Smart Beamforming in Verbal Human-Machine Interaction for Humanoid Robots

Bartlomiej Klin, Michal Podpora, *Member, IEEE*, Ryszard Beniak, Arkadiusz Gardecki, and Joanna Rut

***Abstract*—The ongoing industrial revolution involves an increasing need for human-machine communication in all areas of everyday life. Efficient and flawless human-machine communication becomes a large challenge in crowded environments.The communication scheme used in present robotic and humanoid systems usually does not assume simultaneous interaction with multiple interlocutors (e.g. robot and two talking persons) as well as the associated mechanisms for distinguishing between speakers, which is being simplified to non-overlapping, sequential exchange of questions and answers between two interlocutors. This article presents a novel framework outlines to improve human-machine communication in the humanoid robot area. Concepts of acoustic-visual beamforming are already known, the implementation of which is an acoustic camera. However, according to the authors, combining acoustic and visual perception requires significant changes in the approach to the design of robotic systems in order to be able to take full advantage of the multi-talker capability. The perception of an interlocutor is more natural when a human is able to perceive multi-channel information. This provides the acoustic-visual sensors to be able to support beamforming of audio signal and assign these signals to every interlocutor in the engagement zone of a humanoid robot. This can be useful in a time regime adequate for a conversation.**

***Index Terms*—Software-hardware integration for robot systems, long term interaction, multi-modal perception for HRI, natural dialog for HRI, design and human factors.**

I. Introduction

HUMANS since the mid-twentieth century began to operate machines by sequentially inputting information and after a while receiving a result [1] – in an analogous way to how an operating system's command line works. This scheme of interaction can be named the Read Eval Print Loop (REPL), while it includes distinct stages of communication: a signal of readiness to act, entering a command or input data, information processing, and finally presenting the output information – the result of the processing. Such a processing scheme applies to both the speech-based and textual communication, and is used nowadays, mainly due to the limitations of hardware and software solutions (not only in popular humanoid robot platforms [2], [3]). Voice communication, compared to text-based communication is particularly difficult to implement [4], [5] for machines (and especially for humanoid robots working in public places) for several reasons. The first one is the fact that typically the interlocutors do not get specific instructions for working with a robotic system, presuming that it operates according to the REPL scheme. Whereas, human-human voice communication does not fall in to the REPL scheme [6], but only at most good manners, to which there is no obligation to follow [7], [8]: fixed order of speech, not interrupting the speaker, patiently waiting for the allotted speech time slot, etc. Another reason is the limitations imposed by the usage of (one) static beamformed cone, limiting in advance the service to a single interlocutor, which in turn implies the lack of speaker discrimination in the case of multiple simultaneous interlocutors [9], [10]. Thirdly, existing beamforming solutions used in humanoid robots assume that the sound source does not move [4], which may be a false assumption, especially if the conversation involves more than one human interlocutor, who may engage in conversation with each other and/or lose interest in interaction with the robot (which involves movements of heads, changes in the direction of sound, or changes of relative position/location).

The authors, as a part of their research and development work (hosted and supported gratuitously by the Weegree company), faced numerous challenges in relation to their vision of a humanoid robot working as a front desk officer alongside human employees. The prototype was implemented using the Pepper robot platform from United Robotics Group (Softbank) [11], [12] as the system's Human-Machine Interface.

The versatile development and use of humanoid robots as speech-communicated multipurpose human companions and co-workers is gradually gaining momentum. In line with the trend toward replacing humans in their boring and repetitive jobs (including the economical condition – that the balance of a robot's labor cost is less than that of a human [13]). Two directions in humanoid robots development are evident, and both are extremely important. The first one emphasizes the mobility and agility capabilities [14], and the second one promotes cognitive and communication capabilities [15].

The Pepper robot, as a platform for exploring the above-mentioned interaction-oriented humanoid design trends, seems to be used mainly in the cognitive- and communication-related research, which implies its primary use for communication with

Manuscript received 11 January 2023; accepted 9 June 2023. Date of publication 21 June 2023; date of current version 29 June 2023. This letter was recommended for publication by Associate Editor H. S. Ahn and Editor A. Faust upon evaluation of the reviewers' comments. *(Corresponding author: Bartlomiej Klin.)*

Bartlomiej Klin, Michal Podpora, Ryszard Beniak, and Arkadiusz Gardecki are with the Faculty of Electrical Engineering, Automatic Control and Informatics, Opole University of Technology, 45-758 Opole, Poland, and also with the Weegree Sp. z o.o. S.K., 45-018 Opole, Poland (e-mail: b.klin@student.po.edu.pl; michal.podpora@gmail.com; r.beniak@po.opole.pl; a.gardecki@po.edu.pl).

Joanna Rut is with the Faculty of Production Engineering and Logistics, Opole University of Technology, 45-272 Opole, Poland (e-mail: j.rut@po.edu.pl).

Digital Object Identifier 10.1109/LRA.2023.3288381

humans. It is not a utility robot that performs physical activities – as the manufacturer states, *'A robot designed to interact with humans'* [16], [17]. The Pepper's information inputs are based upon the hardware sensors provided by the manufacturer – including: a four-microphone array placed on the robot's head, 2D and 3D cameras, a set of LIDARs, sonars, and infrared proximity sensors. The Pepper robot also has sufficient information channel capabilities regarding the information output. In addition to the audio/speech output (i.e. the speakers), and the possibility to communicate using non-verbal communication (i.e. gestures and movements), an important input-output device is present on the robot's chest – a multimedia tablet, running the Android operating system integrated with the robots API [16].

