Second Interaction	5 min
Questionnaire	2 min
Third Interaction	5 min
Questionnaire	2 min
Post-Interaction Questionnaire	3 min
Post-Interaction Interview	2 min

Table 4.1: Structure of the study

4.4 Methodology

The efficacy of the backchannel generation modules is evaluated through a multi-dimensional framework that combines subjective user experience with objective interactional metrics. This comprehensive approach ensures that the system’s performance is analysed not only through the user’s conscious perception but also through the subconscious temporal dynamics of the dialogue. The following data collection instruments were utilized: A Per-Block Subjective Assessment following up on each experimental condition, participants completed a questionnaire measuring perceived naturalness, attentiveness, and engagement. This instrument integrated the Coordination and Connection subscale of the Connection-Coordination Rapport (CCR) scale [26], a standalone attentiveness item, and a Mean Opinion Score (MOS)-style metric for perceived naturalness [31]. At the end of each session a global Perception and Manipulation Check was performed, the Partner Modelling Questionnaire (PMQ) was administered to capture the user’s high-level mental model of the agent [13]. The end-of-session questionnaire also included comparative ranking items of the three conditions and a manipulation check to ensure the experimental variables were perceptibly distinct. Alongside the questionnaires a qualitative semi-structured interview took place after each session. Brief (~3 minutes) audio-recorded interviews were conducted to elicit open-ended insights into the participant’s experience, allowing for the identification of nuanced behaviours that quantitative scales might overlook. After the study Objective Behavioural Metrics were extracted. Interactional data were extracted from system logs and audio recordings, focusing on both temporal dynamics like inter-turn latency and conversational structure, including speaking time, turn counts, and the frequency of interruptions or backchannels. This integrated methodology provides a granular understanding of the system’s impact on conversational flow. By correlating subjective rapport with objective timing data, the study can identify the specific architectural components that most effectively contribute to naturalistic human-robot collaboration. In the next section, the complete methodology is explained. Starting from the main constructs this user study was based upon.

4.4.1 Four Constructs

The evaluation framework is centred on four distinct constructs, each grounded in established backchanneling literature and designed to capture specific dimensions of the user’s interactional experience. These constructs bridge the gap between the system’s technical performance and the participant’s subjective perception:

Construct 1: Naturalness The first construct centres on the interactional naturalness, representing the participant’s global impression of how human-like and conversationally “normal” the interaction felt. As the most widely utilized outcome measure in spoken dialogue evaluation, it serves as the primary proxy for the perceived authenticity of the exchange.

This construct is hypothesized to exhibit the most significant variance between the baseline (no-backchannel) condition and the active backchanneling conditions. In human-to-human dialogue, listener feedback is a continuous requirement in order to perceive some sort of understanding at the receivers side [5]. This perceived absence of social presence can make the interaction feel mechanical or uncanny.

Consequently, we anticipate that when the system employs timing algorithms to provide responsive listener feedback, the interaction will be perceived as significantly more natural than the baseline. By aligning the robot’s behaviour with human conversational norms, we expect both the rule-based and ML-based models to outperform the no-backchannel condition in this metric.

Construct 2: Attentiveness The second construct evaluates the perceived attentiveness of the robot, defined as the participant’s belief that the agent is actively listening. This metric assesses whether the robot is perceived as a “present” conversational partner rather than a passive system merely waiting for its turn to speak.

Perceived attentiveness is fundamentally linked to backchanneling behaviour, as these signals serve as the primary indicator of active listening. A robot that remains silent while a user speaks appears inattentive, even if its eventual response is contextually appropriate. This is because attentiveness in human dialogue is judged incrementally during the speaker’s turn, rather than retrospectively based on the final output. Without continuous listener feedback, the speaker lacks the social confirmation that their message is being received and understood.

We hypothesize that this construct will show the most significant effect across the three conditions. Both backchannel-enabled conditions are expected to significantly outperform the no-backchannel baseline. Furthermore, we anticipate that differences in the temporal precision of the rule-based versus ML-based timing approaches may lead to observable variances in attentiveness scores, as poorly timed feedback may be perceived as distracted or mechanical rather than attentive.

Construct 3: Engagement The third construct evaluates conversational engagement, defined as the participant’s subjective investment in the dialogue, specifically their interest, focus, and willingness to sustain the interaction. This metric is distinct from perceived attentiveness (which evaluates the robot) and general enjoyment: a participant can be engaged in a difficult conversation they do not necessarily enjoy.

Engagement represents the primary speaker-side benefit of synchronized listener feedback. In human-to-human interaction, proper backchanneling serves as a social fuel that encourages the speaker to elaborate and maintain their turn. Conversely, interacting with an unresponsive agent is often mentally exhausting.

We hypothesize that both backchanneling conditions will significantly outperform the baseline no-backchannel condition. By ensuring the speaker feels heard, the system is expected to reduce the cognitive and emotional burden on the user, thereby fostering a more resilient and invested state of engagement.

Construct 4: Social Presence The final construct evaluates social presence, defined as the participant’s perception of the robot as a social entity rather than a technological object. While attentiveness and engagement focus on the mechanics of the exchange, social presence addresses the broader partner-modelling question: “To what extent does the user perceive the agent as a ‘someone’ rather than a ‘something’?”

In the absence of backchannels, a robot can feel less understanding, giving the feeling of not really being present. However, the production of contextually appropriate listener feedback serves as a powerful cue for social agency. When the robot provides a well-timed backchannels, it signals a shared psychological space, transitioning from a reactive machine to an active participant in the conversation.

We hypothesize that the backchanneling conditions will yield significantly higher social presence scores compared to the no-backchannel baseline. By adhering to the subtle social protocols of human dialogue, the system is expected to cross the threshold from a speech-processing interface to a perceived social partner.

4.4.2 Validated Scales

To correctly assess these four constructs, three scales were carefully selected. In the next section, all three scales are explained with their rationale for using it in this study. All scale choices follow the Saad et al.

ACM TiHRI guideline and GMU HRI Scale Database¹, which were rated against a 13-item psychometric quality guideline [45]. Only scales $\geq 77\%$ on this guideline were used.

Scale 1: Connection-Coordination Rapport (CCR) To quantify the quality of the human-robot interaction, this study utilizes the Connection-Coordination Rapport (CCR) scale [26]. Grounded in established interpersonal communication literature, the CCR scale is specifically validated for HRI contexts and decomposes rapport into two distinct factors. The first factor, Connection (12 items), evaluates the relational and affective dimensions of the interaction, such as mutual understanding and warmth. Conversely, the other factor, Coordination (6 items), measures the behavioural and synchronic dimensions of the dialogue.

The CCR scale was selected primarily due to its perfect rating (100%) according to the GMU psychometric guidelines. Furthermore, the scale is specifically tailored for HRI, ensuring that the items remain contextually relevant when applied to a human-robot interaction. Specifically, three key dimensions within this subscale are operationalized to measure our primary constructs (defined in section 4.4.1). Perceived attentiveness is measured through the Coordination dimension “Pepper was attentive to what I was saying” and conversational engagement is captured by the Connection dimension “Pepper seemed engaged in the conversation”. Also naturalness-via-smooth-flow can be measured through the smoothness item in the Coordination subscale: “The conversation flowed smoothly”.

CCR was originally published with single-word items (like “Coordination”, “Bond”, “Caring”, ...). Due to standard HRI questionnaire practices, where full sentences are used instead of words for better understanding and the preservation of the construct in each item measure, were the single-word items converted to sentences (e.g. “Pepper and I coordinated well during the conversation”). The full list of item mappings can be found in Appendix A.

Scale 2: Partner Modelling Questionnaire (PMQ) To evaluate the participant’s high-level mental model of the robot, the Partner Modelling Questionnaire (PMQ) was administered at the end of the experimental session [13]. Developed specifically to assess how users conceptualize their conversational partners in speech-interface interactions, the PMQ comprises eighteen items across three subscales: Competence and Dependability, Human-Likeness, and Communicative Flexibility.

The Human-Likeness subscale is of particular relevance to this study, as it measures the warmth-coldness, social-transactional, and personal-generic dimensions, traits that are fundamentally influenced by the presence and quality of backchanneling. While the PMQ was originally validated for screen-based or speaker-based speech interfaces, its constructs align closely with the social presence requirements of humanoid robotics. A notable methodological consideration is the potential for higher baseline human-likeness scores due to the robot’s physical embodiment. However, the scale remains a robust instrument for identifying the incremental social value provided by the backchanneling modules.

The PMQ was intentionally administered as a single end-of-session measurement rather than a per-block assessment. This allows participants to provide a reflective, considered judgment of the system’s overall social identity after experiencing the full range of interactional conditions. To bridge this global perception with specific experimental variables, the PMQ results are complemented by comparative ranking items, wherein participants directly evaluate the three backchanneling conditions against one another.

Scale 3: Single MOS-style naturalness item To directly quantify conversational authenticity, a single Mean Opinion Score (MOS)-style item was incorporated: “How natural did this conversation feel?”, rated on a five-point Likert scale ranging from Very Unnatural to Very Natural [31].

