

Silence is Golden – Making Pauses in Human Utterances Comprehensible for Social Robots in Human-Robot Interaction*

Thomas Sievers

Abstract—People pause when speaking for a variety of reasons – often in the middle of a sentence. It’s not easy for a machine to tell the difference between a pause for thought and an intended turn, but a smooth turn-taking is essential for flawless communication. Pauses within a sentence reveal something about the current emotional state of the speaker, and a correct interpretation of emotions is crucial for the mutual understanding of actors in human-robot interaction (HRI). How can we assess what the pauses a person makes in a dialogue with a social robot tell us about their emotional state? The text-to-speech tool Whisper from OpenAI enables robust speech recognition across different languages and the measurement of pauses between words. These pauses can be used to improve the assessment of the speaker’s emotional state by evaluating human utterances, including speech pauses, by a Large Language Model (LLM) from OpenAI (ChatGPT) using sentiment analysis. The inclusion of pauses as a non-verbal cue provides a helpful component for such an analysis.

I. INTRODUCTION

Understanding the spoken word is still a challenge for social robots in many respects. Difficulties can be caused by background noise, the speaker’s dialect or the robot’s speech-to-text software not taking rare terms into account. Apart from such more technical difficulties, one challenge for the robot in the dialogue is to recognize from the human’s word sequence when the human has finished and expects a response from the robot. Humans sometimes pause in the middle of a sentence for various reasons, even though they have not yet finished speaking. Pauses give the speaker a chance to continue thinking and the audience a chance to think at the same time. Recognizing the difference between a pause and an intended turn is not easy for a machine. However, functioning and fluent turn-taking is essential for smooth communication.

Spoken dialogue systems often detect the end of a turn by a certain amount of silence. However, silence is not a very good indicator of the end of a turn, as pauses within a turn can also cause silence and thus unintentionally trigger a response. Increasing the silence threshold could help, but will lead to delays in responses. For people, silence is clearly not the main signal to switch turns [1]. Many researchers have come to the conclusion that there must be some kind of prediction mechanism for turn-taking. Gaps between turns are short (of the order of 200 ms), but latencies in speech production are much longer (over 600 ms). This seems to indicate that the participants in a conversation need to predict

the end of the current speaker’ turn in order to prepare their response in advance [2].

In addition to the difficulty for a machine to correctly record the spoken word and recognize intentional turn-taking as such, there is also the challenge of correctly classifying what has been said based on non-verbal cues in order to be able to react appropriately. The speaker’s current emotional state generally influences pitch, volume and speaking speed. In this respect, a machine could use such characteristics of the sound to recognize emotions [3]. In addition, pauses within a sentence or utterance also reveal something about the emotional state of the speaker [4].

Robust speech recognition software is needed to overcome the difficulties of speech recognition, such as accents, dialects, interacting with children, difficult environments and recognizing pauses. Currently available social robots are not always equipped with such software. Therefore, in Human-Robot Interaction (HRI), it may be necessary to use external speech-to-text (STT) software. Whisper, developed by OpenAI, is an open source tool for automatic speech recognition (ASR) and translation for multiple languages [5]. Released in September 2022, Whisper was developed based on an encoder-decoder Transformer architecture trained on 680,000 hours of labeled audio data [6].

We use Whisper in an application for the social robot Pepper as a speech-to-text tool for human statements in dialogue with the robot. Whisper records the human utterances and measures the pauses between each word. These intervals are delivered to OpenAI’s Generative Pre-trained Transformer (GPT, commonly known as ChatGPT) [7] in addition to the spoken text. Since ChatGPT is able to perform a sentiment analysis of a text [8], it is then evaluated by the large language model (LLM) with regard to the mood and emotional state of the speaker.

This paper describes a method using the text-to-speech tool Whisper in combination with an LLM to detect and evaluate speech pauses and pauses between words in dialogues between a human and a robot, which can be used by the robot to evaluate the human emotional state.