Comparing the robot's capabilities outlined in the information brochure [18] against the functionality required for a robot-receptionist, set by the company, helps to assess the scope of challenges to be taken. The naive assumption that all that is left to do is to program the robot by filling its database with the necessary content (promoted by the robot manufacturer, and taken for granted by the customers), could not be further from the truth [19]. After basic preliminary research, it became obvious that Pepper's built-in mechanisms (sensors, input information, processing capabilities, access to remote services) need to be improved/extended to achieve satisfactory performance in a natural environment.

This particular paper is devoted solely to the improvement of the audio-based NLP-oriented information acquisition of a humanoid robot. Pepper uses a microphone array arranged in a rectangle shape, placed on the top of the head, essentially for two tasks: to detect the direction of arrival (DOA) of incoming sound (the probable location of interlocutor), and to compose the acquired four audio signals into one (taking a benefit from the robot's head facing the interlocutor). As a result, only the signal received directly from the front of the robot's head at a specific angle (corresponding to the hypothetical location of the interlocutor's speech apparatus) will be amplified [20]. Concluding, Pepper acquires speech signal from the direction where its gaze is focused. It is essential to keep this detail in mind. Even a seemingly insignificant offset from the axis of Pepper's head will significantly degrade the beamforming result from its built-in microphone array [21]. The static beamforming mechanism requires prior localization of the interlocutor [22]. A video subsystem combined with proximity sensors and estimated sound DOA is used to make the necessary adjustments to the robot's head or entire body by rotating them in the direction targeting the selected active speaker. Such a construction requires appropriate workflow management (according to the REPL scheme) of voice interaction with a human since the robot cannot switch between interlocutors simultaneously talking during the ongoing conversation. Therefore, using the appropriate interaction workflow requires individual profiled training which, unfortunately, the potential supplicant at the company's reception desk is not likely to get. The robot's manufacturer outlines the following 'interaction tips'[3]: *'Pepper's hearing area is narrow'*, *'Keep eye-contact when interacting with Pepper'*, *'Stay in front of Pepper'*, *'Pepper is listening when shoulder LEDs are blue and blue line on the tablet'*, *'Do not speak at the same time'*, *'Use short sentences'*, *'Do not speak too fast'*.

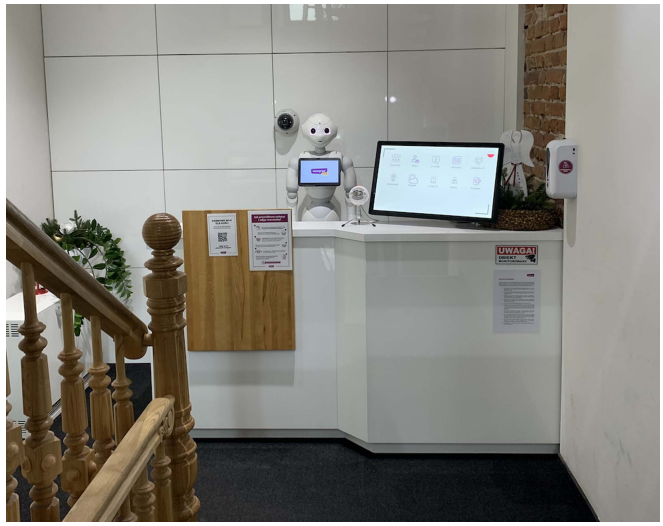

Fig. 1. Front view of robot's workstation.

The authors' motivation for the present research is the perceived need for more commercial humanoid robot solutions suitable for unhindered and simultaneous communication with at least two interlocutors in crowded public spaces while still achieving the budget range conditions of the small company [23]. Such communication is the key to engaging the humanoids for multi-talker applications. Previously this was not possible (not supported), preventing the use of robots e.g. in such applications as a multilingual hostess at trade fairs or as an office worker accepting applications.

The authors' present research is focused on the verification of the following research question: Is smart beamforming enabling multi-talker capabilities in verbal human-machine interaction for humanoid robots? Verification steps begin with presenting the context and results of measurements and conclude with investigating solutions to propose an outline that includes the tasks necessary to achieve multi-talker capabilities. The idea of audio and video sensor fusion combined with beamforming is researched along with all implications. Note that authors are not exploring any specific algorithm of neither beamforming [24] nor NLP [25], relying upon an external library for now. More important to the authors is the emphasis on the examination of real-world cases by collecting input data correctly, thus being able to replay them in another modified execution environment built exclusively to verify the research question and deliver validated proof of concept for future research and development.

II. Materials and Methods

The robotic testbed was built in a natural environment – in the Weegree office (Opole, Poland), and consisted of a Pepper robot placed behind a desk (shown in Fig. 1), supplemented by additional hardware for monitoring the robot's surroundings, to have independent data from outside the robotic system [10]. For this purpose, an external microphone array and an additional camera were installed. Data inputs from external sensors while maintaining a temporal inputs stream synchronization. It was

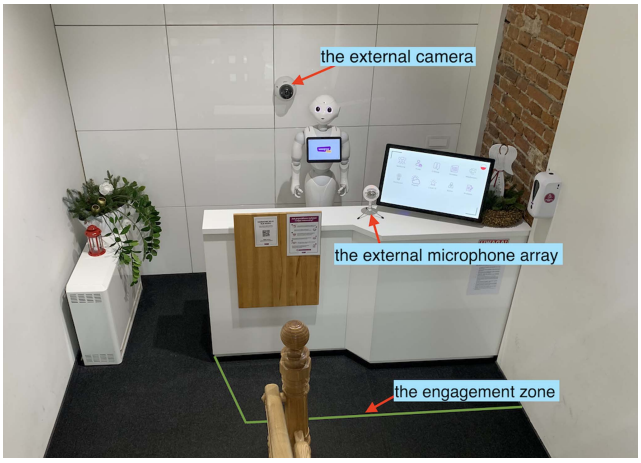

Fig. 2. Layout of robot's workstation.

essential in order to be able to replay recordings later, in a modified execution environment. The layout of robot's workstation is shown in Fig. 2. The process was supervised by the authors.

