

Human-robot conversation with multiple participants in noisy public spaces

Divesh Lala^{1,2} Yogeeswaran Muthukumaran³ Vincent Fernandes^{2,4} Kazushi Kato¹ Shota Fujiki²
Zihao Chi² Masaya Iwasaki² Taiken Shintani² Megumi Kawata² Kazuki Sakai² Koji Inoue¹
Yuichihiro Yoshikawa² and Tatsuya Kawahara¹

Abstract—For noisy real-world environments such as those in open public spaces, spoken dialogue systems for both autonomous robots and avatars should be carefully designed to provide enhanced speech signals. These signals can be used either for speech recognition or, in the case of an avatar system, transmitted as clean speech to a remote operator. This work proposes an audio system that can be used for both these scenarios and was demonstrated as a proof-of-concept at the 2025 World Expo in Osaka. The first scenario is an attentive listening system with the android ERICA, and the second is a conversation support system with mobile Teleco robots, with one of them acting as an avatar for a remote operator. Both systems feature multi-party conversation and use a single multi-channel microphone array. We describe how our audio system not only enhances the speech of multiple speakers in a noisy environment, but provides a form of spatial audio which allows for more immersiveness in avatar-based conversational interactions.

I. INTRODUCTION

Large language models (LLMs) have made natural conversation between humans and artificial intelligence ubiquitous for one-to-one dialogue using text as the primary modality. These systems have largely solved the problem of ensuring coherent and human-like conversation through text. On the other hand, when interacting with physically embodied social robots, humans often use speech as the primary modality. The intuitive solution is to integrate LLMs into these robots by using results of automatic speech recognition (ASR) as an input to an LLM. Although LLMs are largely trained on textual data rather than spoken dialogue, this method is reasonable for free-form conversation.

However, there are several settings which are still yet to be addressed by this framework. The first is the use of systems in uncontrolled, public environments. Conversational robots are often tested in controlled lab settings, where there is little background noise to affect speech recognition. Both user and robot may be seated in front of each other to accomplish a pre-defined task with a clear communication channel. On the other hand, public settings require the robot to converse with a diverse range of people who may interact with the robot in different ways. For example, children require dialogue to

be much simpler, while elderly users may speak slower or repeat themselves.

Another setting in which LLMs are not yet optimized for is multi-party conversation (involving three or more participants, either robot or human). The dynamics of multi-party conversation are much different than dyadic conversation [1]. Utterances from speakers have a greater chance of overlapping or interrupting, and turn-taking in these environments is more complex. Situated multi-party conversation is common in everyday life and social robots should also be able to accommodate this scenario.

Future multi-party interactions may also include cybernetic avatars (CAs) - physically situated robots or virtual agents which are teleoperated by humans [2]. CAs have been implemented as a means for humans to interact from remote locations, and allow them to engage in social activities as well as provide services [3], [4]. These include participation in multi-party conversations where the interaction becomes immersive. A noisy environment provides a challenge for CAs. Humans in the real world are able to conduct conversations in such environments by filtering out background noise and focusing only on particular stimuli (speech during conversation). This is known as the “cocktail party” effect [5] and has been the subject of much auditory research [6], [7]. CAs need to replicate this in settings with high background noise, or else conversation becomes impossible.

The basic requirement for all multi-party settings with a high noise level is a robust audio system. To preserve the naturalness of the interaction, it is desirable that participants do not need to hold a microphone and can speak in a natural manner. One common solution is to use a single multi-channel microphone array through which the speech of multiple participants can be extracted. However, this array alone cannot be used for speech recognition in a noisy environment due to the high level of noise from all directions and the inability to easily separate sources of speech. For CAs, a microphone array makes it very difficult for a remote operator to filter out background noise and focus on the speech of a particular person. This presents a technical challenge in this domain which has not yet been fully addressed in spoken dialogue system research.

Considering the above objectives, we present an audio system that can be used for human-robot multi-party dialogue scenarios in environments with a high level of noise. We describe two different human-robot dialogue scenarios using

¹Kyoto University Graduate School of Informatics

²Osaka University, Graduate School of Engineering Science

³National University of Singapore

⁴ENSEA

Corresponding author: lala@i.kyoto-u.ac.jp

this system which were demonstrated at the 2025 World Expo in Osaka, Japan. Each dialogue scenario involves three participants. The first scenario is two seated human participants engaging in attentive listening conversation with an android robot, ERICA. The second scenario is a conversation with one standing human and two mobile Teleco robots. One robot is tele-operated by another human as a CA, while the other is fully autonomous. The robots and systems used in this work are implemented in the Japanese language.