II. RELATED WORK

In 2018, before the release of OpenAI’s Whisper, Gardecki et al. found the speech-to-text service provided by Google Cloud Services to be the most suitable for implementing natural language processing (NLP) methods in a Pepper robot [9]. In 2024, Pande et al. analyzed five open-source real-time speech-to-text recognition tools (Google Speech Recognition, Vosk, CMUSphinx, DeepSpeech, and Whisper)

*This work was not supported by any organization

Thomas Sievers is with the Institute of Information Systems, University of Lübeck, 23562 Lübeck, Germany
t.sievers@uni-luebeck.de

to determine the most suitable tool to integrate into a social robot to improve its speech recognition [10]. They considered Whisper to be the best option among the selected tools.

An integration of Whisper ASR and ChatGPT for conversations with a Pepper robot was evaluated by Chen et al. [11]. They compared Whisper with Google ASR as a Python library and Google Cloud ASR and found that Whisper performed best in terms of average Word Error Rate and processing time.

Janssens et al. evaluated whether ASR can process the semantic content of children’s speech [12]. They found that Whisper was able to transcribe children’s speech sufficiently to offer the potential for robust spoken interaction between children and robots. In addition, they found that using an external microphone as opposed to a microphone built into the robot resulted in significantly better recognition performance. Whisper was used by Pande et al. to transcribe speech into text to be displayed on the screen of a Pepper robot in a classroom to better understand and convey what the teacher was saying [13]. They used Whisper on a laptop and the Python-based version of Pepper instead of the newer Android version. The Python programmable version of the Pepper robot was able to record and process audio directly with its built-in microphones.

There is a large body of research on improving turn-taking in conversational systems including social robots. Skantze provided an overview of research on turn-taking in conversational systems and related multimodal indicators like verbal cues, prosody, breathing, gaze and gestures [1]. Bilac et al. presented a multimodal turn-taking system for humanoid conversational robots, in which they trained a machine learning system to recognize auditory filled pauses and integrate them into the robot’s dialogue system together with gaze direction observation [14]. They found that auditory pause detection in combination with gaze-tracking significantly reduced the interruption of the human sentence flow by the robot.

There are numerous scientific approaches to an automatic recognition and evaluation of emotions in human language.

In 2008, Vogt et al. examined the feasibility of human-computer interfaces that react in real time to emotions in the user’s voice and presented guidelines for the technical realization of such interfaces [15]. Schuller et al. investigated the automatic recognition of emotions from spoken words using vector space modeling compared to string kernels [16]. They analyzed the performance with an ASR engine based on continuous hidden Markov models. Rázuri et al. performed a parameterization of audio data for the purpose of automatic recognition of emotions in speech [3]. They evaluated six different types of classifiers for predicting six basic universal emotions from nonverbal features of human speech using a decision tree and found their approach suitable for a real-time system that can be used in social robots.

To the best of our knowledge, our approach of automatically analyzing the emotional meaning of pauses in human utterances using an LLM seems to be unique so far.

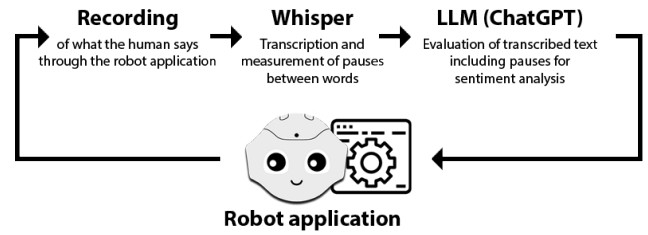


Fig. 1. Architecture and components of the proposed system

III. METHODS

In order to make pauses in human utterances during a dialogue with a social robot tangible and understandable for the robot, these pauses must first be recognized and measured. The humanoid social robot Pepper is optimized for human-robot interaction and should be enabled to recognize pauses in human conversation and assess their significance for the emotional state of the person.