While single-item measures often possess lower psychometric reliability compared to multi-item scales, this metric does not stand in isolation. Instead, it is compared against two additional measures to ensure robust findings. The CCR Coordination Subscale, specifically item: “Smooth flow”, captures the rhythmic and temporal aspect of naturalness. The Comparative Ranking item, “Which conversation felt most natural?”, requires participants to make a direct relative judgment between the three experimental conditions.

By synthesizing these three distinct perspectives, the study establishes convergent evidence. This multi-methodological approach provides high confidence in the final naturalness conclusions, effectively compensating for the inherent variance of any single instrument.

¹<https://hriscaledatabase.psychology.gmu.edu/>

Objective Measures Additionally to the subjective measures collected after each session, transcripts, session recordings and backchannel logs were collected. The session recordings were analysed to determine the turn taking duration and inter-turn latency. Each session recording was annotated with the end-of-turn and start-of-turn indicators (in milliseconds) and processed per condition to retrieve the mean turn duration and inter-turn latency for each condition.

Results

The evaluation was conducted over a one-week testing period utilizing a within-subjects design ($N = 20$, $L1$ -language = Dutch), wherein each participant interacted with all three experimental configurations: the baseline (no backchannels), the rule-based heuristic model, and the ML-based model. This setup was specifically designed to determine the perceived importance, temporal precision and naturalness of automated backchannels within human-robot interaction. The conversational durations across the three experimental conditions were as follows: Baseline Condition (4.63 ± 0.89 minutes), Rule-Based Condition (4.28 ± 0.67 minutes) and ML-Based Condition (4.3 ± 1.03 minutes). The relatively uniform interaction times (~ 3.5 to 5.3 minutes per block) yield an aggregate interaction time of approximately 13.2 minutes per participant.

This chapter presents a comprehensive analysis of the data collected during the experimental evaluation. The findings are structured per subscale and provide insights into the performance of the three different backchannel generation models.

5.1 Subjective Measures

The evaluation instruments were synthesized using a multi-scale approach aligned with the ACM TiHRI guidelines [45]. The per-block evaluations integrated the Connection-Coordination Rapport (CCR) scale [26], utilizing both the Connection subscale (12 items; Cronbach's $\alpha = .841$) to capture affective alignment and the Coordination subscale (6 items; Cronbach's $\alpha = .761$) to isolate behavioural synchrony. These subscales were supplemented by a standalone perceived attentiveness metric and a single-item Mean Opinion Score (MOS)-style naturalness rating. Additionally, at the end of each session, results from a comparative ranking scale together with the PMQ scale were collected to retrieve a retrospective view on all three versions. This combined with objective measured data, results in a highly qualitative dataset from which we could further observe the analytical patterns. The subsequent sections detail the descriptive findings derived from these instruments, prioritizing the most statistically significant variations, user response patterns, and cross-conditional trends.

5.1.1 CCR Coordination

The Coordination subscale evaluates the perceived behavioural synchrony of the interaction, including dimensions of conversational flow, turn-taking efficiency, focused attention and balanced engagement. A repeated-measures ANOVA revealed no statistically significant main effect of the backchannel condition on perceived coordination, $F(2, 38) = 1.68$, $p = .199$, $\eta^2 = .028$. Surprisingly, these statistics (detailed in fig. 5.1) indicated that the baseline configuration (no backchannels) achieved both the highest mean coordination score and the lowest standard deviation. This pattern suggests that participants held a somewhat more uniform and favourable perception of the conversational flow when the robot remained silent during listening. While counter-intuitive, this trend may indicate that omitting backchannels entirely removes the risk of misaligned feedback.

5.1.2 CCR Connection

The connection item relates to the relational and affective dimension of the interaction such as warmth, friendliness, closeness, ... A one-way repeated measures ANOVA was conducted to compare the effect of different backchannel generation models (Rule-based, ML-based, and No Backchannels) on the Connection

subscale of rapport. The main effect of the model version was not statistically significant, $F(2, 38) = 1.60$, $p = .215$, with a medium effect size, $\eta^2 = .025$. As detailed in fig. 5.1, the Connection subscale yielded consistently moderate ratings across all three experimental conditions. Interestingly, the baseline (no-backchannel) configuration demonstrated a marginal directional advantage in mean scores. However, inferential statistical testing indicated that this variation was not significant ($p > .05$). Consequently, there is no empirical evidence to suggest that the manipulation of the backchannel generation models directly influenced the participants’ affective or relational bond with the robot.

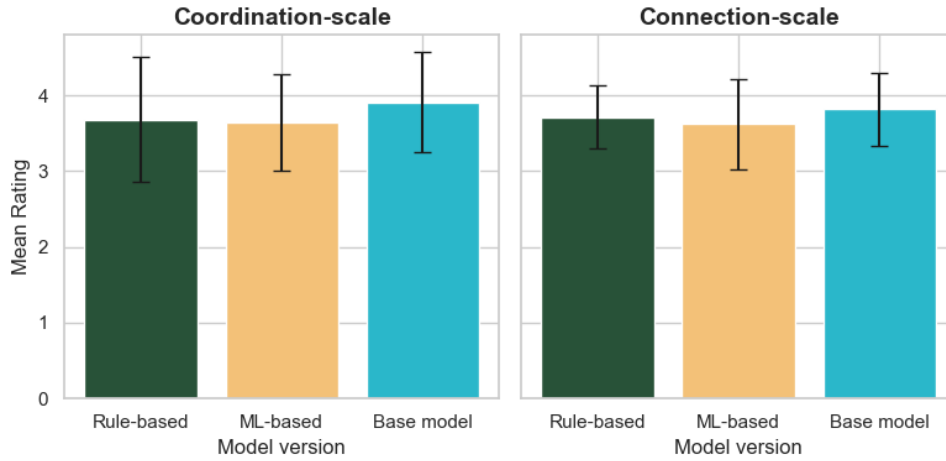
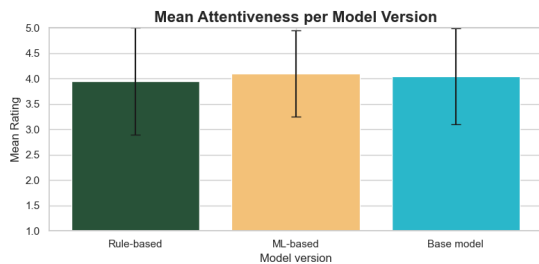
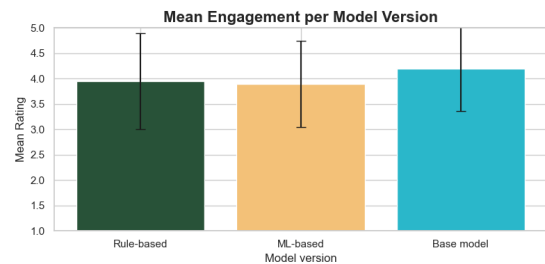


Figure 5.1: CCR-subscale score per Version

To evaluate the perceived attentiveness of the agent, an internal cross-validation was conducted by isolating a specific item from the multi-item Coordination subscale: “Pepper was attentive to what I was saying”. Extracting this item allows for a direct methodological comparison with the standalone attentiveness metric, verifying the stability of the participants’ responses across different questionnaire contexts. A Friedman test performed on this isolated Coordination item revealed no statistically significant differences across the three experimental conditions, $\chi^2(2) = 0.35$, $p = .840$, with a negligible Kendall’s W effect size ($W = .009$). The descriptive means and standard deviations corresponding to this item are visualized in fig. 5.2a.



(a) Attentiveness Item from Coordination Subscale



(b) Engagement Item from Coordination Subscale

Figure 5.2: Attentiveness and Engagement Item from the CCR scale

To complement the global subscale analysis, the single item measuring explicit user engagement within the Coordination framework: “I felt engaged during the conversation”, was isolated and analysed independently. A Friedman test conducted on this item revealed no statistically significant differences across the experimental conditions, $\chi^2(2) = 1.532$, $p = .465$, with a weak Kendall’s W effect size ($W = .038$). The descriptive means and standard deviations for this item are illustrated in fig. 5.2b. Descriptively, the baseline configuration maintained a marginal advantage over the active models.

5.1.3 Standalone Attentiveness Item

While the CCR scale captures the quality of the interaction, the standalone attentiveness item was isolated to evaluate the direct impact of backchannels, which are explicitly used in human conversations

to indicate continuous listening. By isolating this item, it allows direct comparison between all three conditions without the noise from other subscale items. Additionally, this item focused more on the feeling the participant felt during speech: “Pepper felt attentive while I was speaking”, compared to the actual feeling of the robot being understanding to what the participant was saying: “Pepper was attentive to what I was saying”. A Friedman test conducted on this item revealed no statistically significant differences in perceived attentiveness across the three experimental conditions, $\chi^2(2) = 0.760, p = .684$, with a negligible Kendall’s W effect size ($W = .019$). This lack of statistical variance suggests that the presence or absence of vocalized feedback did not significantly alter the participants’ belief that the robot was listening to them.