II. RELATED WORK

Multi-party conversation with robots is increasingly becoming a focus of research [1], particularly as dyadic conversation is well handled by LLMs. Speaker detection is a specific task for multi-party conversation, and one common strategy is to simply have each speaker use separate microphones [8]. For hands-free interaction, microphone arrays are often used, either as external devices [9], [10] or integrated into the robot itself [11], [12], and these require techniques to extract the speech of the current speaker.

The management of multi-party interactions with robots should also be extended to public interactions [13]. Although there are works on service robots for use in public, these are situations where the robot and humans are stationary or designed for dyadic conversation [14], [15]. Our intention is to allow the robot to perform audio processing tasks while moving in a public space. Such systems with mobile robots have been attempted [16] but are still relatively rare and do not handle multi-party interaction. Conversely, simultaneous speaker recognition has been integrated into a mobile robot [17], [18], but this was before the development of LLMs and it is unknown if free-form conversation could be handled appropriately.

One recent work integrated multi-party conversation with LLMs in a hospital environment [19], [20], although it is unknown if the robot was mobile. Speaker identification was achieved through a visual modality, but speakers could not talk simultaneously. Other work also used speaker separation through a microphone array for psychiatric care [21], but this was in a controlled environment. Our work brings together the above research areas to tackle multi-party human-robot conversation in public which can accommodate mobile robots. To our knowledge such a system has not been addressed in previous work.

III. AUDIO SYSTEM

Our audio architecture is based on the use of one low-cost ReSpeaker circular microphone array with four channels. Audio from each channel can be continuously captured, but are not completely separated. There is significant audio bleeding from other channels when a person speaks, making simultaneous speech recognition difficult. This microphone array can be used as a stand-alone device or built into the hardware of the robot itself.

The goal of this system is to be able to reliably extract clean speech signals of multiple people from the microphone array in a noisy environment. Once this has been achieved,

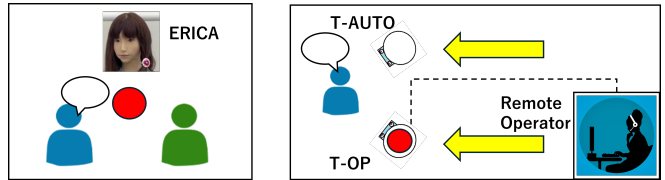


Fig. 1. Overview of the demonstrations for attentive listening ERICA (left) and conversational support Telecos (right). The red circle indicates the location of the microphone array.

we can then use it for multiple purposes. Firstly, it can be used as a cleaner signal for speech recognition. Secondly, it can be used as an audio signal for the remote operator in a CA system.

To achieve these goals we first make use of a system described in [22]. In this system, a multi-channel microphone array source is separated into multiple enhanced audio channels. The listening direction of these channels on the microphone array can also be defined - it is not necessary for the direction of the channels to match the orientation of the individual microphones. This model was trained to predict and enhance voice activity even with a large amount of background noise. From this system we receive enhanced audio from multiple channels simultaneously.

For automatic speech recognition we use a low-latency Japanese incremental speech recognition model which can run locally. The model was also trained noisy speech, such as YouTube videos [23]. The model is lightweight enough that multiple models can run simultaneously on one computer. The result is an audio system which provides noise-robust incremental speech recognition from multiple people. We now describe the two different dialogue scenarios where this system was used.

IV. DIALOGUE SCENARIOS

To test the effectiveness of the system for multi-party conversation, we used it in two dialogue scenarios, each having a different set of requirements. Fig. 1 shows a general overview of the systems and Table I provides a comparison of the major differences.

A. Attentive Listening with ERICA

ERICA is an android robot who can produce realistic facial expressions, speech and upper body gestures. She has been used for several dialogue experiments [24], [25], [26]. In this scenario we use ERICA in a multi-party interaction with two human participants. Fig. 2 shows a photo of ERICA during an interaction. The microphone array is placed on a tripod between the participants to capture their speech.

The first part of the dialogue scenario is attentive listening based on previous work [27]. ERICA asks each participant one at a time to act as the main speaker and chat about a light topic. ERICA acts as an attentive listener to the main speaker and responds to the main speaker's talk by using verbal backchannels, nods and short responses. Backchannels and nods are generated by a trained model based on voice

TABLE I
COMPARISON OF MULTI-PARTY DIALOGUE SCENARIOS

	ERICA	Teleco
Task	Attentive Listening	Conversation Support
Participants	1 robot, 2 humans	1 autonomous robot, 1 tele-operated robot, 1 human
Microphone position	Static, between participants	Dynamic, on top of robot
Participant positions	Static	Moving
Speech recognition	Multi-channel, simultaneous	One channel



Fig. 2. Android ERICA used in the attentive listening demonstration.

activity projection [28]. Short responses are simple elaborating questions (e.g. “What kind of ramen did you eat?”) or empathetic responses (e.g. “Ah, that’s great!”) which are generated by GPT 4.1. Each participant has approximately one minute of attentive listening with ERICA.