The speech recognition that Pepper comes with is not able to distinguish between a turn-taking pause, where the robot is supposed to take over, and a pause that humans make for other reasons, e.g. when they are thinking. In such a mental pause without finishing the sentence, the robot would react to what has been incompletely said so far and formulate an answer. Such behavior would be inappropriate and would make a normal conversation by human standards very awkward. This makes it necessary to develop a speech-to-text system for the robot that can deal more flexibly with pauses in conversation.

Studies have shown that Whisper as a speech-to-text tool has good speech recognition even in difficult situations [10], [11], [12], while Pepper’s built-in speech recognition often fails, for example with accents, background noises or special terms. The text created by Whisper is passed on to ChatGPT to generate the robot’s responses in the dialogue.

Since we use Whisper as an online service via the transcriptions application programming interface (API) [17], there is a delay due to a chained use of online API connections once for speech-to-text by Whisper and then for generating the response by ChatGPT. These latencies are noticeable during the dialogue, although not significantly longer than when using Pepper’s built-in speech recognition in combination with ChatGPT to generate the response.

Figure 1 illustrates the components of the proposed overall system consisting of the robot application, Whisper for transcription and LLM for sentiment analysis.

A. Humanoid social robot Pepper

The humanoid social robot Pepper, shown in Figure 2, was developed by Aldebaran and first released in 2015 [18]. The robot is 120 centimeters tall and optimized for human interaction. It is able to engage with people through conversation, gestures and a tablet on its chest. Pepper is equipped with internal sensors, four directional microphones in his head and speakers for voice output. Speech recognition

and dialogue is available in 15 languages. The robot features an open and fully programmable platform so that developers can program their own applications for Pepper. In the model we use, which is based on the Android operating system, this is done using a software development kit (SDK) and the Java or Kotlin programming languages.

To improve the robot’s built-in speech recognition capabilities, the Whisper speech recognition tool is integrated using the transcriptions API for transcription of an audio file and to recognize and differentiate between pauses in speech. Since the Pepper robot model based on the Android operating system does not allow access to the audio files recorded with the built-in microphones, we alternatively use the microphone of the robot tablet to record the human’s dialogue parts. These audio files are sent to the Whisper API for speech-to-text conversion and measurement of pauses between words.

In addition to the ability to take pauses into account, Whisper also counteracts the limitation of Pepper’s speech recognition to hold multilingual conversations. With the Pepper robot, understanding a different language requires a programmatic change of language package each time, whereby the corresponding language package must be available on the machine. Whisper is able to understand many different languages by default.

To receive a response to a text generated by Whisper from the audio file, the robot application forwards this text as input to OpenAI’s ChatGPT API, which returns a dictionary with the status and response from the API. With each API call, the entire dialogue is transferred to the GPT model. This allows the LLM to constantly ‘remember’ what was previously said and refer to it as the dialogue progresses.

In this way, complex dialogues between humans and machines become possible. The text returned by ChatGPT is forwarded to the robot’s voice output. The general use of such APIs with the Pepper robot is described in more detail by Sievers et al. [19].

B. Measuring speech pauses using Whisper

The Whisper model was trained by OpenAI to predict approximate timestamps at the speech segment and word level for transcripts and video edits. These timestamps are recognized for the start and end of each individual word in the audio file so that a pause between adjacent words can be determined and output in the transcribed text. When calling the transcriptions API, we use *whisper-1* as the model and a sampling *temperature* of 0 for maximum deterministic output. Although generating word timestamps according to OpenAI results in additional latency, “word” must be chosen as the timestamp granularities since we need the word-based timestamps to measure the pause between words.