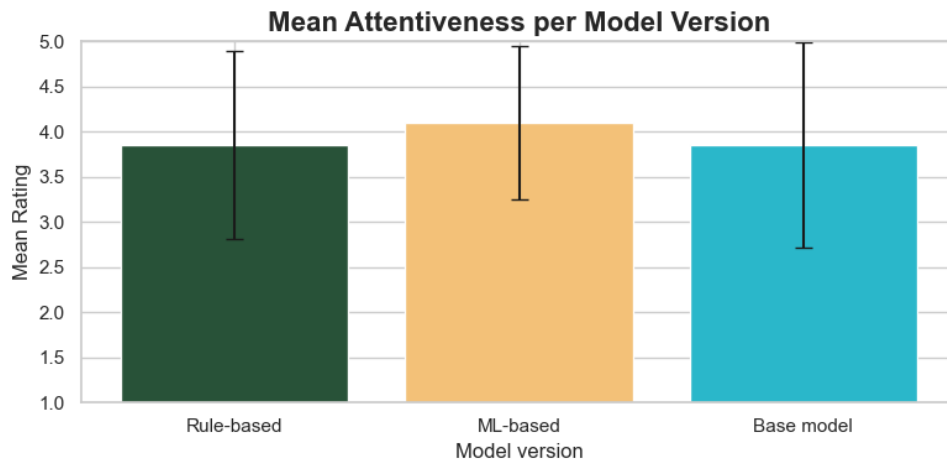


Figure 5.3: Mean Attentiveness per Version

5.1.4 MOS-style Naturalness Item

The final construct from the per-block questionnaires evaluated the naturalness of the interaction. A Friedman test revealed no statistically significant main effect of the experimental condition on perceived naturalness, $\chi^2(2) = 4.20, p = .123$, with a Kendall’s W effect size of $W = .105$.

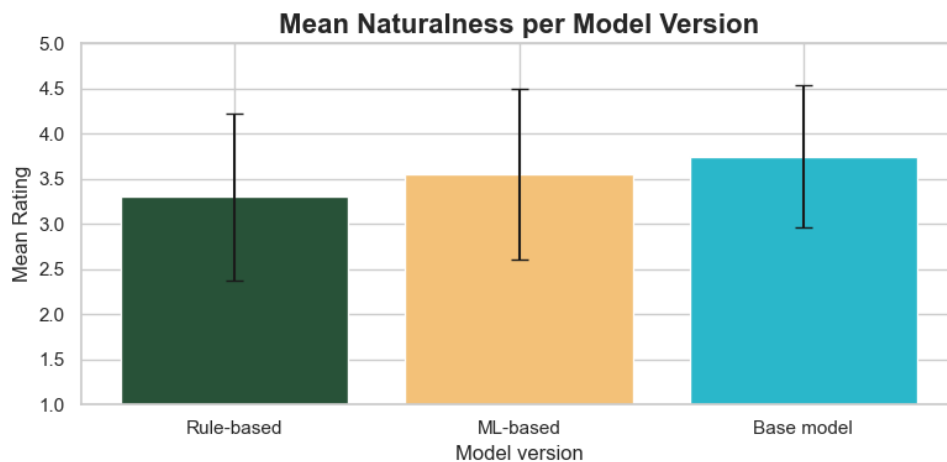


Figure 5.4: Mean Naturalness per Version

5.1.5 End-of-session: PMQ

To capture the participant’s feeling of Pepper as a conversational partner, the Partner Modelling Questionnaire (PMQ) was administered as an end-of-session evaluation. This scale required participants to reflect on their entire interactive experience across all three conditions, rather than evaluating individual experimental blocks. Because this global data structure does not feature measures repeated per condition, inferential statistical testing (such as a repeated-measures ANOVA or a Friedman test) is methodologically

excluded. Consequently, the analysis of the PMQ relies strictly on descriptive statistics. The mean scores and standard deviations for the three PMQ subscales: Competence and Dependability, Human-Likeness, and Communicative Flexibility, are presented in fig. 5.5.

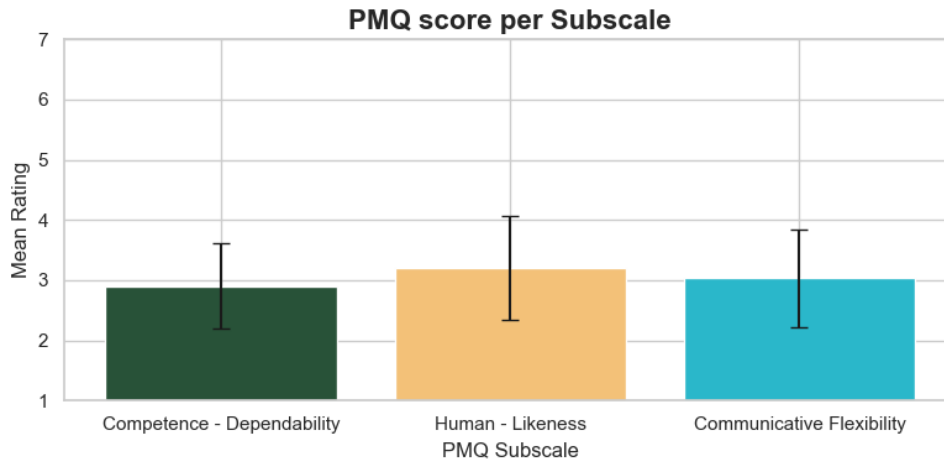


Figure 5.5: End-of-Session PMQ mean scores

When interpreting the PMQ data, it is critical to note the instrument’s inverse scoring architecture, lower numerical values on the seven-point Likert scale denote a more positive perception of the agent. Consequently, the fact that the mean scores for all three subscales fell well below the neutral midpoint of 4 indicates that participants held a globally positive impression of Pepper as a conversational partner.

5.1.6 End-of-session: comparative ranking

To complement the per-block subjective evaluations, participants completed an end-of-session comparative ranking. This measure was designed to capture the users’ retrospective preferences after they had experienced the full range of experimental conditions. The participants had to choose between all three conditions or another option to indicate a similar experience. The distribution of selections for each category is visualized in fig. 5.6. To determine if the distribution of participant choices deviated significantly from chance (an equal three-way split), Chi-squared goodness-of-fit tests were conducted for each ranking category. The inferential results did not reach statistical significance for any of the four questions: Most Natural: $\chi^2(2) = 1.53, p = .466$, Most Attentive: $\chi^2(2) = 0.57, p = .752$, Preferred Again (Future Interaction): $\chi^2(2) = 0.47, p = .790$ and Least Engaging: $\chi^2(2) = 0.47, p = .790$. Despite the lack of statistical significance, which is expected given the high threshold required for Chi-squared tests in a small sample size ($N = 20$), the descriptive frequencies indicate a clear trend. The ML-based model received the highest absolute number of votes for both the “Most Natural” and “Preferred Again” categories.

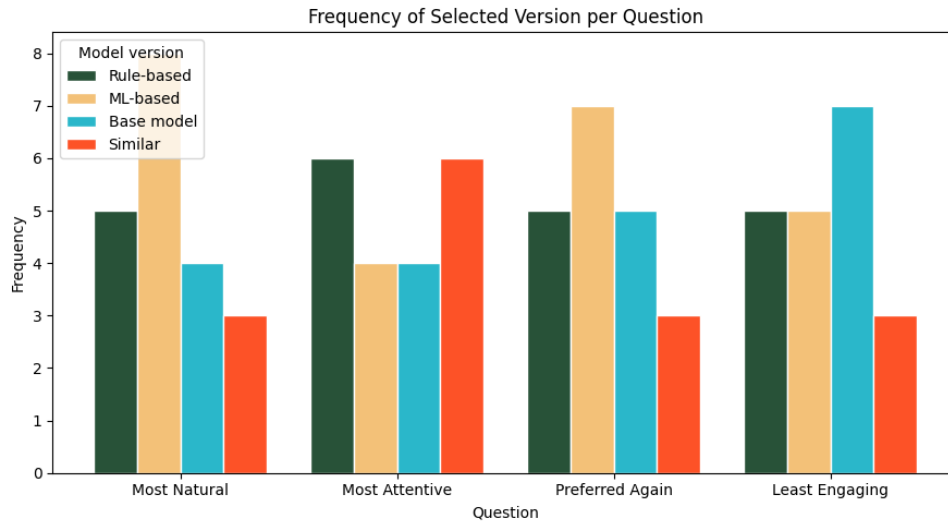


Figure 5.6: Frequencies of Questions from Comparative Ranking

5.1.7 Manipulation Check

To assess the explicit detectability of the independent versions, a post-experimental manipulation check was conducted, asking participants whether they noticed any behavioural variations between the three experimental blocks. Interestingly, 40% of the participants reported that they either did not notice any differences or were uncertain if a change had occurred.

A closer examination of the order of the versions during each session reveals an ordering effect: the vast majority of participants who failed the manipulation check were exposed to either the baseline or the rule-based model during their initial block. When users began their session with the rule-based model, its conservative, or poorly timed placements failed to cross the threshold of conscious awareness, causing participants to mentally group it with the baseline as a single, uniform experience. Participants that were initially exposed to the ML-based version did notice the difference, probably due to the larger contrast with the baseline version compared to the Rule-based version. Additionally, due to the already introduced backchannels in the Rule-based version, participants who tested the ML-based version afterwards did not perceive this contrast, which was probably more subtle between the Rule-based and ML-based version.