The next phase of the dialogue involves ERICA asking how each participant felt about each other’s talks. ERICA uses GPT 4.1 to extract important points from each talk and provides an in-depth question to start the discussion. In this phase participants speak back and forth and reflect on the topics brought up in their talks. Finally, ERICA summarizes the discussion to end the interaction. An example of this is shown below, paraphrased from a real interaction at the Expo between ERICA, a woman, and her mother:

(Attentive listening phase)

ERICA: “What are you looking forward to at today’s Expo?”

A (child): “I’m actually pretty tired, it’s a hot day and I had to line up a lot...”

B (mother): “Lucky that you could bring me here when it was raining, thanks...”

(Discussion phase)

ERICA to A: “You could find this exhibit interesting, thanks to the rain. What did you think about what your mother said?”

A (child): “She goes with me everywhere, and I don’t usually say this to her, but she’s a great mother.”

Both participants may speak to each other or to ERICA at any time. Therefore, continuous and simultaneous speech

recognition is essential. In order for ERICA to generate thoughtful questions, the system must be able to separate the speech of individual speakers so that their dialogue histories can be used as input to the LLM. Turn-taking and backchannelling in this system uses a voice activity projection (VAP) model, which continuously predicts if ERICA should take the turn from the current speaker [29]. This provides smooth, human-like conversation flow and requires that audio channels of both speakers are well enhanced to prevent interruptions.

The physical setup has both ERICA and the participants seated in a triangular formation with the participants approximately 30° to the left and right of ERICA. The microphone is in a fixed position in the center of the conversation. Enhanced audio channels are directed towards each of the participants’ positions to fully capture their speech. The static positioning of participants also means ERICA can do gaze behavior even without needing any visual information.

B. Conversation Support with Teleco

Teleco is a small robot which sits on a mobile rover, and has been used in previous research related to mobile robots [30], [31], [32]. Navigation is done autonomously through the use of SLAM [33] which runs inside a small computer located inside the rover. This computer also controls the movement and other behaviors of the Teleco itself. The face of Teleco is displayed on a screen. It can be made to say text-to-speech utterances by playing audio through a small speaker array below the robot. However, for our demonstration we use the speaker of the computer inside the rover to project its voice. This is because it can project at a much higher volume, which is critical when we consider the noise inside the environment. The Teleco robots are shown in Fig. 3.

For our system, there are two Teleco robots. One is controlled by a remote operator (referred to as **T-OP**), and is a robot avatar. The other robot behaves autonomously (referred to as **T-AUTO**). A camera on the chest of **T-OP** streams real-time video to a remote operator, and their voice is played through the rover’s speaker with real-time mouth synchronization. **T-OP**’s system uses the standard WebRTC protocol to transmit video and audio streams.

The operator of **T-OP** is situated in a separate location and interacts using a computer. The GUI of the system displays **T-OP**’s camera feed and through headphones they listen to audio picked up from the microphone array on **T-OP**’s head. The GUI also allows the operator to manipulate **T-OP** by

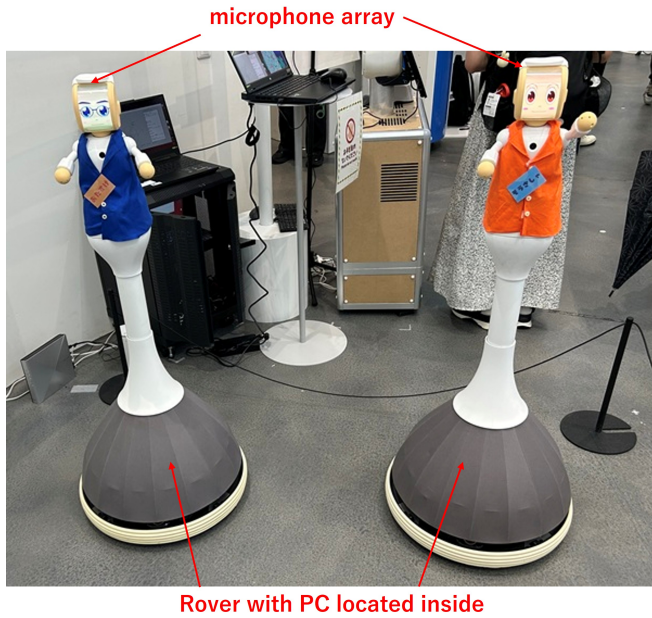


Fig. 3. The Teleco robots used in the public demonstration. The left Teleco in the blue vest is autonomous, while the right Teleco in the orange vest is tele-operated by a human operator.