A text transcribed using Whisper, including pauses in seconds between the words in square brackets, looks like this, for example: “[0.0] Okay [0.26] Houston [0.08000004] we’ve [0.0] had [0.0] a [0.0] problem [0.0] here [0.26000023] This [0.0] is [0.0] Houston [0.29999995] Say [0.0] again [0.07999992] please”

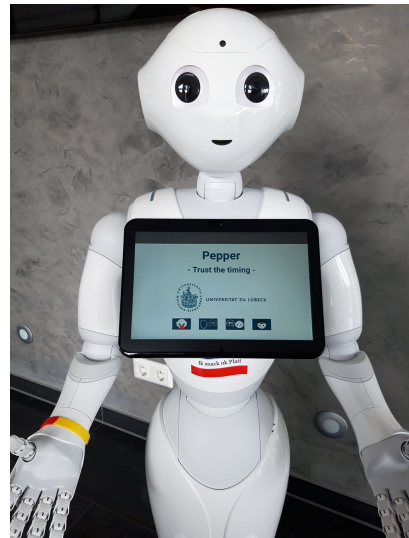


Fig. 2. Humanoid Social Robot Pepper

Most of the pauses in this example are very short or zero, but there are two longer pauses, once when switching from one speaker to another and once before “Say again please”. This is also an example where the longest pause in the dialogue is not where there is a turn of speaker.

C. Enabling turn-taking

At the beginning of a dialogue with the human, the robot application starts the audio recording of what the human is saying. From a certain sound level, for which favorable values must be tested, it is assumed that the human has started to speak. If the sound level subsequently falls below this value again, a time measurement is started in order to record the length of silence in the person’s utterances and to take over the conversation from a certain length. A pause at the beginning of the human speech is ignored.

We have tested pause times up to a turn-taking by the robot of 500 to 1500 ms, whereby a longer pause time clearly disturbs the flow of the conversation. When the robot takes over the conversation, the human utterances saved as an audio file are sent to Whisper for speech-to-text conversion. Once the robot has responded via ChatGPT, an audio recording is started again and the process begins anew. As already mentioned, silence alone is not a good indicator to recognize the end of an utterance, but the focus of this work is on the evaluation of pauses within an utterance and not on the detection of turn-taking. Nevertheless, a functioning turn-taking is required to create a dialog for our tests.

D. Prompting the GPT model

We used prompts for the system role to instruct GPT-4 to execute the tasks as a completion task. Zero-shot prompting [20] was used for this tasks.

1) *Prompting for Responses:* No particular specification was made to the GPT model regarding the response in the dialogue, except that the LLM should assume that it is a robot from the University of Lübeck named Pepper.

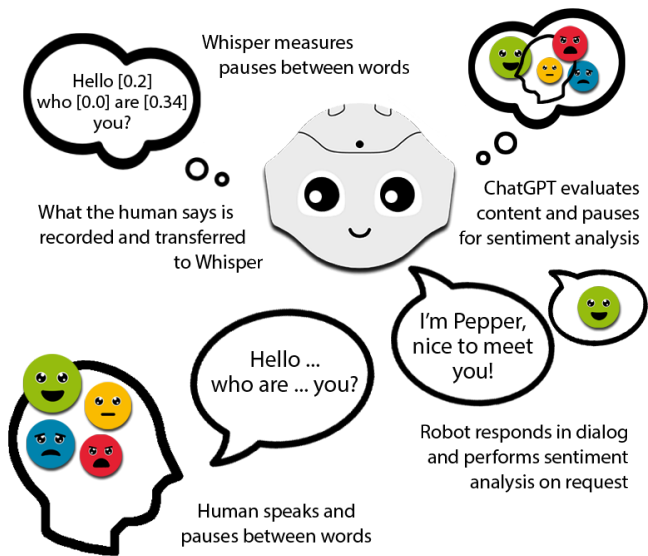


Fig. 3. Dialogue sequences and processing by Whisper and ChatGPT

2) *Prompting for Sentiment Analysis*: To perform the sentiment analysis with GPT-4, we used the following system prompt: “As an AI with expertise in language and emotion analysis, your task is to analyze the mood of the human in the following dialogue between a human and a robot. Take into account the general tone of the discussion, the emotions conveyed by the language used and the context. In the human’s utterances, the pauses in seconds between words are written in square brackets [] before each word. Analyze the person’s emotional state based on the length of these pauses. Explain your analysis briefly.”