5.2 Objective Measures

In addition to the subjective psychometric evaluations, objective behavioural metrics were continuously logged throughout each experimental session. This system-level data was captured to evaluate the influence of the varying backchannel models on the participants' actual conversational behaviour.

Specifically, the system recorded the temporal dynamics of the dialogue, including user turn durations (the length of continuous speaker utterances) and inter-turn latency (the temporal gaps or pauses between conversational turns). By subjecting these objective interaction logs to statistical analysis, this study grounds its findings in measurable behavioural data.

5.2.1 Turn Duration

The first objective metric analysed was the mean turn duration, defined as the average length of continuous speech produced by the participant per conversational turn. It was hypothesized that the introduction of backchannels would correspond with extended user turn durations. A Friedman test conducted on the mean turn durations revealed no statistically significant main effect across the three conditions, $\chi^2(2) = 3.70, p = .157$. However, the analysis yielded a non-negligible Kendall's W effect size of $W = .0925$, indicating a distinct descriptive trend. Notably, the ML-based model elicited turn durations that were nearly 2 seconds longer, on average, than those observed in both the baseline and rule-based conditions (as shown in fig. 5.7). While inferential significance was not achieved, likely due to the high natural variance in human speaking styles within a small sample size, this directional trend is noteworthy. It shows objective behavioural trends suggesting that the backchannels subtly encouraged

participants to elaborate further. This objective increase in speaking time confirms the descriptive trends observed in the subjective engagement.

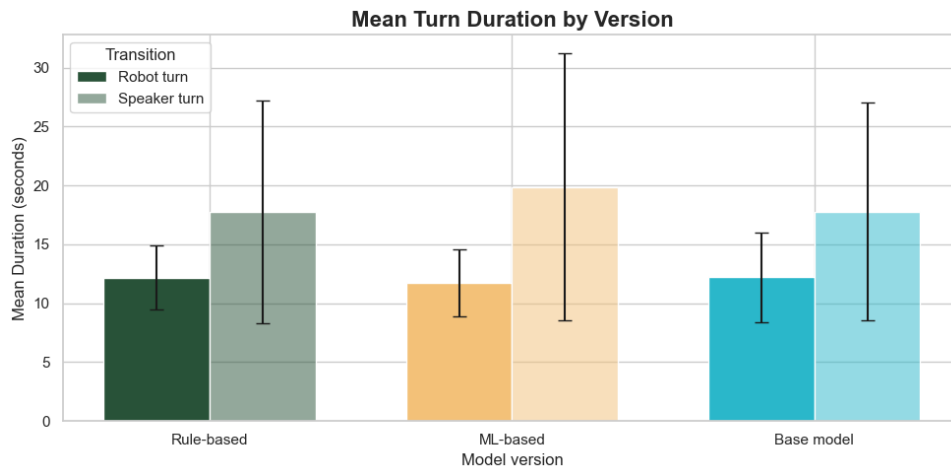


Figure 5.7: Mean Turn Durations for Robot and Human. Darker colour indicates robot turn, lighter colour indicates speaker turn.

5.2.2 Inter-turn latency

Beyond turn speaking durations, the analysis also evaluated human-to-robot inter-turn latency. This metric captures the temporal gap between the participant concluding their speech and the robot initiating its response. During the experimental trials, the system experienced network-induced processing delays. Therefore, it was verified whether this technical latency disproportionately affected any specific experimental condition, which could act as a confounding variable. A Friedman test conducted on the mean human-to-robot latency revealed no statistically significant differences across the three configurations, $\chi^2(2) = 0.70, p = .705$, with a negligible Kendall's W effect size of $W = .018$. As confirmed by the descriptive data in fig. 5.8, the response delays were nearly uniform. In the context of this study, this null result is desirable. It demonstrates systemic consistency, proving that the network latency was a constant factor rather than a variable one. Consequently, the relative scores between the baseline, rule-based, and ML-based models remain valid, as no single condition was unfairly penalized by disproportionate lag.

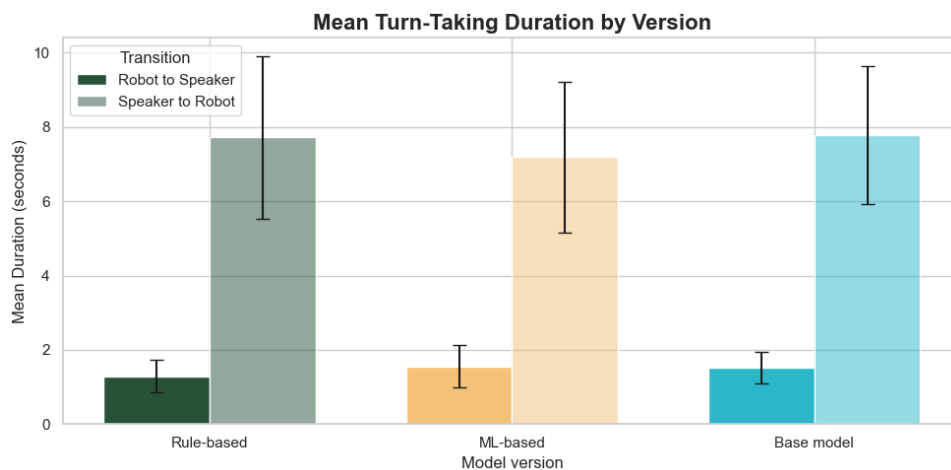


Figure 5.8: Mean Turn Taking Duration between Robot and Human. Darker colour indicates robot turn, lighter colour indicates speaker turn.

Discussion

Having detailed the descriptive and objective data, this section synthesizes these multi-methodological findings to directly address the study’s primary research questions and to evaluate the utility of conversational backchannels across four core constructs: Naturalness, Attentiveness, Engagement, and Social Presence. Specifically, this synthesis resolves the two foundational questions that guided our user study:

- **RQ1:** To what extent does the autonomous generation of backchannels by a social robot enhance perceived naturalness and user engagement?
- **RQ2:** Which algorithmic approach for timing prediction is perceived as most natural by human interlocutors?

By cross-referencing our subjective scales, objective system logs, and qualitative participant feedback, the following subsections try to find an explanation for the results found in chapter 5, providing a structured resolution to each research question.

6.1 Do Backchannels Enhance the Perceived Naturalness?

Research Question 1 (**RQ1**) sought to determine to what extent the inclusion of backchannels enhances the perceived naturalness and user engagement of human-robot dialogue compared to a baseline model. By cross-examining our four primary evaluation constructs, we can systematically unpack how the presence of listener feedback shifts the conversational dynamics.

6.1.1 Naturalness

The naturalness was measured through the MOS-style Naturalness item during the per-block questionnaires. Although the results did not have any kind of statistical significance, the descriptive hierarchy contradicted our initial hypotheses: predicting that backchanneling would enhance conversational authenticity. This unexpected outcome may be attributed to an expectation violation effect driven by user novelty. Post-session interviews revealed that participants simply did not expect a robotic platform to produce continuous backchannels.

P0: “I don’t personally think you would expect it from a robot.”

P12: “When the robot started saying ‘uh’, I didn’t expect the robot to do that. In the beginning, I thought the robot would start speaking, but he kept silent.”

In the absence of these cues, the interaction aligned better with the users’ existing expectations of a robot. When backchannels were introduced, their novelty may have created immediate cognitive friction. Even well-timed backchannels may have been perceived as unusual or distracting simply because they violated the baseline assumption of how a robot behaves, thereby depressing the naturalness scores for both active conditions. Several participants also noted that the frequency of the robot’s vocalizations felt “too much”, while others perceived the feedback as being generated at random intervals rather than in response to their narrative structure.

P1: “These small sounds Pepper made, sometimes came too often.”

P17: “The sounds pepper made seemed to come at quite random intervals.”

In human dialogue, backchannels are highly context-dependent [35, 50]. If a system injects them too frequently or misses the subtle turn-endings of the speaker, the user experiences a loss of interactional alignment. Consequently, rather than reinforcing the illusion of an attentive listener, the over-production of signals caused the robot’s behaviour to clash with the participant’s internal expectations of a natural conversational flow.

While the per-block naturalness ratings initially appeared counter-intuitive, the lack of statistical significance precludes any definitive claim of a meaningful decline in naturalness for the active backchannel models. Crucially, this immediate reactive data is contrasted by the end-of-session comparative ranking. When asked retrospectively which interaction felt “Most Natural”, the active backchanneling configurations received more votes over the baseline version. This divergence between the two evaluation methods points toward a temporal adaptation effect. The per-block Likert scales captured participants’ immediate, real-time responses to each condition. In the early stages of the experiment, the sudden introduction of robotic vocalizations might have triggered a high degree of novelty friction, as it violated their baseline expectations of machine behaviour. Conversely, the end-of-session ranking required participants to evaluate the entire experience. By the conclusion of the study, participants could already have grown accustomed to the robot’s vocal feedback. This suggests that while artificial backchannels may initially feel unusual to an unaccustomed user, familiarity breeds a retrospective appreciation for the conversational responsiveness they provide.

