Using Respiration for Enhancing Human-Robot Dialogue

Takao Obi and **Kotaro Funakoshi**
Tokyo Institute of Technology, Tokyo, Japan
{smalltail, fukanoshi}@lr.pi.titech.ac.jp

Abstract

This paper presents the development and capabilities of a spoken dialogue robot that uses respiration to enhance human-robot dialogue. By employing a respiratory estimation technique that uses video input, the dialogue robot captures user respiratory information during dialogue. This information is then used to prevent speech collisions between the user and the robot and to present synchronized pseudo-respiration with the user, thereby enhancing the smoothness and engagement of human-robot dialogue.

1 Introduction

For spoken dialogue robots to be effectively used in various scenarios, it is crucial for human-robot dialogue to be as natural as human-human dialogue. In human-human dialogue, communication occurs not only through verbal language but also through non-verbal cues. Thus, incorporating non-verbal information is essential for enhancing the naturalness of human-robot dialogue. Previous research has shown that integrating non-verbal cues such as nodding and body movements into robots can improve dialogue fluency (Watanabe et al., 2002), confirming the benefits of including non-verbal information in dialogue robots.

Our focus is on a specific type of non-verbal information: respiration. We believe that integrating respiratory information can significantly enhance human-robot dialogue, as respiration is intimately connected to speech. Our research has demonstrated that respiratory information is effective in predicting user speech onset (Obi and Funakoshi, 2023). Based on this finding, we developed a spoken dialogue system that uses user respiratory information to predict user speech onset, helping to prevent speech collisions in human-robot dialogue (Obi and Funakoshi, 2024). Furthermore, to ensure accurate capture of user respiratory information in real-world dialogue settings, we have implemented a respiratory estimation method that uses video input. This method employs the first-ever deep learning model to provide a robust estimation of the respiratory waveform, even in the presence of speech movements, marking an improvement over an existing method (Obi and Funakoshi, 2023).

Building on these developments, we created a spoken dialogue robot that uses respiration. The dialogue robot estimates user respiratory waveform values and uses them for enhancing human-robot dialogue. Initially, the dialogue robot predicts user speech onset and initiates dialogue responses only when user utterances are not predicted, thereby preventing speech collisions. This approach is designed to facilitate smoother human-robot dialogue by ensuring that the conversation flows without interruptions. A previous study in human-robot dialogue has confirmed that avoiding speech collisions contributes to smoother turn-taking (Funakoshi et al., 2008). Secondly, the dialogue robot presents synchronized pseudo-respiration with user respiration. This approach is designed to enhance the dialogue robot's impression. Research in robots has shown that the presentation of pseudo-respiration can enhance the robot's impression (Terzioğlu et al., 2020). Furthermore, a study in human-human dialogue has demonstrated that synchronized respiration during turn-taking leads to smoother transitions between speakers (Rochet-Capellan and Fuchs, 2014).

To confirm the impact of these approaches on human-robot dialogue, we conducted a dialogue experiment in which 50 participants each interacted individually with the dialogue robot. We used actual respiratory waveform values obtained from a respiratory measurement device as user respiration (Obi and Funakoshi, 2024). Preliminary analysis, conducted after the initial report, indicates that adjusting the timing of robot speech onset using user speech prediction effectively reduces speech

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 325–328
September 18–20, 2024.

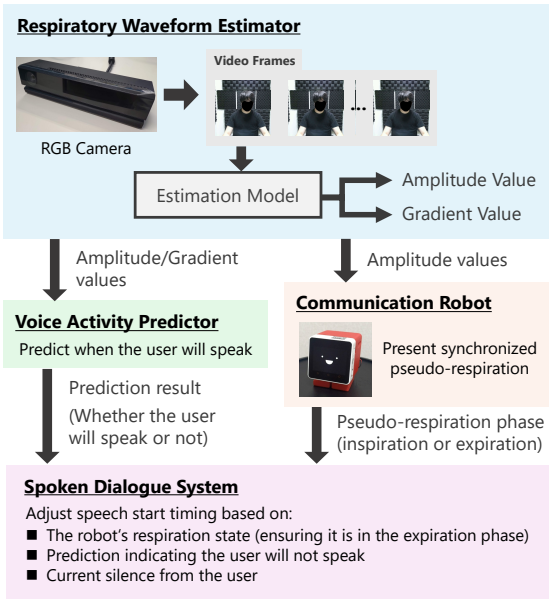

Figure 1: Overview of spoken dialogue robot using respiration

collisions. Additionally, initial user impression evaluations suggest that both the robot speech adjustment and synchronized pseudo-respiration presentation make users feel that their speech is less likely to overlap with the robot's responses. These findings suggest that when the respiratory waveform estimation operates ideally, it can facilitate smooth human-robot dialogue.