The robot was programmed (by the authors) to act as a receptionist whose task is to show and explain the way to visitors. The robot's built-in mechanisms (ALDialog module) [26] were used for this to provide answers about the location of the target room based on a query vocalized by the interlocutor: – The purpose of the visit (e.g.: "I am looking for a job as a car painter"), – or a specific employee or meeting (e.g.: "I have an appointment with Joe Doe").

Pepper's screen was used for indicating the robot's state according to the REPL scheme and presenting the speech to text (STT) transcription result along with the sentences currently being spoken by the robot.

The field study took four days, resulting in acquisition of 140 conversations of English-speaking interlocutors (office visitors/employees who consented to converse with the robot). Two variants of interaction initiation models were implemented to verify the interlocutor's awareness of possible communication schemes. In the first scenario, the robot initiated the conversation after detecting the presence of a probable interlocutor and encouraged him to ask the question. In the second one, on the other hand, after detecting the interlocutor, it directed its gaze but remained silent, surrendering the opportunity to initiate the conversation.

In addition to the audio and video data, the following relevant factors were considered during the data acquisition phase:

- the position of the interlocutor's speech apparatus in relation to the robot – was determined as the deviation from the axis of the beamformed cone in space in pixels measured with the camera built into the robot's head combined with a readout of the proximity sensors [26]. The center of the beamformed cone axis is considered to coincide with the center of the camera (robots head's built-in camera) image,
- familiarity with the robot's operation (in particular, the REPL scheme) – was defined as the fact of reaching a timeout of 10 seconds of elapsed time since interlocutor detection to conversation initiation. In the first case, when the robot is waiting in silence, it notes the time the interlocutor's first words were vocalized. In the second case, when the robot actively invited the interlocutor to the dialog, the time of the given answer matched in a programmed dialog tree is noted,
- the presence of an adequate signal level at the input of the speech processing system – was defined as the ratio of the duration of audio samples in seconds to those in which there was a sufficient power level taking as a rule of thumb that the number 1 refers to 100% filling of the samples window with at least a minimum required power level signal, and 0 means no samples below a minimum power level (i.e. all samples were categorized as noise),
- and the number of interlocutors participating in the interaction with the robotic system at one time – by counting the number of people detected in the engagement zone [26] by the robotic system in combination with faces counted using an external camera.

The authors have identified a roadmap of checkpoints required to develop the capability to handle multi-talker conversations. Table I presents the most important insights, arranged from the most general to more specific ones:

1) disregarding the 'tradition' to design the communication in compliance with the REPL scheme is possible and purposeful – there is no need to know the workstation manual due to the shifting the workflow towards a natural workflow for humans (such as in a human-to-human communication) [27],
2) the lack of a REPL communication scheme implies design changes in data processing so that the following becomes possible: (a) stand not in front of the robot, (b) speak at the same time as the robot, (c) speak longer sentences, (d) interact with another interlocutor and a robot during the same session,
3) support for two interlocutors requires adaptive beamforming [28], [29] (as opposed to permanently fixed virtual axes of the static beamformed cone and a moving head/body). The ability to move the center of the virtual axes in space without moving the head/body, combined with multiple data pipelines, allows any number of streams (interlocutors) to be extracted,
4) processing multiple streams of the beamforming module invalidates the concept of single STT service instance for delivering text for further NLP (Natural Language Processing) [30],
5) the multiple occurrences of STT and NLP instances require linking them to a relevant interlocutor and persisting the context related to specific parties [31],
6) distinguishing interlocutors and having keys to identify individual conversation instances/parties requires linking (digital fingerprints of the interlocutors') voices with their faces (digital fingerprints of their images), as well as tracking their movement,
7) interaction with multiple interlocutors may require a speech task queuing system that is able to manage the STT queue and to cancel STT tasks at any time,

TABLE I
ISSUES AND SOLUTIONS

No.	Issue description	Results	Solutions
1	An interlocutor does not know when to speak to the robot and tries to speak when the robot speaks.	No audio stream at the input of the STT system.	Need to overcome lack of knowledge of the REPL communication scheme. Turn on the echo cancellation mechanism at the input of the speech recognition system, and do not turn off the speech recognition system while the robot is synthesizing speech.
2	Case 1: An interlocutor does not stand in the axis of the beamformed cone. Case 2: Two interlocutors, but neither in the axis of the beamformed cone.	No audio stream at the input of the STT system.	Need to overcome static beamformed cones and the requirement to position the robot against the participant. Adaptive beamforming that allows to steer the beamforming parameters, in particular, steering the cone to target participants.
3	Two interlocutors, one in the axis of the beamformed cone.	Only one audio stream is present at the input of the STT system.	The need to overcome static beamformed cones and introduce adaptative beamforming with multiple-talker capabilities. Also, monitoring participants' signal levels will be required to detect incorrect states.
4	Two interlocutors properly aligned against beamformed cones, however, with overlapping speech.	Absence or malformed audio stream at the input of the STT system.	Need to overcome the lack of queuing capabilities in the conversation subsystem by introducing multi-instance processing pipelines (one pipeline for every interlocutor) linked with the context recovery subsystem based on unique fingerprints (face, voice).