Furthermore, the high turn-taking latency provides context for the relatively low naturalness scores. While the latency was equal across conditions, its high absolute magnitude meant that participants frequently encountered an awkward, silent processing gap before the robot replied. This slow floor taking introduced a sense of artificiality into the dialogue, possibly suppressing the naturalness ceiling for the entire experiment regardless of the backchanneling model employed.

6.1.2 Attentiveness

The evaluation of perceived attentiveness achieved similar scores across all subjective metrics. This dimension was measured through three independent layers of data collection: an embedded item within the Coordination subscale, a standalone per-block rating item, and a retrospective ranking at the conclusion of the session. All three metrics yielded statistical insignificant results. Neither the immediate per-block evaluations nor the global inferential tests revealed a difference in how attentive the robot was perceived across the three conversational modes. This lack of variance is best illustrated by the retrospective comparative ranking data. When asked to select which configuration felt “Most Attentive”, the majority of participants selected the “No Preference” option. This behavioural data indicates that even with the benefit of direct cross-conditional comparison, users could not meaningfully distinguish the attentive capacity of the backchannel models from that of the baseline model.

A probable explanation for this uniform perception lies in the baseline multimodal behaviour of the robot. Even in the baseline condition, Pepper’s physical embodiment, characterized by continuous eye contact, subtle micro-movements, and responsive head tracking, may have already established a ceiling effect for perceived attentiveness.

Although, the results show no clear difference between the three models, one participant explained the following:

P19: “I do think that during the third conversation (Baseline) with Pepper, it felt like as if he did listen less to me.”

6.1.3 Engagement

While the explicit, single-item subjective metric did not reveal statistical significance, user engagement in socially situated dialogue systems can be operationalized and measured through alternative behavioural channels. Crucially, Inoue et al. established that user turn duration serves as a direct, reliable objective indicator of engagement during conversational interactions [22]. When evaluated through this behavioural lens, the results were statistically insignificant, though a subtle trend emerges that contradicts the relatively flat result of the subjective Likert scale. As demonstrated in our system logs, the ML-based backchannel model increased the mean user turn duration by nearly 2 seconds relative to the other

conditions. This divergence highlights a methodological construct mismatch: the explicit, self-reported question failed to capture the subtle shifts in interactional dynamics, whereas the objective temporal logs indicated a possible trend towards the ML-based condition.

Similarly, in the end-of-session comparative ranking a negative attribute prompt asking participants to identify which interaction felt the “Least Engaging” was included. Crucially, the distribution of responses revealed that fewer participants indicated the backchanneling models as the least engaging compared to the baseline. This retrospective pattern is aligned with the objective turn duration data extracted from our system. The fact that the baseline model was most frequently isolated as the least engaging behaviour could match the reality that participants seem to have a shorter interaction time when interacting with it.

6.1.4 Social Presence

While social presence was not directly measured after each block, this was measured through the PMQ scale after each session. Notably, the Human-Likeness subscale achieved the lowest numerical mean (3.21 ± 0.86) compared to the Competence-Dependability subscale (2.91 ± 0.7) and the Communicative Flexibility subscale (3.03 ± 0.82). While this might initially seem counter-intuitive, this result aligns with established HRI design principles, given that Pepper is explicitly engineered to be less anthropomorphic than highly realistic androids. By utilizing stylized, abstract human features, Pepper avoids the Uncanny Valley effect. This deliberate hardware design translated into the participants’ lower score for Human-Likeness. Additionally, due to the latency issues, some participants mentioned that it still felt like a robot mainly because of the long waiting times after every turn.

P4: “It feels the most like a robot because I have to wait for its response.”

6.1.5 Conclusion

Ultimately, the collective results reveal a consistent trend indicating that the active backchannel models could enhance perceived naturalness, user engagement, and social presence compared to a baseline model. While caution must be maintained, given that the global inferential tests did not yield any statistical significance within our participant, it is highly notable that nearly every descriptive result leans toward the active models. Rather than viewing these non-significant p -values as a nullification of the architecture, these directional patterns across both subjective surveys and objective system logs suggest that the backchanneling pipeline is actively moving the interaction in the intended direction. Consequently, larger sample based studies are needed to confirm this trends.

6.2 Which Timing Algorithm is Best?

While **RQ1** establishes whether backchannels are actually perceived as natural during a conversation, **RQ2** explicitly evaluates the algorithmic execution of those signals, investigating which timing determination framework is optimized for human-robot dialogue. By isolating the performance of the static, rule-based heuristic against the machine learning architecture across our multi-methodological datasets, we can try to get more insights.

6.2.1 Naturalness

While the presence of listener feedback broadly shifts conversational dynamics, a more nuanced distinction lies in the perceived naturalness between the two timing prediction models. When evaluated through independent, block-by-block psychometric measures, such as the MOS-style naturalness item, the differences between the two algorithmic approaches remained subtle, yielding similar descriptive profiles. However, when looking at the retrospective evaluation the difference is no longer subtle. When participants had to evaluate directly the configurations side-by-side at the conclusion of the session, a clear preference for the ML-based model emerged.

6.2.2 Engagement

A differentiation between the two algorithmic approaches manifests within the engagement construct. When operationalized as continuous user turn duration, a subtle divergence emerges: the ML-based architecture extended participant speech by nearly 2 seconds per conversational turn relative to the rule-based heuristic. In spoken dialogue systems, such a increase in utterance length could indicate interactional entrainment. The robot’s well-timed feedback prompted users to elaborate on their thoughts rather than closing their conversational turns earlier.

Also some participants explicitly remarked on this phenomenon, noting that the ML-based model felt more engaging.

P18: “But with the last one (ML), I felt it was a bit more pleasant, or that there was much more engagement.”

Rather than experiencing the robot as a passive listener or an algorithmic interrupter, the predictive neural network proved that precise, data-driven backchannel timing is a primary requirement for sustained human engagement.

6.2.3 Attentiveness

Although the Friedman test on the Attentiveness scale did not reach statistical significance, a descriptive pattern within the data is worth noting in the standalone Attentiveness and the single Attentiveness item in the Coordination subscale. As illustrated in figs. 5.2a and 5.3, the mean score for the ML-based model was a little higher than both the rule-based and the baseline condition.

While this directional trend must be interpreted with caution due to the lack of statistical significance, the upward trajectory provides an early indication that participants descriptively favoured the ML-based model, suggesting that the nuanced, variable timing generated by the neural network may align more closely with human conversational expectations than the predictable, rigid triggers of a rule-based system.

6.2.4 Social Presence

Because a dedicated Social Presence instrument was not included in the per-block questionnaires to minimize participant fatigue, this construct was evaluated retrospectively at the conclusion of the session. Specifically, the comparative ranking featured an item evaluating future interaction preference: “Which version of Pepper would you prefer to converse with again?” In human-robot interaction, this desire for re-engagement serves as an indication for social presence, capturing the degree to which a machine is perceived as a legitimate, socially interactive entity rather than a computational tool. While the resulting Chi-squared goodness-of-fit test did not reach statistical significance, which could be caused by the small sample size ($N = 20$), the descriptive voting distribution reveals a subtle directional trend. The ML-based architecture received the highest amount of positive selections compared to both the baseline and the rule-based version. When participants evaluated the robot across the entirety of the session, the organic variability and precise placement of the backchannels subtly elevated Pepper’s perceived social presence.

6.2.5 Conclusion

In summary, while the lack of strict inferential significance across several metrics precludes making definitive, hard claims, this study reveals a consistent descriptive trend toward the ML-based backchannel module. Across subjective rankings, qualitative participant feedback, and objective behavioural metrics, most notably the 2-second increase in user turn duration, this model consistently improved the rigid rule-based heuristic across most of the constructs.

Conclusion

7.1 Limitations

While this study offers valuable insights into automated listener feedback, several methodological and technical limitations must be acknowledged to contextualize the findings.

Network-Induced Inter-Turn Latency: The most prominent technical challenge encountered during the study was a high, network-induced inter-turn latency. This latency created an artificial, awkward gap during turn transitioning. Crucially, this delay was an artifact of the systemic processing pipeline and network infrastructure, completely independent of the real-time backchannel generation models, which were generated asynchronously and locally on the machine. However, its presence introduced a sense of artificiality that lowered the naturalness scores across all conditions. Mitigating this bottleneck through securing a reliable network connection is essential for further studies.

User Personality Profiles The conversational architecture generated backchannels based on generalized backchannel generation models without personalizing the feedback to individual user profiles. In human conversations, the optimal frequency of listener feedback is highly dependent on individuals [8]. Because the system did not adapt to the context, several participants reported an over-use of backchannels during the interaction. The model’s inability to dynamically adapt its behavioural backchannel intensity to match the user’s specific conversational preference highlights a structural limitation in this study.

Language Generalization Constraints Finally, the experimental evaluation was conducted exclusively with native Dutch (*L1*) speakers, while the conversations could be performed in both Dutch and English. Backchanneling is recognized as a deeply culture-specific and language-bound conversational phenomenon [50]. Realizations of listener feedback vary drastically in frequency and placement across different languages. Consequently, the descriptive trends and behavioural dynamics observed in this Dutch study cannot be generalized globally.

