Improving Multiparty Interactions with a Robot Using Large Language Models

Prasanth Murali
Northeastern University
Boston, MA, USA
murali.pr@northeastern.edu

Ian Steenstra
Northeastern University
Boston, MA, USA
steenstra.i@northeastern.edu

Hye Sun Yun
Northeastern University
Boston, MA, USA
yun.hy@northeastern.edu

Ameneh Shamekhi
Nuance Inc.
Burlington, MA, USA
Ameneh.shamekhi@nuance.com

Timothy Bickmore
Northeastern University
Boston, MA, USA
t.bickmore@northeastern.edu

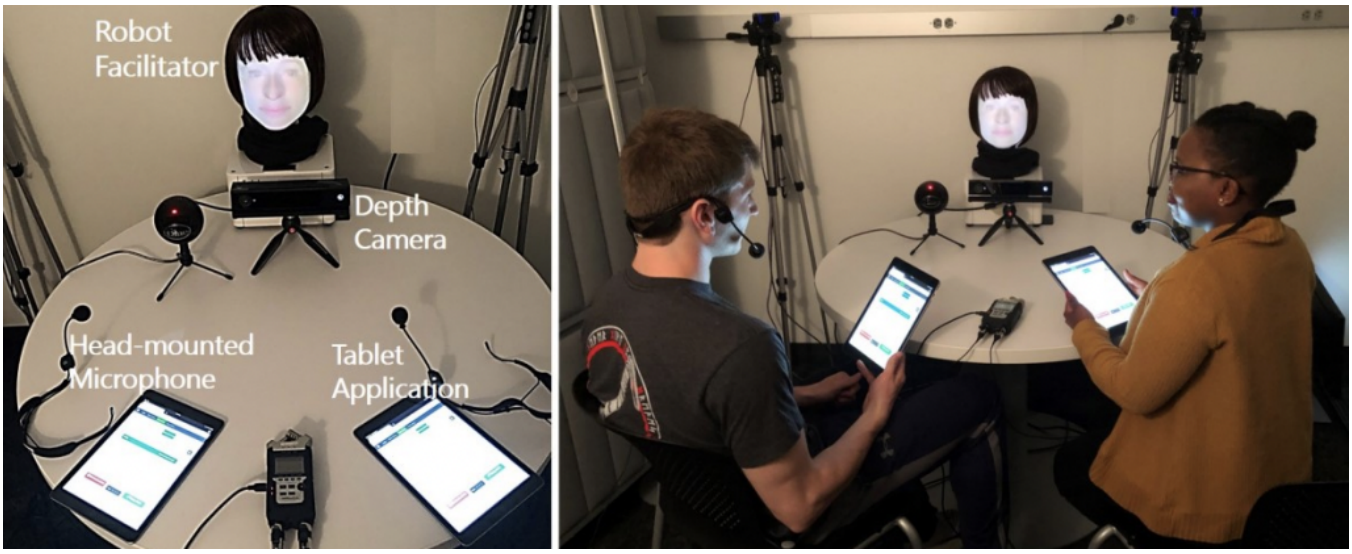

Figure 1: Meeting Facilitation Robot from Ameneh and Bickmore, 2019 [1]

ABSTRACT

Speaker diarization is a key component of systems that support multiparty interactions of co-located users, such as meeting facilitation robots. The goal is to identify who spoke what, often to provide feedback, moderate participation, and personalize responses by the robot. Current systems use a combination of acoustic (e.g. pitch differences) and visual features (e.g. gaze) to perform diarization, but involve the use of additional sensors or require overhead signal processing efforts. Alternatively, automatic speech recognition (ASR) is a necessary step in the diarization pipeline, and utilizing the transcribed text to directly identify speaker labels in the conversation can eliminate such challenges. With that motivation, we leverage large language models (LLMs) to identify speaker labels from transcribed text and observe an exact match of 77% and a word level accuracy of 90%. We discuss our findings and the potential use of LLMs as a diarization tool for future systems.

KEYWORDS

Large Language Models (LLMs), ChatGPT, Diarization, Social Robots, Meeting Facilitation

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585602

1 INTRODUCTION

Systems that can interact with multiple co-located users have many important applications, including classroom education [2, 3], in-person presentations [4–6], business meeting support [1, 7], and facilitating group counseling [8]. These systems track ongoing interactions among users and intervene to provide relevant information and guidance to the interactants to improve the interaction quality and task outcomes. For example, a business meeting facilitation system can enforce equity by ensuring that all users have a chance to be heard or increase meeting efficiency by providing the information needed by the group before it is requested.

Embodied conversational agents or humanoid robots may be particularly effective in the role of meeting organizers, by emulating the behavior of good human facilitators, educators, or counselors. These interfaces can use human verbal and nonverbal conversational behavior to intervene using speech, gesture, and prosody, and direct questions and statements to individual users or subsets of users through gaze and deictic gestures [9]. A key requirement for most multiparty support systems is the ability to identify which user is currently speaking. This is essential when responses to individual user queries or statements need to be provided, or if the system needs to know whether the prompted user is the one who responded. This is particularly challenging when responses are co-constructed by multiple users [8].

There are several approaches to speaker identification from speech, also referred to as speaker diarization. Several solutions require dedicated hardware, such as microphones for each user [1], or microphone arrays that can detect the direction of the audio source [10–12]. Most research in this area involves the identification of the unique audio qualities of each user's voice [11, 13]. However, accuracy with these approaches varies depending on the number of participants and the particular voice qualities of the individuals [14, 15].