The entire previous dialogue flow was then passed to the LLM as input for the user role, with the brackets being removed from the usual ChatGPT JSON notation and the role “user” being replaced by the term “human” and the role “assistant” by “robot” in order to obtain a textual impression of a dialogue flow. The time indications in square brackets for the pauses between the words have also been transferred. This results in input for sentiment analysis in the following form, for example: “Human: [2.86] Hello [0.0] Pepper [0.18000007] I [0.0] am [0.0] Thomas Robot: Hello Thomas. How can I help you?”

Sentiment analysis is started manually by clicking a button on the tablet. Figure 3 shows sequences in the dialogue and the effect of Whisper and ChatGPT.

E. Hyperparameters

The OpenAI API provides various hyperparameters for the GPT model that can be used to control the model behavior during an API call. As with the Whisper transcriptions API, the *temperature* was set to 0 in order to obtain consistent responses and exclude any randomness. The *top-p* hyperparameter determines the proportion of tokens with the highest probability from which to select. It ranges from 0 to 2, with lower values increasing determinism and higher values increasing randomness. It was set to zero. Positive values

of *presence penalty* would penalize new tokens, depending on whether they already appear in the text, increase the probability that the model will talk about a new topic. We set this value to zero. Positive values of frequency penalty would penalize new tokens based on their previous frequency in the text, reducing the likelihood that the model will repeat the same line verbatim. This value was also set to zero.

IV. RESULTS

Two different short dialogues between a human and a social robot and a third using an existing audio file without human-robot interaction show exemplary results of the assessment of the emotional state of human speakers.

A. First dialogue example

“Human: [0.0] Hello [0.0] Pepper [0.27999985] I [0.0] am [0.0] Thomas Robot: Hello Thomas! How can I help you? Human: [0.0] You [0.0] can [0.0] tell me [0.099999905] what [0.0] you [0.0] do here [0.0] Robot: I’m here to answer questions and provide information. How can I help you Thomas? Human: [0.0] Tell [0.0] me [0.13999999] who [0.0] you [0.0] are Robot: I am Pepper, a humanoid robot developed at the University of Lübeck. My main task is to help and support people.”

This led to the following emotional assessment by GPT-4: “The mood of the person in this dialogue seems to be neutral to positive. The pauses between words are very short or non-existent, indicating a relaxed and comfortable interaction. No signs of frustration, anger or confusion as the person’s questions are clear and direct. The human seems curious and interested in the robot and its functions. There are no signs of negative emotions or stress in the human’s speech or tone.”

B. Second dialogue example

“Human: [2.9] Hello [0.0] Pepper [0.48000002] how [0.0] are [0.0] you [0.0] today [0.18000003] I [0.0] am [0.0] Thomas [0.0] by the way Robot: Hello Thomas! As a robot I have no feelings but I am ready to help you.”

Emotional assessment of the LLM: “The person’s mood seems calm and friendly. The longest pause at the beginning of the conversation could indicate some uncertainty or hesitation, but the fact that most words are pronounced without pauses indicates a fluent, relaxed speaking style. The greeting and introduction (‘I’m Thomas, by the way’) show an open and polite attitude. There are no signs of anger, frustration or sadness in the person’s speech.”

C. Third example with existing audio file

In this example, a more dramatic dialogue in an unusual situation is analyzed. The basis for this is an existing audio file, which is sent to Whisper for transcription and to ChatGPT for evaluation.