7.2 Future works

The engineering and evaluation of robotic backchannels remains an evolving domain in social robotics. While this study successfully designed and validated a data-driven model predicting backchannel timing and basic type classification, several multi-modal and contextual dimensions remain open for exploration. Future iterations must expand the system’s architecture.

Contextual Integration Future architectures should focus on integrating conversational context and user-specific constraints into the generation pipeline.

A promising framework for this is demonstrated by Park et al., who utilized deep context-aware representations alongside personalized user embeddings to synthesize highly tailored listener behaviour [35]. By embedding individual speaker preferences, the robot can transition away from generic vocalizations toward personalized listener responses, directly eliminating the acoustic overuse reported by several participants in this study.

Lexical-Prosodic Alignment Furthermore, future synthesis engines must bridge the gap between the lexical form of a backchannel and its specific prosodic execution. In natural human conversations, identical lexical words communicate different meanings based entirely on their intonation (e.g. “Yeah!”, signalling enthusiastic agreement versus “Yeah?” signalling surprisal or non-understanding).

To achieve this level of communicative nuance, future systems should look toward dual-modality alignment architectures, such as the contrastive large language model framework proposed by Qian et al. [40]. By systematically mapping joint representations of linguistic context and audio features, subsequent models can dynamically control the pitch contours and duration of generated responses, ensuring that the robot’s vocal feedback matches the immediate emotional and narrative state of the human speaker.

Visual Backchanneling Finally, qualitative insights from our semi-structured interviews suggest a divergence in modality acceptance. While the reproduced backchannels sometimes felt weird due to misplacements, some participants reported that Pepper’s physical head-nodding added clear interactive value.

Isolating purely visual backchannels by completely removing text-based or auditory vocalizations, could possibly give even better human-like results [36]. Given that verbal signals carry heavy contextual and linguistic friction, deploying a gestural model could maximize user engagement through continuous bodily alignment while entirely bypassing the friction, language constraints and uncanny-valley acoustics associated with artificial vocal tracking.

7.3 Conclusion

In summary, this study successfully designed, implemented, and validated a fully operational end-to-end conversational pipeline capable of predicting both the optimal temporal points and semantic types for robotic backchannel reproduction. Through a within-subjects user study, this architecture was tested between three different versions: a rule-based heuristic approach, a ML approach and a baseline model.

While individual inferential metrics did not give any statistical significant results, possibly due to the relatively small sample size, consistent patterns emerged across the datasets. The integration of active backchanneling consistently had a better performance compared to the non-backchannel baseline across three core constructs: naturalness, user engagement, and perceived social presence.

Furthermore, the algorithmic comparison revealed a subtle preference for the ML-based architecture over the rule-based model. This conclusively validates the utility of neural-timed predictive modelling over rigid heuristics, offering a robust structural foundation for the future development of highly fluid, engaging, and human-centric conversational agents.

Societal Reflections

Social isolation and loneliness among older adults represent an increasing global public health challenge. According to recent directives from the World Health Organization, approximately one in ten elderly individuals experiences loneliness, while one in four is fundamentally socially isolated [1]. Consequently, this isolation results in problems that extend far beyond subjective emotional distress. Loneliness operates as a critical determinant of health, carrying severe impacts on overall mortality. It actively exacerbates chronic physical conditions, such as cardiovascular disease and type 2 diabetes, while severely undermining mental well-being through heightened rates of depression and anxiety.

To address this societal gap, this thesis explores the potential of socially assistive robotics to mitigate isolation through natural, fluid interaction. Rather than attempting to replace human-to-human companionship, social robots are engineered to play a vital, complementary role in an individual's care ecosystem. By building sustained social relationships, these agents can actively stimulate a sense of connection and reduce feelings of loneliness among the elderly [38].

The primary barrier to widespread user acceptance, however, lies in the communicative artificiality of current human-robot dialogue systems. This research directly tackles that friction by implementing automated backchannel behaviour. By narrowing the communicative gap between machine interaction and natural human discourse, this pipeline enhances the robot's perceived humanity. Reducing this cognitive and conversational barrier makes the interaction feel fundamentally safer and more authentic, thereby encouraging older adults to engage in a conversation.

While this study demonstrates promising results for utilizing automated backchannels to enhance conversational alignment and perceived humanity, deploying social robots within real-world environments introduces fundamental technical and ethical risks. Foremost among these challenges is the critical need of user data privacy and confidentiality.

Under the system's current architecture, the complete conversational pipeline, including raw acoustic audio input, intermediate textual representations, and long-term dialogue memory, must be transmitted to external cloud-based servers for processing. This remote pipeline is a logistical necessity to leverage the computational power of state-of-the-art machine learning models. However, it introduces significant vulnerabilities regarding data sovereignty and user privacy.

When deploying these systems within sensitive, domestic, or clinical environments, this cloud dependency represents an architectural trade-off. To ethically bridge this gap, future deployments must operate under strict frameworks of informed consent and continuous transparency. Users, particularly vulnerable older adults, must be explicitly aware of the underlying data-routing infrastructure and the privacy risks associated with cloud-assisted dialogue processing.

Ultimately, to transition this research into a scalable, trustworthy care solution, future engineering iterations must prioritize the development of localized or edge-computing architectures. Processing speech recognition, timing predictions and conversational tracking locally on the robot's hardware will mitigate external data vulnerabilities, improving the conversational naturalness does not come at the cost of personal privacy.

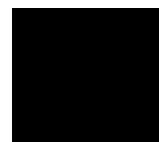
Bibliography

- [1] June 2025. URL: <https://www.who.int/publications/i/item/978240112360>.
- [2] Lukasik A. and A. Gut. “rom robots to chatbots: unveiling the dynamics of human-AI interaction.” In: (2025). URL: <https://doi.org/10.3389/fpsyg.2025.1569277>.
- [3] Amalia Istiqlali Adiba, Takeshi Homma, and Toshinori Miyoshi. “Towards Immediate Backchannel Generation Using Attention-Based Early Prediction Model”. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. June 2021, pp. 7408–7412. DOI: 10.1109/ICASSP39728.2021.9414193. URL: <https://ieeexplore.ieee.org/document/9414193> (visited on 11/13/2025).
- [4] Kyle Arnold. “Humming Along”. In: *Contemporary Psychoanalysis* 48.1 (2012), pp. 100–117. DOI: 10.1080/00107530.2012.10746491. eprint: <https://doi.org/10.1080/00107530.2012.10746491>. URL: <https://doi.org/10.1080/00107530.2012.10746491>.
- [5] Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. “Modeling Feedback in Interaction With Conversational Agents—A Review”. In: *Frontiers in Computer Science* Volume 4 - 2022 (2022). ISSN: 2624-9898. DOI: 10.3389/fcomp.2022.744574. URL: <https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2022.744574>.
- [6] Stefan Benus, Agustín Gravano, and Julia Hirschberg. “THE PROSODY OF BACKCHANNELS IN AMERICAN ENGLISH”. en. In: (2007).
- [7] Philippe Blache et al. “An integrated model for predicting backchannel feedbacks”. en. In: *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. Virtual Event Scotland UK: ACM, Oct. 2020, pp. 1–3. ISBN: 978-1-4503-7586-3. DOI: 10.1145/3383652.3423948. URL: <https://dl.acm.org/doi/10.1145/3383652.3423948> (visited on 04/21/2026).
- [8] Peter Blomsma et al. “Backchannel behavior is idiosyncratic”. en. In: *Language and Cognition* 16.4 (Dec. 2024), pp. 1158–1181. ISSN: 1866-9808, 1866-9859. DOI: 10.1017/langcog.2024.1. URL: <https://www.cambridge.org/core/journals/language-and-cognition/article/backchannel-behavior-is-idiosyncratic/F75D0AEEAF258399166A58E3DDCC7D7E> (visited on 01/19/2026).
- [9] Shiye Cao et al. *Interruption Handling for Conversational Robots*. arXiv:2501.01568 [cs]. Apr. 2025. DOI: 10.48550/arXiv.2501.01568. URL: <http://arxiv.org/abs/2501.01568> (visited on 11/04/2025).
- [10] Junjie Chen et al. *FireRedChat: A Pluggable, Full-Duplex Voice Interaction System with Cascaded and Semi-Cascaded Implementations*. arXiv:2509.06502 [cs]version: 1. Sept. 2025. DOI: 10.48550/arXiv.2509.06502. URL: <http://arxiv.org/abs/2509.06502> (visited on 10/20/2025).
- [11] Yong-Seok Choi, Jeong-Uk Bang, and Seung Hi Kim. “Joint streaming model for backchannel prediction and automatic speech recognition”. en. In: *ETRI Journal* 46.1 (2024). eprint: <https://onlinelibrary.wiley.com/doi/abs/10.4218/etrij.2023-0358>. DOI: 10.4218/etrij.2023-0358. URL: <https://onlinelibrary.wiley.com/doi/abs/10.4218/etrij.2023-0358> (visited on 01/19/2026).
- [12] Ronald Cumbal et al. “Let Me Finish First - The Effect of Interruption-Handling Strategy on the Perceived Personality of a Social Agent”. In: *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. IVA ’24. New York, NY, USA: Association for Computing Machinery, Dec. 2024, pp. 1–10. ISBN: 979-8-4007-0625-7. DOI: 10.1145/3652988.3673916. URL: <https://dl.acm.org/doi/10.1145/3652988.3673916> (visited on 11/04/2025).
- [13] Philip R. Doyle et al. “The Partner Modelling Questionnaire: A Validated Self-Report Measure of Perceptions toward Machines as Dialogue Partners”. In: *ACM Trans. Comput.-Hum. Interact.* 32.4 (Aug. 2025), 39:1–39:33. ISSN: 1073-0516. DOI: 10.1145/3729170. URL: <https://dl.acm.org/doi/10.1145/3729170> (visited on 05/14/2026).