2 System Overview

Our dialogue robot comprises a respiratory waveform estimator, a voice activity predictor, a communication robot, and a spoken dialogue system. Figure 1 provides an overview of these components and their arrangement.

Respiratory Waveform Estimator: We developed a respiratory waveform estimator using a deep-learning model comprising 3DCNN-ConvLSTM, which is robust against speaker motion (Obi and Funakoshi, 2023). This estimator uses RGB video frames of a user as input to estimate the user's respiratory waveform amplitude and gradient at the time of the final frame. The model was trained using VRWiDataset [1]. The estimated values are then transmitted to both the voice activity predictor and the communication robot.

Voice Activity Predictor: We developed a voice activity predictor using a single-layer Long Short-Term Memory (LSTM) network that processes estimated respiratory waveform values as input. This model predicts whether user speech will occur within the next 200 ms during non-speaking periods. It was trained on a dataset created using the VRWiDataset, which extracts data from user non-speaking intervals. This dataset pairs the respiratory waveform values over a specific period with the user's voice activity occurring 200 ms later. The prediction results are then transmitted to the spoken dialogue system.

Communication Robot: We use an open-source robot named stack-chan[2] for the communication robot. The communication robot performs a pseudo-respiratory movement, represented by its vertical motion. The movement is based on the user's respiratory waveform amplitude values obtained from the respiratory waveform estimator. The communication robot uses these values to synchronize the timing of its inspiration and expiration with the user. Additionally, the communication robot's inspiration/expiration phase information is transmitted to the spoken dialogue system to determine the speech timing.

Spoken Dialogue System: We developed a spoken dialogue system facilitating dialogues on arbitrary topics. This system uses GPT-4 Turbo[3] for the generation of dialogue responses. For speech processing, it employs both Google Cloud speech-to-text[4] and say command in macOS. The system initiates responses when the voice activity predictor confirms no imminent user speech onset, and only during the communication robot's expiration phases. Additionally, it is designed not to respond during user speaking, ensuring that there are no response overlaps between the user and the dialogue robot in dialogue. If no user speech is detected, it will autonomously initiate responses continuously to maintain dialogue. The intervals between the continuous responses are set randomly between 0.5 and 3.5 seconds to simulate a realistic dialogue pace.

3 Use Case

Our dialogue robot is designed to demonstrate the effectiveness of integrating respiratory information into human-robot dialogue.

[1] https://github.com/fnkslab/VRWiDataset

[2] https://github.com/meganetaaan/stack-chan

[3] https://openai.com/gpt-4

[4] https://cloud.google.com/speech-to-text

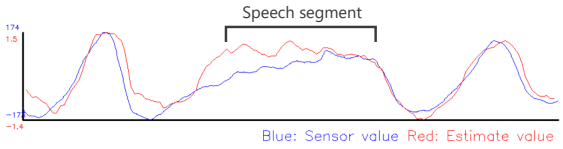

Figure 2: Example of real-time respiratory waveform comparison's window with added speech segment

3.1 Scenario

Initially, a user sits in a chair positioned so that the upper body is visible to the camera, which captures the movements of the chest and abdomen associated with respiration. The respiratory waveform estimator continuously uses the captured video frames to estimate the user's respiratory waveform values. As each estimation completes, the estimated values are immediately and continuously transmitted in real-time to both the voice activity predictor and the communication robot. During this process, the estimated waveform can be plotted and visually verified, enabling real-time confirmation of the accuracy of the estimations. Once the estimator begins transmitting the estimated values, both the voice activity predictor's prediction and the communication robot's pseudo-respiration presentation are initiated. These components start sending data to the spoken dialogue system simultaneously with their activation. Upon receiving these values, the system initiates a greeting, beginning the dialogue with the user. The system engages in dialogues on a variety of topics, capturing user speech through a microphone and considering it to generate contextually relevant responses.

3.2 Advanced Validation Features

The dialogue robot is equipped with various features to explore the effectiveness of using respiratory information.