8) a successful multi-talker conversation may be possible under certain conditions. First of all, the robot must be aware of the addressee of every speech utterance to avoid confusing interlocutors with unwanted responses. Such detection can be made by combining the signal level at the output of the beamforming module with body/face posture angle tracking. Secondly, the biggest challenge would be to parse, store, link with context, and reuse the acquired sentences in conversation and conveyed knowledge in case the robot is not an addressee. It would significantly leverage the quality of conversation between the system and a human.

For the experiment, a set of add-on devices (a microphone array and a camera), was connected to an external system, implementing the basic functionality indicated in the roadmap toward the adaptive beamforming [29] linked to a video system. The authors propose to refer to such a combination by the "Smart Beamforming" term. The first results were produced using the robot's embedded system. Secondly, external add-on equipment was used to provide a better quality version of raw data to be processed using an external execution environment (see Fig. 4).

A visualization of the idea of simultaneous data acquisition using two separate pipelines, basic and experimental, working on the same input, is shown in Figs. 3 and 4.

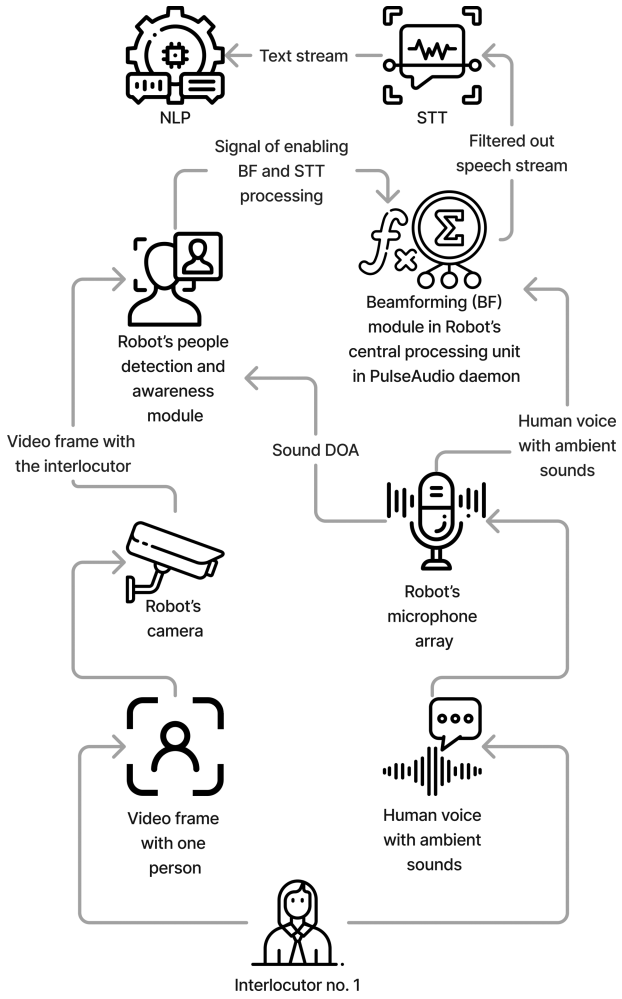

Fig. 3. Information flow in the basic pipeline.

TABLE II
CASE 1: ONE INTERLOCUTOR, SPEAKING IN THE ROBOT'S DIRECTION, NOT IN LINE WITH BEAMFORMED CONE AXES (NUANCE-BASIC PIPELINE)

No.	Ground truth	STT result	WER
1	Good morning, I'm looking for a job	Good morning I'm	7.892
2	Hey robot, I've appointment with recruiter	Robot recruiter	15.642
3	I am here to deliver post	I am pot	8.672
4	Good morning, what's your name	Good you name	9.975
5	Hi I came to see Tom for a recruitment interview	come recruit	25.030

III. RESULTS

The following tables (Tables II, III, and IV), present selected interesting illustrative results/cases regarding the analysis of the correctness of STT translation for three available engines and pipelines (Figs. 3 and 4), while taking into account some of the commonly observed interaction scenarios [32].

The designed and constructed pipelines are using:

- ALSpeechRecognition (Nuance internal) [33], the out-of-the-box solution embedded in the Pepper robot,