Fig. 4. Screenshot of operator's GUI, with the user on the left and **T-AUTO** on the right. Green buttons at the top of the GUI are for sending commands to **T-AUTO**. The scenario is designed so that both participants are displayed within the camera frame.

raising or lowering the pole to adjust its height. **T-OP** can also be rotated left and right to enable the operator to change the viewpoint of their avatar. Due to safety issues, we did not allow the operator to move **T-OP** forward. The other major component of the GUI are buttons to send commands to **T-AUTO**. A screenshot of the GUI is displayed in Fig. 4.

T-AUTO is a Teleco robot which behaves autonomously. Although it also contains a camera and microphone array they are not used in this demonstration. **T-AUTO** receives commands from the tele-operator indicating when they should generate and say an utterance. The role of **T-AUTO** is to act as a conversational support robot for the operator during the conversation.

The dialogue scenario involves one human participant, **T-OP** and **T-AUTO**. Initially both robots are situated about

four meters away from the human participant. The operator begins the dialogue scenario by pushing a button on the interface. Both robots then approach the human and form a triangular F-formation indicating a group conversation. The human and the operator (who is controlling **T-OP**) then converse about a casual topic which has been selected from a fixed list beforehand (e.g. “What animal do you think is best to keep as a pet?”). Although the initial topic is chosen in advance, the actual flow of conversation is unrestricted and the participants are free to move to other topics if they wish.

During the conversation, the operator of **T-OP** has an option to push one of three buttons requesting conversation support from **T-AUTO**. These buttons have the following functions:

- Ask (the human participant) for more details
- Suggest something different
- Back me (the operator) up

Once a button is pushed, **T-AUTO** will produce an utterance based on what type of request was given. This utterance is generated by GPT 4.1 with the input being the entire dialogue history as well as the type of request. As soon as it is an appropriate time to speak, **T-AUTO** moves towards the user and says their utterance before returning to its initial position to preserve the formation. The operator may request support from **T-AUTO** multiple times during the interaction. Each interaction lasts for approximately 5 minutes. At the end of the interaction, **T-AUTO** summarizes the dialogue and says a farewell message. Both robots return to their initial positions. An example of dialogue in this scenario is given below:

T-OP: “Dogs are really cute, right?”

User: “But dogs take a lot of work, you have to walk them every day...”

(operator pushes “Back me up” button)

T-AUTO: “I agree that dogs are cute. Also, taking them for walks is good exercise. If they are well trained they will bond with you.”

T-OP: “That’s right!”

The microphone array used to capture the speech of human participant is fixed on top of **T-OP**’s head. The speech of the operator is also captured from a close-talk microphone. The speech recognition results of the operator, the human participant, and **T-AUTO**’s dialogue are used in the prompt to generate **T-AUTO**’s response.

This scenario also uses the same backchannelling and turn-taking model as ERICA’s scenario. For gaze behavior, the system identifies the current speaker and both Telecos direct their gaze towards that participant (either the user or the other Teleco). This means **T-OP**’s microphone array will change orientations as **T-OP**’s head and body turns. The next section describes how the system handles this situation.

V. REQUIREMENTS FOR MOBILE TELECO AVATAR SYSTEM

Although the base audio system is identical for both scenarios, there are several extensions which need to be implemented for the Teleco scenario, due to its mobile nature and the perceptual requirements of the remote operator.

Unlike ERICA, where the microphone array is in a fixed position, the microphone is directly on top of **T-OP**'s head. Consequently, when the operator speaks, their voice (transmitted through the speaker inside **T-OP**'s rover) would automatically be picked up by the speech recognition system, since the sound source is loud and close to all channels. We implemented echo cancellation of the microphone array by using specialized software¹. This proved to be effective for removing the operator's voice from the received audio.

In the Teleco scenario, participants may change their position and orientation during the interaction, so this must be tracked. The rovers for both Telecos track their own positions inside the environment using SLAM. The orientation of a Teleco is calculated by summing the yaw values of its rover, waist joint, and neck joint. The human is tracked by a depth camera connected to a YOLO vision model [34] placed approximately two meters high and observing the whole environment. We found that the pre-trained YOLO model often misclassified Teleco robots as people, so this was fine-tuned with data collected in a training environment. Therefore we can gather continuous information on the positions and orientations of both Telecos and the position of the human.

A. Audio channel selection

The major task for the Teleco scenario is to accommodate the changing position and orientation of the microphone array on **T-OP**, which is responsible for speech enhancement and user speech recognition. In ERICA's static setting we can continuously listen on the same array channels because this will always be directed towards seated participants. For Teleco's scenario we have an extra task to determine this "listening" channel, which will change with the Teleco's position and orientation.