Real-time Waveform Comparison: A user can attach a respiratory measurement device to their upper body, enabling real-time comparisons of actual waveform amplitudes with the estimated ones (Figure 2). Since the estimated waveforms and the actual waveforms have different ranges, they are displayed overlaid in a manner that aligns them to the same scale for comparison. This feature enables the user to directly observe the accuracy with which the respiratory waveform estimator is able to capture the user's respiratory waveform. This real-time feedback is crucial for validating the performance of the respiratory waveform estimator.

Using Actual Respiratory Waveform: The respiratory waveform estimator can also transmit actual respiratory waveform values obtained from a respiratory measurement device instead of the estimated ones. When the actual waveform values are transmitted, they are normalized to align with the scale of the estimated waveforms before transmission. Using these actual waveform values, we can verify the effectiveness of the dialogue robot in using the user respiratory information for human-robot dialogue, assuming that the waveform estimation is accurate.

Customizing Pseudo-Respiration Modes: The communication robot's pseudo-respiration presentation features three distinct modes: synchronized, steady, and no-presentation. In the steady mode, the communication robot follows a consistent, internally generated waveform, presenting pseudo-respiration independent of the user respiration. In the no-presentation mode, the communication robot does not move, and the spoken dialogue system responds based solely on the input from the voice activity predictor and the current absence of user speech, without considering the communication robot's respiratory phase. These options allow for a comprehensive evaluation of how respiratory synchronization and pseudo-respiration presentation affect human-robot dialogue.

Options for Voice Activity Predictor: The voice activity predictor offers a choice between using amplitude or gradient values as inputs. This feature enables to verify which input is more effective in real-world dialogue settings. In our experimental environment, using the estimated gradient values as inputs resulted in higher prediction accuracy than using the estimated amplitude ones (Obi and Funakoshi, 2023). Additionally, the predictor can be turned off, allowing one to observe the impact of its presence or absence on human-robot dialogue.

4 Conclusion and Future Work

In pursuit of facilitating smooth and engaging human-robot dialogue, we have developed a spoken dialogue robot that uses respiration. This dialogue robot employs a respiratory estimation method using video input to capture user respiratory information, which serves two primary purposes: predicting user speech onset to prevent speech collisions in dialogues, and presenting pseudo-respiration synchronized with the user's respiration. These approaches are expected to enhance the smoothness

and engagement of human-robot dialogue.

While adjusting speech timing contributes to smoother human-robot dialogue, reducing the number of robot utterances could detract the naturalness of the dialogue. To address this concern, future work will focus on incorporating non-verbal cues such as gaze into the voice activity prediction model, aiming to enhance its accuracy and ensure the dialogue robot does not unnecessarily remain silent. Additionally, accurate capture of user respiration is essential for the prediction in natural dialogue, so we will also work on developing a more robust respiratory waveform estimation method. Furthermore, we aim to develop a pseudo-respiration presentation that considers robot utterances, preventing a decrease in robot utterances while maintaining pseudo-respiration. Through these enhancements, we aim to use respiratory information more effectively, achieving more natural human-robot dialogue.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP22H04859. We thank Dr. Ludovico Minati, formerly with Tokyo Tech, for his help with the respiratory measurement device.

References

Kotaro Funakoshi, Kazuki Kobayashi, Mikio Nakano, Seiji Yamada, Yasuhiko Kitamura, and Hiroshi Tsujino. 2008. Smoothing human-robot speech interactions by using a blinking-light as subtle expression. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, page 293–296. Association for Computing Machinery.

Takao Obi and Kotaro Funakoshi. 2023. Video-based respiratory waveform estimation in dialogue: A novel task and dataset for human-machine interaction. In *Proceedings of the 25th International Conference on Multimodal Interaction*, ICMI '23, page 649–660. Association for Computing Machinery.

Takao Obi and Kotaro Funakoshi. 2024. Respiration-enhanced human-robot communication. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 813–816, New York, NY, USA. Association for Computing Machinery.

Amélie Rochet-Capellan and Susanne Fuchs. 2014. Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658):20130399.

Yunus Terzioğlu, Bilge Mutlu, and Erol Şahin. 2020. Designing social cues for collaborative robots: The role of gaze and breathing in human-robot collaboration. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 343–357, New York, NY, USA. Association for Computing Machinery.

T. Watanabe, R. Danbara, and M. Okubo. 2002. Interactor: Speech-driven embodied interactive actor. In *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*, pages 430–435.