- [14] P. Drew and J. Heritage. *Conversation Analysis: Turn design and action formation*. Conversation Analysis. Sage, 2006. ISBN: 978-1-4129-1848-0. URL: <https://books.google.be/books?id=tEicAQAAIAAJ>.
- [15] Erik Ekstedt and Gabriel Skantze. “TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. arXiv:2010.10874 [cs]. 2020, pp. 2981–2990. DOI: 10.18653/v1/2020.findings-emnlp.268. URL: <http://arxiv.org/abs/2010.10874> (visited on 11/04/2025).
- [16] Erik Ekstedt and Gabriel Skantze. *Voice Activity Projection: Self-supervised Learning of Turn-taking Events*. 2022. arXiv: 2205.09812 [eess.AS]. URL: <https://arxiv.org/abs/2205.09812>.
- [17] Olov Engwall, Ronald Cumbal, and Ali Reza Majlesi. “Socio-cultural perception of robot backchannels”. In: *Frontiers in Robotics and AI* 10 (Jan. 2023), p. 988042. ISSN: 2296-9144. DOI: 10.3389/frobt.2023.988042. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9909394/> (visited on 01/19/2026).
- [18] Yoshinori Fukunaga et al. “Backchannel prediction for natural spoken dialog systems using general speaker and listener information”. en. In: *Interspeech 2025*. ISCA, Aug. 2025, pp. 1078–1082. DOI: 10.21437/Interspeech.2025-1316. URL: https://www.isca-archive.org/interspeech_2025/fukunaga25_interspeech.html (visited on 02/17/2026).
- [19] Kushal Garg, Lincy Meera Mathews, and Prashanth Kambli. “InterConv: Conversational Agent That Handles Interruptions Like Humans”. In: *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*. Vol. 3. Mar. 2025, pp. 1–6. DOI: 10.1109/IATMSI64286.2025.10985683. URL: <https://ieeexplore.ieee.org/abstract/document/10985683> (visited on 11/04/2025).
- [20] M. Grimm and K. Kroschel. *Robust Speech: Recognition and Understanding*. IntechOpen, 2007. ISBN: 978-3-902613-08-0. URL: <https://books.google.be/books?id=OuSgDwAAQBAJ>.
- [21] Koji Inoue et al. *Prompt-Guided Turn-Taking Prediction*. 2025. arXiv: 2506.21191 [cs.CL]. URL: <https://arxiv.org/abs/2506.21191>.
- [22] Koji Inoue et al. “Towards Objective Evaluation of Socially-Situated Conversational Robots: Assessing Human-Likeness through Multimodal User Behaviors”. In: *International Conference on Multimodal Interaction*. ICMI ’23. ACM, Oct. 2023, pp. 86–90. DOI: 10.1145/3610661.3617151. URL: <http://dx.doi.org/10.1145/3610661.3617151>.
- [23] Koji Inoue et al. *Yeah, Un, Oh: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection*. arXiv:2410.15929 [cs]. Feb. 2025. DOI: 10.48550/arXiv.2410.15929. URL: <http://arxiv.org/abs/2410.15929> (visited on 11/04/2025).
- [24] Divesh Lala et al. “Attentive listening system with backchanneling, response generation and flexible turn-taking”. In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Ed. by Kristiina Jokinen et al. Saarbrücken, Germany: Association for Computational Linguistics, Aug. 2017, pp. 127–136. DOI: 10.18653/v1/W17-5516. URL: <https://aclanthology.org/W17-5516/> (visited on 01/20/2026).
- [25] Borui Liao et al. *FlexDuo: A Pluggable System for Enabling Full-Duplex Capabilities in Speech Dialogue Systems*. arXiv:2502.13472 [cs] version: 2. May 2025. DOI: 10.48550/arXiv.2502.13472. URL: <http://arxiv.org/abs/2502.13472> (visited on 10/21/2025).
- [26] Ting-Han Lin et al. “Connection-Coordination Rapport (CCR) Scale: A Dual-Factor Scale to Measure Human-Robot Rapport”. In: *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*. HRI ’25. Melbourne, Australia: IEEE Press, Mar. 2025, pp. 869–879. URL: <https://dl.acm.org/doi/10.5555/3721488.3721594> (visited on 05/14/2026).
- [27] Yuxin Lin et al. *Predicting Turn-Taking and Backchannel in Human-Machine Conversations Using Linguistic, Acoustic, and Visual Signals*. en. arXiv:2505.12654 [cs]. May 2025. DOI: 10.48550/arXiv.2505.12654. URL: <http://arxiv.org/abs/2505.12654> (visited on 11/02/2025).
- [28] Chao Liu et al. “Toward Enabling Natural Conversation with Older Adults via the Design of LLM-Powered Voice Agents that Support Interruptions and Backchannels”. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI ’25. New York, NY, USA: Association for Computing Machinery, Apr. 2025, pp. 1–22. ISBN: 979-8-4007-1394-1. DOI: 10.1145/3706598.3714228. URL: <https://dl.acm.org/doi/10.1145/3706598.3714228> (visited on 10/14/2025).

- [29] Long Mai and Julie Carson-Berndsen. *Real-Time Textless Dialogue Generation*. arXiv:2501.04877 [cs]. Jan. 2025. DOI: 10.48550/arXiv.2501.04877. URL: <http://arxiv.org/abs/2501.04877> (visited on 10/27/2025).
- [30] Arunima Maitra, Dorothea French, and Katharina von der Wense. “Dialogue Acts as a Lens on Human–LLM Interaction: Analyzing Conversational Norms in Model-Generated Responses”. In: *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*. Ed. by Su Lin Blodgett et al. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 317–325. ISBN: 979-8-89176-353-1. DOI: 10.18653/v1/2025.hcinlp-1.25. URL: <https://aclanthology.org/2025.hcinlp-1.25/> (visited on 12/15/2025).
- [31] *Methods for subjective determination of transmission quality*. Recommendations. Geneva: IUT, 2000. URL: <https://cds.cern.ch/record/789314>.
- [32] H. Moravec. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, 1988. ISBN: 9780674576186. URL: <https://books.google.be/books?id=56mb7XuSx3QC>.
- [33] Openai. *Voice activity detection (VAD) - OpenAI API*. nl-NL. URL: <https://platform.openai.com> (visited on 11/04/2025).
- [34] Michael Paierl, Martin Hagmüller, and Barbara Schuppler. “Continuous prediction of backchannel timing for human-robot interaction”. en. In: *Interspeech 2025*. ISCA, Aug. 2025, pp. 3020–3024. DOI: 10.21437/Interspeech.2025-770. URL: https://www.isca-archive.org/interspeech_2025/paierl25_interspeech.html (visited on 11/04/2025).
- [35] Yo-Han Park et al. “Backchannel prediction, based on who, when and what”. en. In: *Interspeech 2024*. ISCA, Sept. 2024, pp. 3570–3574. DOI: 10.21437/Interspeech.2024-2523. URL: https://www.isca-archive.org/interspeech_2024/park24b_interspeech.html (visited on 01/19/2026).
- [36] Yoonvin Park, Doha Kim, and Hayeon Song. “I Know You’re Listening: Designing Visual Backchannels for Voice User Interfaces”. In: *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*. IUI ’25 Companion. New York, NY, USA: Association for Computing Machinery, Mar. 2025, pp. 56–59. ISBN: 979-8-4007-1409-2. DOI: 10.1145/3708557.3716343. URL: <https://dl.acm.org/doi/10.1145/3708557.3716343> (visited on 11/04/2025).
- [37] Maria J. Pinto and Tony Belpaeme. “Predictive Turn-Taking: Leveraging Language Models to Anticipate Turn Transitions in Human-Robot Dialogue”. In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. Aug. 2024, pp. 1733–1738. DOI: 10.1109/RO-MAN60168.2024.10731379. URL: <https://ieeexplore.ieee.org/document/10731379/> (visited on 10/21/2025).
- [38] M J Pinto-Bernal, K Boels, and T Belpaeme. “Fostering Social Connection and Emotional Engagement to Mitigate Social Isolation in Care Homes Using Empathetic Human-Robot Conversation”. en. In: ().
- [39] Ronald Poppe et al. “Backchannel Strategies for Artificial Listeners”. en. In: *Intelligent Virtual Agents*. Ed. by Jan Allbeck et al. Vol. 6356. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 146–158. ISBN: 978-3-642-15891-9. DOI: 10.1007/978-3-642-15892-6_16. URL: http://link.springer.com/10.1007/978-3-642-15892-6_16 (visited on 01/19/2026).
- [40] Livia Qian and Gabriel Skantze. *Aligning Backchannel and Dialogue Context Representations via Contrastive LLM Fine-Tuning*. arXiv:2604.16622 [cs]. Apr. 2026. DOI: 10.48550/arXiv.2604.16622. URL: <http://arxiv.org/abs/2604.16622> (visited on 04/27/2026).
- [41] Nikhil R. *Mastering Turn Detection and Interruption Handling in Voice AI Applications*. en. Oct. 2025. URL: <https://comparevoiceai.com> (visited on 10/21/2025).
- [42] Andrew Reece et al. “The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation”. en. In: *Science Advances* 9.13 (Mar. 2023), eadf3197. ISSN: 2375-2548. DOI: 10.1126/sciadv.adf3197. URL: <https://www.science.org/doi/10.1126/sciadv.adf3197> (visited on 02/13/2026).
- [43] Rajarshi Roy et al. “PERSONAPLEX: VOICE AND ROLE CONTROL FOR FULL DUPLEX CONVERSATIONAL SPEECH MODELS”. en. In: (2026).
- [44] Sam O’Connor Russell and Naomi Harte. *Visual Cues Enhance Predictive Turn-Taking for Two-Party Human Interaction*. Version Number: 2. 2025. DOI: 10.48550/ARXIV.2505.21043. URL: <https://arxiv.org/abs/2505.21043> (visited on 11/01/2025).