Our system so far has described the integration of existing components. We possess a system which both enhances and separates multiple audio channels using a low-cost consumer-grade microphone array. Our novel contribution in this work is the extension of this to a mobile robot, which requires the ability to select the audio channel(s) to listen to. This contribution can benefit similar types of mobile robot systems which need to operate in a dynamic and unpredictable environment. For our implementation the positions of all participants in the environment must be tracked.

Let the human position and **T-OP**'s position be (h_x, h_y) and (t_x, t_y) , respectively. We also denote four audio channels as $A_1 \dots A_4$. The value of each of these audio channels is the principal listening direction in degrees, d , where it is focused for speech enhancement. This d changes every time

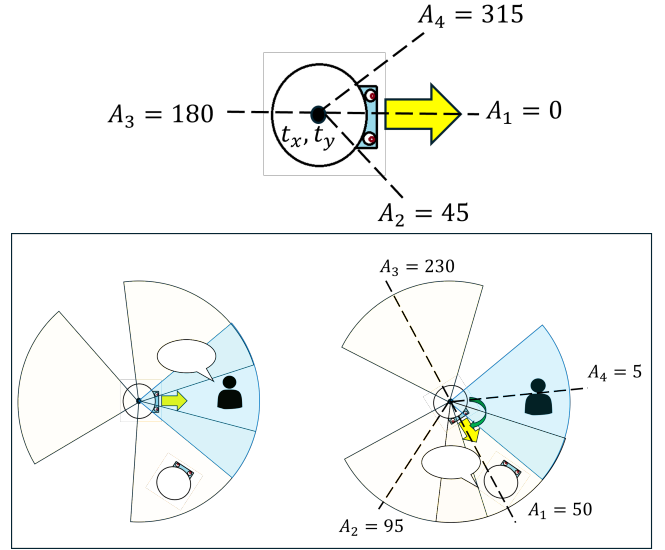


Fig. 5. Diagram of how the audio channel is selected for speech recognition. The top figure shows the orientations of the audio channels and the facing direction of **T-OP** (yellow arrow). The bottom-left figure shows the sectors of each channel and the selected channel A_1 (blue) when a human enters the sector. The bottom-right figure shows how the selected channel changes to A_4 when **T-OP** changes its orientation 50° to the right to face **T-AUTO**.

the orientation of the **T-OP** is changed. Assume that the initial facing direction of **T-OP** is 0. The initial d values for $A_1 \dots A_4$ are 0, 45, 180, and 315, respectively. In other words, **T-OP** will enhance speech mainly from the front and 45° to its left and right, as well as directly behind itself.

For each A_i , the system creates a sector centered at (t_x, t_y) , with a central angle of 60°, a radius of 2 meters, with a orientation of d . The central angle and radius are both adjustable in real-time. The system continuously identifies the sector which contains the closest human to the Teleco. If there is no user in the vicinity of **T-OP** then no listening is done. Pseudocode for this is shown in Algorithm 1.

Sectors for the audio channels overlap, so it is possible that the human is contained in more than one sector. In this case, the algorithm keeps the current listening channel or selects the first candidate in the case there is no current listening channel.

Now we have a method that can extract speech from a person situated in the environment. This can also be extended to multiple speakers if they are in different listening channels. For a scenario with a CA, the enhanced signal is also transmitted to the remote operator. It enables the operator to clearly hear the user's speech. This also functions correctly even if the user moves while they speak.

For further immersion, spatial audio for the avatar should be implemented. For example, if someone is speaking on the left of the robot, this signal should be heard by the operator on the left side of their headphones. Human tracking allows us to produce this effect. When the audio signal is transmitted to the operator, it also sends the relative position of the user. The audio is then panned appropriately to simulate the spatiality of the sound.

¹<https://github.com/voice-engine/ec>

Algorithm 1 Audio channel selection algorithm

```
1:  $c \leftarrow$  current listening channel
2: speech recognition channels  $\leftarrow []$ 
3: listening channels  $\leftarrow []$ 
4: for all  $A$  in audio channels do
5:    $d \leftarrow$  orientation of  $A$ 
6:    $s(\theta, C, d) \leftarrow$  sector where  $C = (t_x, t_y), \theta = 60$ 
7:   for all  $h$  in humans do
8:     if  $s$  contains  $(h_x, h_y)$  then
9:       add  $A$  to speech recognition channels
10:      add  $A$  to listening channels
11:    end if
12:  end for
13:  for all  $r$  in other robots do
14:    if  $s$  contains  $(r_x, r_y)$  then
15:      add  $A$  to listening channels
16:    end if
17:  end for
18: end for
19: if speech recognition channels is empty then
20:   current listening channel  $\leftarrow None$ 
21: else if  $c$  not in speech recognition channels then
22:    $n \leftarrow$  first entry in speech recognition channels
23:   current listening channel  $\leftarrow n$ 
24: end if
```

We also extend this to multiple speakers. When **T-AUTO** speaks, the operator hears their voice emanating from the relative direction of the robot compared the avatar **T-OP**. The effect of this is that the operator hears two voices clearly, each coming from the appropriate direction. This arguably gives the operator of a CA more of a sense of being in the conversation and preserving the audio-spatial relationship between participants, without having to contest with high background noise.