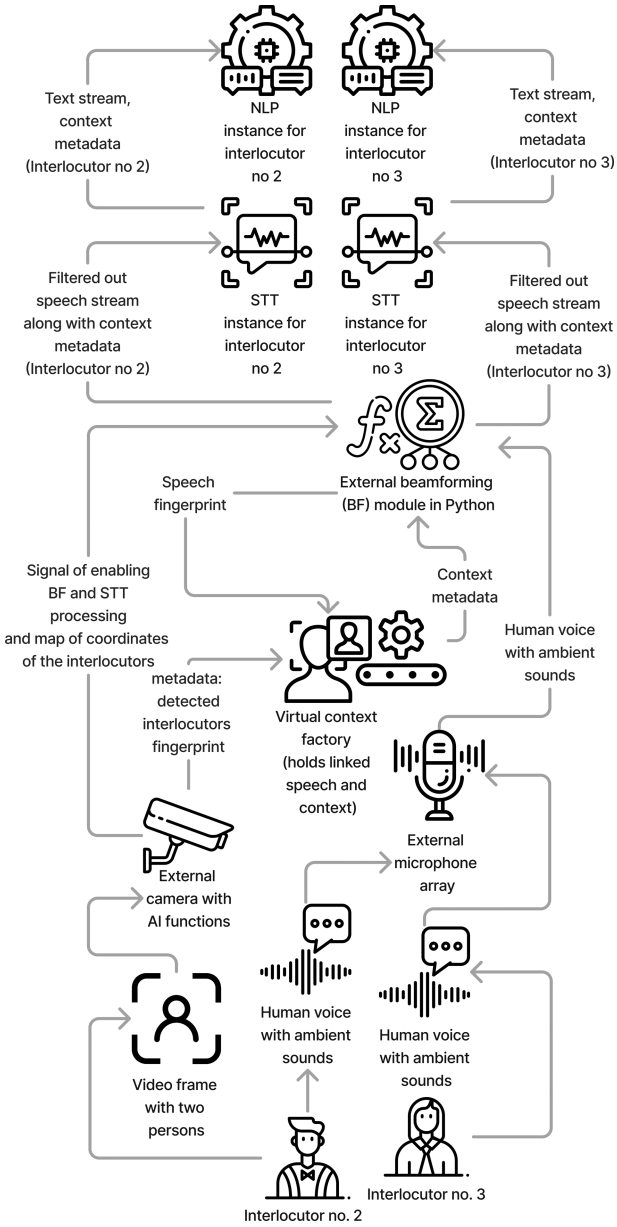

Fig. 4. Information flow in the modified experimental pipeline.

- an external open-source STT engine – Mozilla Deepspeech [34],
- another external open-source STT engine – OpenAI Whisper [35].

STT correctness (/quality) was assessed by using standard word error rate (WER) [36] parameter calculated using Leven-Phrase [37] approach (proposed and developed by authors [38]), adjacent to additional parameters to give a better understanding of causes and sources of emerging differences in the results and allow their better interpretation:

- position offset of interlocutor in relation to the robot,
- the fact of reaching speech input data timeout,
- the number of acquired samples validated as no-noise (carrying potential speech utterances),
- the number of interlocutors.

TABLE III
CASE 2: ONE INTERLOCUTOR, SPEAKING IN THE ROBOT'S DIRECTION, IN LINE WITH BEAMFORMED CONE AXES (MOZILLA DEEPSPEECH-MODIFIED PIPELINE)

No.	Ground truth	STT result	WER
1	Hi, where I can find Anna	Hi where I can find Anna	0
2	How long I have to wait	How long I have wait	0.8572
3	I have big package for you	I have big package for	0.889
4	Where is public toilet	Where is public toy	2.992
5	Hi I came to see Tom for a recruitment interview	Hi came to see Tom for recruitment	7.468

TABLE IV
CASE 3: TWO INTERLOCUTORS, ONE OF THEM IS NOT IN LINE WITH BEAMFORMED CONE AXES (OPENAI WHISPER-MODIFIED PIPELINE)

No.	Direction	Ground truth	STT result	WER
1	A-robot	Pepper what's up	Peter what's	0.884
2	A-B	Look over there at the screen	Look over there screen	2.798
3	A-B	It's creepy while it moves	creepy	9.602
4	A-robot	Tell me a joke	Tell me a joke	0
5	B-robot	What time is it now	what	8.407
6	B-robot	Pepper tell me where i can find IT department	Pepper toll	27.449
7	B-A	Allright lets go	All	15.000
8	B-A	I am leaving in five minutes	leaf	19.855
9	A-robot	Hi I came to see Tom for a recruitment interview	Hi came to see Tom for recruitment intervene	6.704

Interlocutor A is in line with beamformed cone axes. Interlocutor B is not in line with beamformed cone axes.

WER results were collated by the STT engines, providing a view of the difference between the embedded solution (basic pipeline) and the modified pipelines.

The first case, presented in Table II, is a conversation between one interlocutor and a robot using the Nuance engine built into the robot. However, there is an offset of the beamformed cone relative to the interlocutor's speech apparatus. The robot has initiated a conversation, and the participant is aware of the REPL interaction scheme.

The second case, presented in Table III, is a conversation between one interlocutor and a robot using Mozilla Deepspeech's STT external engine. Compared to the first case (Table II), despite the offset of the axes of robot's head in relation to interlocutor's speech apparatus, there is no shift of the beamformed cone due to active adaptative beamforming. The participant starts the conversation and takes over the initiative within the dialogue.

The third case, presented in Table IV, is a conversation between two interlocutors with a robot and each other, translated using external STT engine – OpenAI Whisper. In this case, as for the first interlocutor, there is no shift of the beamformed cone relative to his speech apparatus. However, while he is not facing the robot, combined with incorrect settings and a lack of sufficient monitoring of the signal level at the input of the STT system, the beamformer cone shifts for the second interlocutor.

The results in the last rows of Tables II, III, IV describe the same conversation replayed in every pipeline for comparison.

IV. Discussion and Conclusion

The examples presented in Tables II, III, and IV were selected not as representative results, but rather as meaningful use-case examples, as these happened to be surprisingly common scenarios during the research experiment. The authors have selected them for illustration purposely.

Comparison of the results in Table II against the results in Table III indicates, as expected, that the shift of the beamformed cone axis has a significant impact on the result of speech-to-text translation due to greedy filtering strategy, cutting out relevant pieces of information. Logbook analysis revealed that the cones' axes were affected by the vision system operating too slowly and erroneously. It also caused the beamformer to take too long to set up, therefore not capturing the beginning of interlocutor's speech.

Table III presents cases when the video processing system performs well, allowing the setting of the beamformer correctly. The translation of speech to text was significantly better. However, in sentence No. 2, a lack of the word "to" is observed, and sentences 3 and 4 are truncated (by reaching the time limit of the speech). The analysis of event sequences hints at the main cause for this – an unintentional speech delay made by interlocutors (up to a few seconds) while the system was already acquiring the silence instead of the speech, respectively shrinking the time span left of the available utterances buffer. Such behavior is expected in systems that work in accordance with the REPL scheme, where the data acquisition phase is time limited and may be exceeded by unaware interlocutors during first interactions.