“[0.28] Okay [0.26] Houston [0.08000004] we’ve [0.0] had [0.0] a [0.0] problem [0.0] here [0.26000023] This [0.0] is [0.0] Houston [0.29999995] Say [0.0] again [0.07999992] please [1.06] Uh [0.0] Houston [0.059999943] we’ve [0.0] had [0.0] a [0.0] problem [1.5999999] We’ve [0.0] had [0.0]

a [0.0] main [0.0] beam [0.0] undervolt [1.4200001] Roger [0.06000042] main [0.0] beam [0.0] undervolt [2.6200008] Okay [0.119999886] stand [0.0] by [0.0] at [0.0] 13 [0.06000042] We're [0.0] looking [0.0] at [0.0] it."

Emotional assessment of the LLM: "The mood of the person in this dialogue seems to be tense and worried. This is indicated by the use of words such as 'problem' and 'undervolt' in a technical context, suggesting a possible malfunction or error. The pauses between words vary, but the longer pauses (1.06, 1.6, 1.42 and 2.62 seconds) could indicate some uncertainty or thoughtfulness as the human tries to understand or explain the problem. The repeated use of the word 'Houston' indicates that the individual may be communicating with a control center, which increases pressure. The request for repetition and the request to wait ('Say again please', 'stand by') could also indicate some urgency or concern."

In general, it can be stated that the LLM essentially analyzes content aspects for an evaluation of the emotional meaning of human utterances and thus the emotional state of the speaker, but also takes into account pauses in the utterances. Pauses can therefore be used as an additional criterion and non-verbal cue to carry out a sentiment analysis of human speech. Such an analysis can form the basis for appropriate utterances and reactions of the social robot.

V. DISCUSSION

Setting up a technical system for robust and reliable speech recognition for social robots in everyday situations is no easy task. Existing solutions built into robots often have their limitations. The speech-to-text tool Whisper from OpenAI is generally well suited for flexible and reliable speech recognition. However, using it via the transcriptions API causes additional latencies, especially if times are to be measured at word segment level in order to determine pauses between words.

Occasionally, Whisper creates hallucinations during a lengthy silence and produces texts such as "thank you for watching" or "thanks for watching", although nothing was actually said. These false sequences can be filtered out before further processing if their content is known. It is also conceivable to carry out Voice Activity Detection (VAD) before using Whisper.

Reliable turn-taking adapted to human behavior remains difficult if the only criterion is a certain period of silence. Whisper can provide helpful hints for this by recognizing pauses, but other aspects must be added to capture the meaning – e.g. gaze direction analysis or possibly a kind of prediction of an upcoming change.

VI. CONCLUSIONS AND FUTURE WORK

In order to be able to derive a comprehensive assessment of a person's emotional state from their utterances, the inclusion of pauses between words is a helpful component of the analysis. Such an analysis can be carried out by determining the pauses between words and with the help of an LLM using sentiment analysis. However, it should be

checked for reliability by comparing it with analyses carried out by humans.

Thanks to its word timestamps, Whisper is well suited to determining pauses and is also a useful tool for social robots in HRI due to its almost universal applicability for different languages and robust speech recognition. ChatGPT allows the inclusion and consideration of pauses to evaluate emotional aspects of what is said in addition to purely content-related aspects.

In future work, a multimodal approach, e.g. through an additional interpretation of facial expressions and gaze directions, could help to differentiate between end-of-sentence, turn-taking and confusion or thinking pauses. The reduction of latency times would be helpful for a practicable application with human-like reaction times. Insights from sentiment analysis can be flexibly incorporated into the reactions of a social robot and its utterances and behaviors can be tailored more precisely to the person with whom it is interacting.