VI. SUPPORTING THE SENSE OF AGENCY

In this work we have described our system which is intended to enable speech recognition from multiple users in highly noisy environments. We also propose that the Teleco system in particular can support the sense of agency of the operator.

Firstly, the operator is provided agency by allowing them to control a robotic avatar (**T-OP**). The operator can make **T-OP** speak and change their viewpoint. The immersiveness of their interaction is supported by the spatially enhanced audio system. The goal is to reduce the cognitive load on the operator in terms of trying to listen to a conversation in a noisy environment. While humans in a situated interaction can naturally benefit from the “cocktail party effect” to focus on their conversation partner, this is more difficult in a remote interaction with constant audio from all directions. By using our system, our intention is to try and simulate this effect to increase the feeling of presence and agency. For example, turning **T-OP** also changes the relative direction of speech and allows some form of audio agency in the interaction.

Furthermore, the inclusion of the autonomous robot **T-AUTO** adds an extra dynamic to the interaction. The operator is able to not only control their own avatar but also influence the actions of **T-AUTO** which acts as a support robot. It should be noted that the operator does not control all aspects of **T-AUTO**. It would be very difficult to maintain a multi-party conversation while being fully in control of two or more participants. Therefore we frame this as a form of shared agency. The operator’s control of the autonomous robot only extends to their general request for support, while the robot itself is responsible for conversational behaviors such as response generation and turn-taking. This makes it possible for the operator to control other robots within the conversation space while maintaining an acceptable cognitive load.

VII. RESULTS

We now describe the actual testing and implementation of the proposed system at the 2025 World Expo in Osaka, Japan. Note that due to privacy issues, we could not collect data at the expo itself, so no formal subjective evaluation was conducted. However we did conduct pilot testing of the speech recognition functionality in a controlled laboratory environment.

A. Pilot testing

The speech recognition tests involved the same user speaking under several conditions. These conditions were:

- The user holding a hand-held microphone (no robot)
- The user speaking at a stationary Teleco
- The user speaking at a Teleco which was constantly rotating back and forth
- The user speaking at a Teleco with a large amount of simulated background noise coming from speakers

For conditions which involved a Teleco robot, the user spoke from a distance of approximately one meter, approximately the same that for casual conversation. We then measured the character error rate (CER) of the system under these conditions. Our findings are presented in Table II.

TABLE II
CER OF SPEECH RECOGNITION UNDER VARIOUS CONDITIONS

Condition	CER (%)
Hand-held microphone	7.8
Stationary Teleco	6.7
Moving Teleco	6.7
Teleco in noisy environment	7.1

We found that the CER of the speech recognition system is reasonably well-performing, and comparable to a hand-held microphone. Although we just used one user for these pilot tests, we also found similar performance for other users.

B. Live expo demonstration

The systems were run simultaneously in the same building (pavilion) for six days during the month of August. The Expo itself had many visitors and members of the

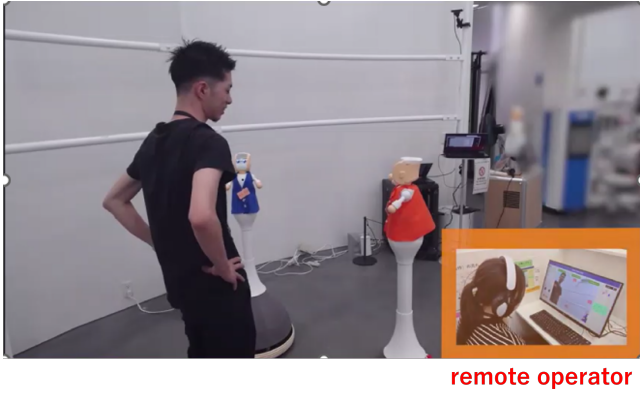


Fig. 6. A participant engaging in a multi-party conversation with the two Teleco robots.

public were invited to use the systems. The pavilion also contained multiple other robot demonstrations. Across six days, approximately 7300 people entered the pavilion. There were 513 interactions with ERICA (both pairs of people and individuals), however we did not log the number of interactions with the Teleco robots.

The pavilion environment contained a large amount of reverberation and high noise levels, with thousands of members of the public entering the pavilion on a daily basis. Background speech from other exhibits and occasionally heavy rainfall would create extra noise inside the pavilion. Our audio system was necessary to deal with this extreme environment.