Table IV presents selected examples from conversations between a robot and two human participants. One of the human interlocutors was correctly positioned to the microphone array, and the corresponding beamformer instance was properly set up. Contrary to the second human interlocutor, who was facing the first interlocutor, not the robot. The results show that the system correctly detects and marks the direction of the interlocutors' speech and later purposely drops from the NLP answering queue utterances not intended for the robot.

The phrases in which the addressee is not a robot are meant to be ignored silently by the system (human interlocutors interact without a robot, although they still stand within the engagement zone). A research testbed was designed with this functionality in mind. Regardless, additional use cases emerged in an experiment run. It would be valuable and practical if the robot could memorize facts spoken in its proximity/engagement zone. The analysis of performed interactions supplied hints that such speech should be processed carefully and silently (without producing an answer) to extract facts by the NLP module and store them along with the proprietor's and addressee's details. Such functionality could be introduced as an additional module extending awareness functions – working in the background and providing only contextual input. The authors would like to raise one more issue: the extent to which the system should trust facts extracted from the background. Validation of such facts, in relation to general knowledge, is a software engineering challenge – similarly to verifying e.g. the correctness of the latest "Chat GPT" implementation [39]. However, verification of any abstract knowledge or information which is not available publicly is difficult.

The challenge of leveraging human-machine communication by using all available means is complicated and demanding in resources. Likewise, introducing the "smart beamforming" feature implies significant changes in processing pipelines. The full potential of acquired surrounding speech-sourced knowledge in an improved system also involves extensive development of NLP features, especially access to general public knowledge and ways of evaluating the trustworthiness of gathered information. Abstract and subjective knowledge also should be at least detected and marked as such. The below-mentioned barriers and risks were identified and enumerated regarding beamforming and processing pipelines:

A. The Risk of Over-Filtering

The high risk of over-filtering [40] is a critical aspect of beamforming with microphone arrays caused by aggressive filtering strategies erasing useful audio information. Beamforming includes two tasks: isolating target speech, and denoising of ambient sounds. The denoising effect relies on wisely used filters cutting out frequency ranges that do not carry information about the human voice (necessary for the STT services). The isolated interlocutor's speech should be lossless, at least in a well-known range of speech-carrying frequencies. Relying on beamformer cone settings configured only by the video system is often faulty and delayed below the required threshold of accommodation to the environmental changes. Therefore, the video module needs to be fast enough and consistently deliver new parameters to the beamformer on time. Otherwise, it may cause discrepancies between the current cone settings and the interlocutors' actual locations, resulting in obstructed beginnings or ends of phrases. As a possible remedy, feedback loops measuring signal level may be proposed or a sensor fusion of video and parallel/supplementary sound-based localization may be applied. The additional sound localization can be performed by the following of sound sources based on triangulation with conditions of preponderant share of frequencies typical for the human speech.

B. Audio and Video Processing Time

The real-time/near real-time processing of audio and video Maintaining the time domain synchronization [41] necessary for the control of beamforming settings is important for the NLP and STT operation, as well as for the audio-video files recorded for offline validation/research. The synchronisation should provide lip-sync required by selected monitoring/recognition algorithms.

C. The Risk of Low Reliability

While the proposed approach may have a high number of probable points of failures [42], therefore a less complicated pipeline, being more reliable in unforeseen situations, is strongly suggested. During the research, the authors noted that in the case of unfortunate lighting conditions (e.g. low light, a disability glare), relying too much on video processing results increases

the number of failures in (heavily dependent on video analysis) processing pipeline. Extending the system by additional means to monitor and handle unexpected failures would lead to even bigger complexity, and therefore create subsequent points of failures.

D. Continuous Speech Recognition and On-Demand Delivery Of transcription Results

STT services available on the market, supporting real-time speech-to-text transcription, are designed to operate in a processing sequence of at least three states: (a) data acquisition, (b) data processing (while the interlocutor is supposed to wait idly for a result), and (c) providing the results. The end of the speech (transition from state a to b) is detected by ongoing silence or the limits of the utterances buffer. Then the final result (state c) is delivered (however, it can be preceded by the approximated intermediate one while still being in state a or b). Such architecture conforms to the REPL scheme. The delays in the delivery of transcription results (sum of time passed on states b and c) and the unavailability of the on-demand requests for results (mandatory detection of the end of speech) are additional barriers to overcome.

E. The Risk of Low Performance

Achieving a high level of system responsiveness, which is the foundation of the proposed architecture's advantages compared to the REPL workflow, is crucial, yet challenging. Responsiveness is recognized to be a time-based parameter – represented as the time elapsed from vocalizing the interlocutors' query to synthesizing the robot's response. The lower the value, the better – high responsiveness means keeping the elapsed time minimal, thus delivering an impression of instantaneous response [43]. Due to the aggregated processing time of every component in a pipeline, the response given by the robot may be delayed significantly. From the interlocutor's point of view, the system's overall performance will be seen as lagging or stuck at some point, causing confusion, and perceived as impairment and spoilage of the interaction experience. It is hard to define the typical/universal threshold value of responsiveness in seconds, because such a factor is a matter of subjective reception of the overall impressions of the interaction and the robot. However, to reach the goal of mimicking a truly natural conversation as far as possible by all available means, and to go beyond the REPL pattern, the responsiveness should be comparable to that of a human being in conversation.

Considering the roadmap mentioned earlier, facilitating the improvements implementing the proposed "smart beamforming" paradigm is expected to increase the software structure complexity. However, it is necessary to break free from the REPL workflow scheme into a more natural workflow for humans and increase overall performance in multi-talker scenarios. A robust and widely used REPL scheme is intended for single-talker scenarios. It is semi-interactive in design as a processing phase consumes some time, and signaling the end of the speech is done by detecting silence or reaching the bounds of utterance buffer. Such interaction is perceived as being similar to using walkie-talkie – a non-overlapping sequential exchange of questions and answers with the notorious in-between delay.