REFERENCES

- [1] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review", *Computer Speech & Language*, Volume 67, doi: 10.1016/j.csl.2020.101178, 2021.
- [2] S. C. Levinson, and F. Torreira, "Timing in turn-taking and its implications for processing models of language", *Frontiers in Psychology*, Volume 6, doi: 10.3389/fpsyg.2015.00731, 2015.
- [3] J. F. Rázuri, D. Sundgren, R. Rahmani, A. Larsson, A. Moran, and I. Bonet, "Speech emotion recognition in emotional feedback for Human-Robot Interaction", *International Journal of Advanced Research in Artificial Intelligence*, Volume 4, doi: 10.14569/IJARAI.2015.040204, 2015.
- [4] C. Sobin, and M. Alpert, "Emotion in Speech: The Acoustic Attributes of Fear, Anger, Sadness, and Joy", *Journal Psycholinguist Research*, Volume 28, pp. 347–365, doi: 10.1023/A:1023237014909, 1999.
- [5] OpenAI. (2024) Introducing Whisper [Online]. Available: <https://openai.com/index/whisper/>
- [6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision", *Proceedings of the 40th International Conference on Machine Learning*, PMLR, pp. 28492–28518, 2023.
- [7] OpenAI. (2024) The most powerful platform for building AI products [Online]. Available: <https://openai.com/api/>
- [8] T. Sievers, and N. Russwinkel, "Interacting with a Sentimental Robot – Making Human Emotions tangible for a Social Robot via ChatGPT", *Workshop on Advanced Robotics and its Social Impacts (ARSO)*, IEEE, doi: 10.1109/ARSO60199.2024.10557749, 2024.
- [9] A. Gardecki, M. Podpora, R. Beniak and B. Klin, "The pepper humanoid robot in front desk application", *Progress in Applied Electrical Engineering (PAEE)*, doi: 10.1109/PAEE.2018.8441069, 2018.
- [10] A. Pande, B. Shrestha, A. Rani, and D. Mishra, "A Comparative Analysis of Real Time Open-Source Speech Recognition Tools for Social Robots", In: Marcus, A., Rosenzweig, E., Soares, M.M. (eds) *Design, User Experience, and Usability. HCII 2023. Lecture Notes in Computer Science*, vol 14033, Springer, Cham, pp. 355–365, doi: 10.1007/978-3-031-35708-4_26, 2023.
- [11] X. Chen, K. Luo, T. Gee and M. Nejati, "Does ChatGPT and Whisper Make Humanoid Robots More Relatable?", doi: 10.48550/arXiv.2402.07095, 2024
- [12] R. Janssens, E. Verhelst, G. Abbo, Q. Ren, M. Pinto, and T. Belpaeme, "Child Speech Recognition in Human-Robot Interaction: Problem Solved?", doi: 10.48550/arXiv.2404.17394, 2024
- [13] A. Pande and M. Mishra, "The Synergy between a Humanoid Robot and Whisper: Bridging a Gap in Education", *Electronics*, Volume 12, doi: 10.3390/electronics12193995, 2023.
- [14] M. Bilac, M. Chamoux, and A. Lim, "Gaze and filled pause detection for smooth human-robot conversations", *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pp. 297–304, doi: 10.1109/HUMANOIDS.2017.8246889, 2017.

- [15] T. Vogt, E. André, and J. Wagner, "Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation", *Affect and Emotion in Human-Computer Interaction*, Lecture Notes in Computer Science, Volume 4868, doi: 10.1007/978-3-540-85099-1_7, 2008.
- [16] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion recognition from speech: Putting ASR in the loop", 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4585–4588, doi: 10.1109/ICASSP.2009.4960651, 2009.
- [17] OpenAI. (2024) Create transcription [Online]. Available: <https://platform.openai.com/docs/api-reference/audio/createTranscription>
- [18] Aldebaran, United Robotics Group and Softbank Robotics. (2024) Pepper [Online]. Available: <https://www.aldebaran.com/en/pepper>
- [19] T. Sievers, M. Bender, and R. Möller, "Connecting AI Technologies as Online Services to a Humanoid Service Robot", 15th International Conference on Computer and Automation Engineering (ICCAE), IEEE, pp. 431–435, doi: 10.1109/ICCAE56788.2023.10111181, 2023.
- [20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, and P. Dhariwal et al., "Language models are few-shot learners", *Adv. neural Inf. Process. Syst.* 33, pp. 1877–1901, 2020.