Our system was able to successfully recognize conversational speech in the majority of cases. Coherent and relevant responses from both ERICA and Teleco showed that conversation flow could be maintained in both scenarios. Monitoring of Teleco’s audio system showed that selection of the appropriate audio channel based on the position and orientation of participants was successful. Several participants who took the role of the operator remarked that they could clearly hear the conversation despite the high level of noise. This shows that using enhanced audio channels to stream to operators rather than raw audio should be considered in similar public spaces.

C. Limitations

While the implementation was successful, there is room to improve the system. One major issue is that the dialogue system itself is not truly multi-party. In attentive listening, the participants interact with ERICA one at a time, while in the conversational support scenario the remote operator explicitly commands the robot to interject during the conversation. Multi-party conversation requires the autonomous robot to be able to engage in smooth turn-taking, in particular addressee detection. There have been attempts to use LLMs to solve this [35], [20], however this would probably require a multimodal approach using vision models.

Speech recognition was generally acceptable, although we observed situations, particularly in the conversational support

scenario, when the user’s speech was not picked up by the microphone array. We identified two main sources of this problem. The first is that the speaker’s voice was too quiet to be recognized. The second issue was that some users would turn and talk to **T-AUTO**. This is a natural behavior, but it also meant their voice was not able to be processed correctly. Since **T-AUTO** would only produce dialogue when instructed to by the operator, it would remain silent and the conversation would stagnate.

VIII. CONCLUSION

This work described two multi-party conversational robot systems with different settings, but using audio processing based on the same microphone array system. The settings varied in terms of the distribution of humans and robots, the amount of mobility, and the location of the microphone. We showed that we could accommodate both these scenarios through speech enhancement and appropriate selection of microphone array audio channels. We demonstrated both systems as proofs of concept at the 2025 World Expo in Osaka, Japan and found that they were robust to noise and could facilitate multi-party conversation.

ACKNOWLEDGMENT

This work was supported by JST Moonshot R&D JP-MJPS2011.

REFERENCES

- [1] S. Fang, X. Liu, T. Igarashi, and K. Yatani, “Unraveling multi-party conversations: From human interaction mechanisms to conversational agent challenges and persona design,” *International Journal of Human-Computer Studies*, p. 103719, 2025.
- [2] H. Ishiguro, F. Ueno, and E. Tachibana, *Cybernetic avatar*. Springer, 2025.
- [3] R. Cisneros-Limón, A. Dallard, M. Benallegue, K. Kaneko, H. Kamimaga, P. Gergondet, A. Tanguy, R. P. Singh, L. Sun, Y. Chen, *et al.*, “A cybernetic avatar system to embody human telepresence for connectivity, exploration, and skill transfer,” *International Journal of Social Robotics*, vol. 17, no. 3, pp. 535–562, 2025.
- [4] Y. Horikawa, T. Miyashita, A. Utsumi, S. Nishimura, and S. Koizumi, “Cybernetic avatar platform for supporting social activities of all people,” in *2023 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2023, pp. 1–4.
- [5] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *Journal of the acoustical society of America*, vol. 25, pp. 975–979, 1953.
- [6] S. Haykin and Z. Chen, “The cocktail party problem,” *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.
- [7] A. R. Conway, N. Cowan, and M. F. Bunting, “The cocktail party phenomenon revisited: The importance of working memory capacity,” *Psychonomic bulletin & review*, vol. 8, no. 2, pp. 331–335, 2001.
- [8] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, “Furhat: a back-projected human-like robot head for multiparty human-machine interaction,” in *Cognitive behavioural systems: COST 2102 international training school, dresden, Germany, february 21-26, 2011, revised selected papers*. Springer, 2012, pp. 114–130.
- [9] G. A. Abbo, M. J. Pinto-Bernal, M. Catrycke, and T. Belpaeme, “Fast multi-party open-ended conversation with a social robot,” *arXiv preprint arXiv:2503.15496*, 2025.
- [10] M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. P. Petrick, “Two people walk into a bar: Dynamic multi-party social interaction with a robot agent,” in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012, pp. 3–10.
- [11] I. Nishimuta, K. Itoyama, K. Yoshii, and H. G. Okuno, “Toward a quizmaster robot for speech-based multiparty interaction,” *Advanced Robotics*, vol. 29, no. 18, pp. 1205–1219, 2015.