The video and audio sensor fusion concept is key to the reliable utilization of beamforming by robotic systems. Processing and synthesizing data from multiple channels and sources is key to achieving sufficient speech separation for every interlocutor in multi-talker scenarios in publicly working robotic systems. Human sensations also result from processing information from multiple channels. The sensors-fusion approach seems natural if we consider the processes of interpreting stimuli carried out by the biological brain of a living organism. Stimulus signals acquired through different sensors (e.g., vision, hearing), retrieved as multimodal information, is combined to use the redundant data necessary to amplify the meaningful information and filter noisy, fuzzy, and defective signals from inputs [44]. An equivalent combination of multimodal data streams has been incorporated in the most delinquent neural networks model utilized for speech enhancements [45].

Based on the presented results, it may be concluded that in favorable conditions of video and audio processing, the experimental pipeline of "smart beamforming" is able to deliver performance levels that expect to exceed the standard pipeline results.

The experimental research proved that combining video processing and beamforming in the proposed framework can support multiple interlocutors simultaneously speaking in conversation, as shown in Tables III and IV. However, it is still required to put efforts into the extensive development of multi-instance STT/NLP components and reliable mechanisms for matching interlocutors with internal context-holding conversation data attributed to the correct addressees.

V. Future Work

The results obtained so far, encourage the authors to explore the topic of beamforming further. The following research areas seem to be specifically important: methods of beamforming signals from microphone arrays, methods of localizing sound sources, methods of generating unique audio-visual imprints of interlocutors (enabling precise identification of the sound source among all the signals of the robotic station environment), and stable real time vision module operation. Another critical topic is providing reliable results of video processing (under any circumstances) or the temporal synchronization of results delivered continuously, though with jitter.

During the research, numerous research challenges and technical barriers emerged. However, the significantly increased complexity of the processing pipeline may also require analysis in the quality/management field, especially to prepare for unforeseen scenarios, where failure/error recovery strategies should be ensured in the framework's scope.

References

[1] T. Sheridan, "Forty-five years of man–machine systems: History and trends," *IFAC Proc. Volumes*, vol. 18, no. 10, pp. 1–9, 1985.

[2] T.-C. Phan, A.-C. Phan, H.-P. Cao, and T.-N. Trieu, "Content-based video Big Data retrieval with extensive features and deep learning," *Appl. Sci.*, vol. 12, no. 13, 2022, Art. no. 6753. [Online]. Available: https://www.mdpi.com/2076-3417/12/13/6753

[3] United Robotics Group, "Talking with PEPPER," Accessed: Oct. 30, 2022. [Online]. Available:https://support.unitedrobotics.group/support/solutions/folders/80000686318

[4] J. Sun et al., "Performance analysis of beamforming algorithm based on compressed sensing," *Appl. Acoust.*, vol. 198, 2022, Art. no. 108987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003682X22003619

[5] X. Wenmeng, B. Changchun, J. Maoshen, and J. Picheral, "Speech enhancement with robust beamforming for spatially overlapped and distributed sources," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 30, pp. 2778–2790, 2022.

[6] E. Jäckel et al., "NegotiAct: Introducing a comprehensive coding scheme to capture temporal interaction patterns in negotiations," *Group Org. Manage.*, 2022, Art. no. 10596011221132600, doi: 10.1177/10596011221132600.

[7] S. Seok, E. Hwang, J. Choi, and Y. Lim, "Cultural differences in indirect speech act use and politeness in human-robot interaction," in *Proc. IEEE/ACM 17th Int. Conf. Hum.-Robot Interact.*, 2022, pp. 1–8.

[8] S. Kumar et al., "Politeness in human–robot interaction: A multi-experiment study with non-humanoid robots," *Int. J. Social Robot.*, vol. 14, no. 8, pp. 1805–1820, 2022.

[9] A. Gardecki and M. Podpora, "Experience from the operation of the pepper humanoid robots," in *Proc. Prog. Appl. Elect. Eng.*, 2017, pp. 1–6.

[10] M. Podpora et al., "Human interaction smart subsystem–extending speech-based human-robot interaction systems with an implementation of external smart sensors," *Sensors*, vol. 20, no. 8, 2020, Art. no. 2376.

[11] A. Gardecki, M. Podpora, R. Beniak, and B. Klin, "The pepper humanoid robot in front desk application," in *Proc. Prog. Appl. Elect. Eng.*, 2018, pp. 1–7.

[12] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind," *IEEE Robot. Automat. Mag.*, vol. 25, no. 3, pp. 40–48, Sep. 2018.

[13] N. Wu, "Restrict foreigners, not robots": Partisan responses to automation threat," *Econ. Politics*, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ecpo.12225

[14] A. Morenville, L. Vermeulen, P. Schaus, and S. Nijssen, "Simultaneous localization and mapping and autonomous navigation on spot robot from Boston dynamics," M.S. thesis, Ecole polytechnique de Louvain, Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium, 2022. [Online]. Available:https://dial.uclouvain.be/memoire/ucl/object/thesis:35688

[15] D. Ahirwar, J. Purohit, V. B. Semwal, S. Gawre, and M. Rajpurohit, "The recent advancements in humanoid robot technology," in *Proc. IEEE Int. Students' Conf. Elect., Electron. Comput. Sci.*, 2022, pp. 1–6.