- [45] Laura Saad et al. “Choosing the “Perfect” Scale: A Primer to Evaluate Existing Scales in HRI”. In: *J. Hum.-Robot Interact.* 15.2 (Jan. 2026), 42:1–42:30. DOI: 10.1145/3772066. URL: <https://dl.acm.org/doi/10.1145/3772066> (visited on 05/14/2026).
- [46] Mohan Shi et al. *Semantic VAD: Low-Latency Voice Activity Detection for Speech Interaction*. arXiv:2305.12450 [eess]. May 2023. DOI: 10.48550/arXiv.2305.12450. URL: <http://arxiv.org/abs/2305.12450> (visited on 10/21/2025).
- [47] Bandhav Veluri et al. “Beyond Turn-Based Interfaces: Synchronous LLMs as Full-Duplex Dialogue Agents”. en. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 21390–21402. DOI: 10.18653/v1/2024.emnlp-main.1192. URL: <https://aclanthology.org/2024.emnlp-main.1192> (visited on 10/21/2025).
- [48] Thi Hoang Mai Vu and Thi Bich Ngan Tran. “Backchannel Across Cultures: A Review and Implications”. en. In: *VietTESOL International Convention Proceedings 1* (2021). URL: <https://proceedings.viettesol.org.vn/index.php/vic/article/view/189> (visited on 01/19/2026).
- [49] Jinhan Wang et al. *Turn-taking and Backchannel Prediction with Acoustic and Large Language Model Fusion*. arXiv:2401.14717 [cs]. Jan. 2024. DOI: 10.48550/arXiv.2401.14717. URL: <http://arxiv.org/abs/2401.14717> (visited on 11/15/2025).
- [50] Nigel Ward. *Backchannel Facts*. May 2017. URL: <https://www.cs.utep.edu/nigel/bc/> (visited on 01/19/2026).
- [51] Nigel Ward and Wataru Tsukahara. “Prosodic features which cue back-channel responses in English and Japanese”. In: *Journal of Pragmatics* 32.8 (2000), pp. 1177–1207. ISSN: 0378-2166. DOI: [https://doi.org/10.1016/S0378-2166\(99\)00109-5](https://doi.org/10.1016/S0378-2166(99)00109-5). URL: <https://www.sciencedirect.com/science/article/pii/S0378216699001095>.
- [52] Victor H. Yngve. “ON GETTING A WORD IN EDGEWISE”. In: *The Sixth Regional Meeting [of the] Chicago Linguistic Society*. 6th ed. 1970, pp. 567–578.
- [53] Aohan Zeng et al. *GLM-4-Voice: Towards Intelligent and Human-Like End-to-End Spoken Chatbot*. en. Dec. 2024. URL: <https://arxiv.org/abs/2412.02612v1> (visited on 10/25/2025).
- [54] Dong Zhang et al. *SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities*. arXiv:2305.11000 [cs]. May 2023. DOI: 10.48550/arXiv.2305.11000. URL: <http://arxiv.org/abs/2305.11000> (visited on 10/26/2025).
- [55] Hao Zhang et al. *LLM-Enhanced Dialogue Management for Full-Duplex Spoken Dialogue Systems*. en. arXiv:2502.14145 [cs]. Feb. 2025. DOI: 10.48550/arXiv.2502.14145. URL: <http://arxiv.org/abs/2502.14145> (visited on 10/20/2025).



Item-by-Item Mappings

This part of the appendix shows every item that appears in the questionnaire bundle, the original CCR phrasing it derives from, the full-sentence version we administer and what aspect of the construct each item captures.

CCR Coordination subscale (6 items)

Original CCR item	Full-sentence wording	What it captures
Coordination	Pepper and I coordinated well during the conversation.	Joint timing and rhythm
Focus	Pepper stayed focused on the conversation.	Topical attentiveness
Attentiveness	Pepper was attentive to what I was saying.	Listening behaviour
Smooth flow	The conversation flowed smoothly.	Absence of awkward gaps
Equal participation	Pepper and I participated equally in the conversation.	Turn balance
Engagement	Pepper seemed engaged in the conversation.	Active investment

CCR Connection subscale (12 items)

Original CCR item	Full-sentence wording	What it captures
Mutual interest	Pepper and I were both interested in the conversation.	Shared engagement
Friendliness	Pepper came across as friendly.	Warmth
Liking	I liked talking with Pepper.	Hedonic positivity
Closeness	I felt close to Pepper during the conversation.	Interpersonal closeness
Mutual understanding	Pepper and I understood each other.	Shared meaning
Acceptance	I felt accepted by Pepper.	Non-judgemental tone
Personal connection	I felt a personal connection with Pepper.	Affective bond
Shared perspective	Pepper and I seemed to share a perspective.	Cognitive alignment
Bond	I felt a sense of bond with Pepper.	Relational bond
Genuineness	Pepper came across as genuine.	Authentic presentation
Comfort	I felt comfortable talking with Pepper.	Relaxation
Caring	Pepper seemed to care about what I was saying.	Other-orientation

Stand-alone attentiveness item

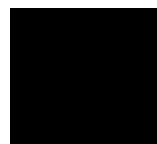
Item wording	What it captures
Pepper felt attentive while I was speaking.	Single-item attentiveness measure for direct cross-condition comparison.

Naturalness item

Item wording	What it captures
How natural did this conversation feel?	MOS-style global naturalness judgement on a 5-point scale (Very unnatural → Very natural). Field standard in speech and dialogue research

PMQ Human-Likeness sub-dimension (6 semantic-differential pairs)

Left anchor	Right anchor	What it captures
Human-like	Machine-like	Global anthropomorphism judgement.
Life-like	Tool-like	Animacy / presence as an entity.
Warm	Cold	Affective warmth, the dimension most likely affected by good backchannels.
Empathetic	Apathetic	Perceived emotional responsiveness.
Personal	Generic	Whether the robot felt addressed-to-me vs. a generic chatbot.
Social	Transactional	Whether the interaction felt like socialising vs. a transaction.



Use of Generative Artificial Intelligence

Throughout the development of this thesis, Generative Artificial Intelligence (AI) assistance was selectively utilized to support and optimize specific phases of the research workflow. In accordance with institutional guidelines for academic integrity, a human-in-the-loop validation process was maintained. All AI-generated outputs were critically evaluated, fact-checked, and cross-referenced against primary sources to mitigate the risk of computational hallucinations or factual inaccuracies. More specific, AI models were used across the following stages:

Software Development and Engineering: AI-assisted development tools were used as coding agents to support the implementation, debugging, and optimization of the end-to-end conversational pipeline for the Pepper robot.

Statistical and Methodological Consultation: AI was used as an analytical aid during the experimental design phase to explore potential psychometric scales, cross-validate statistical assumptions, and determine appropriate testing frameworks (e.g., the Friedman test).

Writing Refinement: AI tools functioned as editorial assistants, formally restructuring syntax, correcting grammatical variances, and elevating the academic tone of the manuscript to ensure maximum reading clarity. Every rewriting was re-evaluated to remove incorrect parts or reformulate specific sentences.

Literature Synthesis: AI text summarization was leveraged to generate rapid, high-level structural overviews of complex academic literature, which were subsequently cross-verified through manual reading of the primary papers. Additionally, the use of embeddings were used to create a semantic search engine for all the listed references, to easily find relative sources.

The final interpretation of the experimental datasets, the architectural decisions, and the intellectual ownership of the narrative remain entirely the work of the author.