- [12] R. Gomez, T. Kawahara, K. Nakamura, and K. Nakadai, "Multi-party human-robot interaction with distant-talking speech recognition," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 439–446.
- [13] A. Müller and A. Richert, "No one is an island-investigating the need for social robots (and researchers) to handle multi-party interactions in public spaces," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 1772–1777.
- [14] E. J. Hwang, B. K. Ahn, J. Y. Lim, B. A. Macdonald, and H. S. Ahn, "Robot dialog system in the context of hospital receptionist and its demonstration," *International Journal of Social Robotics*, vol. 15, no. 4, pp. 679–687, 2023.
- [15] N. Gunson, D. H. Garcia, W. Siefińska, A. Addlesee, C. Dondrup, O. Lemon, J. L. Part, and Y. Yu, "A visually-aware conversational robot receptionist," in *Proceedings of the 23rd annual meeting of the special interest group on discourse and dialogue*, 2022, pp. 645–648.
- [16] A. Yuguchi, S. Kawano, K. Yoshino, C. T. Ishi, Y. Kawanishi, Y. Nakamura, T. Minato, Y. Saito, and M. Minoh, "Butsukusa: A conversational mobile robot describing its own observations and internal states," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 1114–1118.
- [17] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system 'HARK' - open source software for listening to three simultaneous speakers," *Advanced Robotics*, vol. 24, no. 5-6, pp. 739–761, 2010.
- [18] J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H. G. Okuno, "Robust recognition of simultaneous speech by a mobile robot," *IEEE Transactions on Robotics*, vol. 23, no. 4, pp. 742–752, 2007.
- [19] A. Addlesee, N. Cherakara, N. Nelson, D. Hernández García, N. Gunson, W. Siefińska, M. Romeo, C. Dondrup, and O. Lemon, "A multi-party conversational social robot using llms," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 1273–1275.
- [20] A. Addlesee, N. Cherakara, N. Nelson, D. H. Garcia, N. Gunson, W. Siefińska, C. Dondrup, and O. Lemon, "Multi-party multimodal conversations between patients, their companions, and a social robot in a hospital memory clinic," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 62–70.
- [21] K. Ochi, D. Lala, K. Inoue, T. Kawahara, and H. Kumazaki, "Robot-mediated multi-party conversation aimed at affect improvement for psychiatric patients," *IEEE Transactions on Affective Computing*, 2025.
- [22] Y. Ishikawa, K. Konaka, T. Nakamura, N. Takamune, and H. Saruwatari, "Real-time speech extraction using spatially regularized independent low-rank matrix analysis and rank-constrained spatial covariance matrix estimation," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 730–734.
- [23] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, "JTubeSpeech: corpus of Japanese speech collected from YouTube for speech recognition and speaker verification," *arXiv preprint arXiv:2112.09323*, 2021.
- [24] E. Ishii, G. I. Winata, S. Cahyawijaya, D. Lala, T. Kawahara, and P. Fung, "ERICA: An empathetic android companion for COVID-19 quarantine," in *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue*, 2021, pp. 257–260.
- [25] K. Inoue, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara, "An attentive listening system with android ERICA: Comparison of autonomous and woz interactions," in *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*, 2020, pp. 118–127.
- [26] K. Inoue, K. Hara, D. Lala, S. Nakamura, K. Takanashi, and T. Kawahara, "A job interview dialogue system with autonomous android erica," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*. Springer, 2021, pp. 291–297.
- [27] D. Lala, P. Milhorat, K. Inoue, M. Ishida, K. Takanashi, and T. Kawahara, "Attentive listening system with backchanneling, response generation and flexible turn-taking," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 127–136.
- [28] K. Inoue, D. Lala, G. Skantze, and T. Kawahara, "Yeah, un, oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 7171–7181.
- [29] K. Inoue, Y. Okafuji, J. Baba, Y. Ohira, K. Hyodo, and T. Kawahara, "A noise-robust turn-taking system for real-world dialogue robots: A field experiment," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 874–879.
- [30] M. Iwasaki, K. Masuda, K. Sakai, M. Ban, Z. Chi, M. Kawata, A. Meneses, H. Ishiguro, and Y. Yoshikawa, "Following to talk: Effects of mobile robot guidance on dialogue motivation in a field trial," *IEEE Robotics and Automation Letters*, 2026.
- [31] A. Guo, S. Mochizuki, S. Yamashita, H. Kenya, J. Baba, and R. Higashinaka, "A comparative study of human-operated and AI-driven guidance with a teleoperated mobile robot," in *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2025, pp. 1615–1627.
- [32] M. Iwasaki, Z. Chi, K. Masuda, A. Meneses, K. Sakai, M. Kawata, H. Ishiguro, and Y. Yoshikawa, "Hospitable guide robot: Demonstrating the impact of vertical oscillation and looking back motion," in *Proceedings of the 11th International Conference on Human-Agent Interaction*, 2023, pp. 257–263.
- [33] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.
- [34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [35] L. Hilgert and J. Niehues, "Next speaker prediction for multi-speaker dialogue with large language models," in *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, 2025, pp. 60–71.