[16] United Robotics Group, "A robot designed to interact with humans," Accessed: Oct. 30, 2022. [Online]. Available: https://www.aldebaran.com/en/pepper

[17] United Robotics Group, "Can the robot carry objects?," Accessed: Oct. 30, 2022. [Online]. Available:https://support.unitedrobotics.group/support/solutions/articles/80000945766-can-the-robot-carry-objects-

[18] United Robotics Group, "Pepper design outline," Accessed: Oct. 30, 2022. [Online]. Available: 30.10.2022, https://qisdk.softbankrobotics.com/sdk/doc/pepper-sdk/ch6_ux/index_ux.html

[19] M. Marge et al., "Spoken language interaction with robots: Recommendations for future research," *Comput. Speech Lang.*, vol. 71, 2022, Art. no. 101255. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821000620

[20] United Robotics Group, "Pepper's microphones," Accessed: Oct. 30, 2022. [Online]. Available:http://doc.aldebaran.com/2-5/family/pepper_technical/microphone_pep.html

[21] W. Stommel, L. de Rijk, and R. Boumans, "Pepper, what do you mean? miscommunication and repair in robot-led survey interaction," in *Proc. IEEE 31st Int. Conf. Robot Hum. Interactive Commun.*, 2022, pp. 385–392.

[22] United Robotics Group, "Pepper and beamforming," Accessed: Oct. 30, 2022. [Online]. Available: http://doc.aldebaran.com/2-5/naoqi/audio/alsoundlocalization.html

[23] United Robotics Group, "Pepper offer," Accessed: Oct. 30, 2022. [Online]. Available:https://www.aldebaran.com/en/our-offers

[24] D. Veerendra and B. Jalal, "A fast adaptive beamforming technique for efficient direction-of-arrival estimation," *IEEE Sensors J.*, vol. 22, no. 23, pp. 23109–23116, Dec. 2022.

[25] I. Lauriola, A. Lavelli, and F. Aiolli, "An introduction to deep learning in natural language processing: Models, techniques, and tools," *Neurocomputing*, vol. 470, pp. 443–456, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221010997

[26] United Robotics Group, "NAOqi documentation," Accessed: Dec. 31, 2022. [Online]. Available: http://doc.aldebaran.com/2-5/naoqi/

[27] M. Gilles and E. Bevacqua, "A review of virtual assistants' characteristics: Recommendations for designing an optimal human–machine cooperation," *J. Comput. Inf. Sci. Eng.*, vol. 22, no. 5, 2022, Art. no. 050904, doi: 10.1115/1.4053369.

[28] T. Gburrek et al., "Informed vs. blind beamforming in ad-hoc acoustic sensor networks for meeting transcription," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2022, pp. 1–5.

[29] X. Qian, Q. Zhang, G. Guan, and W. Xue, "Deep audio-visual beamforming for speaker localization," *IEEE Signal Process. Lett.*, vol. 29, pp. 1132–1136, 2022.

[30] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Comput. Speech Lang.*, vol. 72, 2022, Art. no. 101317. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821001121

[31] M. Gomez-Barrero et al., "Biometrics in the era of COVID-19: Challenges and opportunities," *IEEE Trans. Technol. Soc.*, vol. 3, no. 4, pp. 307–322, Dec. 2022.

[32] A. Rodrigues et al., "Analyzing the performance of ASR systems: The effects of noise, distance to the device, age and gender," in *Proc. XX Int. Conf. Hum. Comput. Interact.*, New York, NY, USA, 2019, pp. 1–8. [Online]. Available: https://doi.org/10.1145/3335595.3335635

[33] United Robotics Group, "NAOqi documentation–NAOqi ALSpeechRecognition module," Accessed: Oct. 30, 2022. [Online]. Available: http://doc.aldebaran.com/2-5/naoqi/audio/

[34] Mozilla, Project DeepSpeech," Accessed: Oct. 30, 2022. [Online]. Available:https://github.com/mozilla/DeepSpeech

[35] OpenAI, "Project whisper," Accessed: Oct. 30, 2022. [Online]. Available:https://github.com/openai/whisper

[36] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2003, pp. 577–582.

[37] X. He, L. Deng, and A. Acero, "Why word error rate is not a good metric for speech recognizer training for the speech translation task?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 5632–5635.

[38] M. Podpora et al., "LevenPhrase: A bi–space approach for assessment of the distance of two phrases for interaction applications," in review.

[39] M. Zhang and J. Li, "A commentary of GPT-3 in MIT technology review 2021," *Fundam. Res.*, vol. 1, no. 6, pp. 831–833, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667325821002193

[40] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[41] S. Zheng, W. Huang, X. Wang, H. Suo, J. Feng, and Z. Yan, "A real-time speaker diarization system based on spatial spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7208–7212.

[42] Y. Zeng et al., "An analytical method for reliability analysis of hardware-software co-design system," *Qual. Rel. Eng. Int.*, vol. 35, no. 1, pp. 165–178, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/qre.2389

[43] M. Weideman, *Website Visibility: The Theory and Pract. of Improving Rankings* (ser. Chandos Internet series Website visibility). Amsterdam, The Netherlands: Elsevier Science, 2009. [Online]. Available: https://books.google.pl/books?id=sqCjAgAAQBAJ

[44] B. Wahn et al., "When eyes beat lips: Speaker gaze affects audiovisual integration in the McGurk illusion," *Psychol. Res.*, vol. 86, no. 6, pp. 1930–1943, Sep. 2022. doi: 10.1007/s00426-021-01618-y.

[45] L. A. Passos, J. P. Papa, J. Del Ser, A. Hussain, and A. Adeel, "Multimodal audio-visual information fusion using canonical-correlated graph neural network for energy-efficient speech enhancement," *Inf. Fusion*, vol. 90, pp. 1–11, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253522001385