Another approach involves first transforming user speech to text for all users, and then analyzing the stream of words to identify speaking turn boundaries [16–20]. Such an approach enables the use of a single microphone and is not dependent on the voice qualities of particular users. Although this approach does rely on accurate speech-to-text, automatic speech recognition (ASR) accuracy is now very high, comparable to human performance [21].

In this paper, we describe an approach to speaker diarization in which the stream of words for all speakers in a meeting is analyzed using a Large Language Model (LLM). This approach enables speaker identification in multiparty interaction from a single microphone source, obviating the need for special equipment. This also eliminates the need for speaker voice identification based on voice quality, making it more robust for large numbers of users or for users whose voices are very similar. Our application domain is multiparty business meetings, and this exploratory work uses a corpus previously collected for a study on robot-mediated group decision-making.

2 RELATED WORK

In this section, we review previous work through the lens of *design of social robots for multiparty conversations* to support our research framing, current *speaker diarization solutions* to position the novelty in our work, and *LLMs and their ability to perform tasks without finetuning* to lend claim to our proposed solution.

2.1 Design of Social Robots for Multiparty Conversations

Conversational agents and robots are effective candidates for mediating multiparty conversations. Their abilities to communicate in natural language and nonverbal cues, display non-judgemental and neutral behaviors towards all interlocutors, and access vast knowledge bases to guide conversations have led researchers to explore them in several multiparty settings. The growing body of work in this space has shown that social robots are effective in playing different roles in various multiparty domains such as promoting participant conversation in a group setting [22–24], moderating gameplay in group gaming settings [25], and improving a group's engagement and attentiveness [26] by understanding the multiparty dialog input.

Prior work has demonstrated that nonverbal cues can be used to mediate multiparty conversations. Parreira et al. showed that a social robot can utilize non-verbal behavior such as locking and averting gaze, to seamlessly mediate a multiparty dialog and ensure users participate equally in the conversation [27]. Furthermore, studies such as by Cassel and Thorisson [28], Chao and Thomas [29, 30], and Bohus and Horvitz [31, 32], have led to the development of computational models of turn-taking for human-agent interaction by leveraging gaze and other nonverbal cues. Most recently, Moujahid et al. [33] developed a robot receptionist that coordinates conversations between multiple users by engaging and disengaging with a particular user based on the number of people in the scene. This research identified that the tailored engagement behavior of the robot was perceived as intelligent, conscious, polite, and respectful of social norms. In the same vein, Skantze et al. explored how a robot can manage turn-taking through attention management and gaze when multiple players are working with a shared display [34].

In addition to nonverbal cues, researchers have developed approaches to understand dialog to mediate conversations in multiparty human-agent settings. In fact, early research on human-agent interaction positions dialog management as one of the core goals of the field [35]. While this may not be trivial, improved dialog management in this context can lead to increased user satisfaction, task efficiency, and ultimately task success as well. With that goal, Utami and Bickmore [8] developed a couples counselor robot to improve the quality of relationships between intimate partners. The robot seamlessly manages turn-taking in this multiparty setting, by automatically detecting when it can merge into the conversation between the two humans. Similarly, Shamekhi and Bickmore [1] developed an automated group facilitation system, in which a robot facilitator uses multimodal sensory inputs to moderate a small group decision-making session. The system uses two microphones to detect speech input from each user and manages the meeting to ensure equal participation from both users. This research platform also serves as our dataset for the speaker diarization task in our work, explained in greater detail in Section 3.1.

2.2 Speaker Diarization Solutions

Researchers have explored the use of deep learning approaches on audio or video data to identify the speech of each unique speaker. Recent research directions in this space include using speaker embeddings [36–38], voice activity [39], transcribed lexical data [16–20], and even end-to-end systems [40–42], in order to perform speaker diarization. Lexical data from ASR was primarily used for refining output from speaker activity detection models [39], segmenting transcripts into speech by different speakers - through identifying word boundaries [17] or inserting speaker tags by training attention-based encoder-decoder systems [42]. Perhaps the closest to our work is the research by Flemotomos et al. [19], where

the authors demonstrated that linguistic cues improve the performance of speaker diarization systems, in cases where the roles of the speakers are distinct and well-defined. The authors trained a model to identify individual speaker segments by detecting change in speaker role. In the same vein, our work considers the use of ASR output that captures linguistic patterns to segment and identify speaker boundaries to determine speaker labels. To the best of our knowledge, this is the first case study to explore the use of pre-trained LLMs to perform speaker segmentation and ultimately speaker diarization, based solely on the ASR output.

2.3 Large Language Models (LLMs)

In natural language processing (NLP), pre-trained large language models, such as BERT [43], GPT-2 [21], and GPT-3 [44], have revolutionized a wide range of natural language tasks from understanding to generation. LLMs are essentially deep neural networks [45] with millions of parameters pre-trained on large amounts of unlabeled text. In the past few years, using pre-trained LLMs and fine-tuning them for different tasks has become the dominant paradigm for NLP applications [43, 44, 46–50]. Furthermore, these models, trained on massive corpora have shown to perform well on a variety of target domains and tasks [43, 44, 47, 51].For instance, GPT-3 was able to learn a broad set of skills and pattern recognition abilities during training and then use those abilities during inference to rapidly adapt to or recognize the desired task [44]. Building on this, OpenAI introduced ChatGPT[1], which is an LLM trained using reinforcement learning from human feedback to follow a instruction in a prompt and provide a detailed response. ChatGPT was specifically trained from GPT-3.5 to interact conversationally with the user. The dialog format of the output makes it possible for the model to answer follow-up questions, admit its mistakes, and reject inappropriate requests. Due to ChatGPT's promising zero-shot performance (where no demonstrations or examples are used and only an instruction in natural language is given to the model [44]) and training on datasets in dialog format, we chose ChatGPT to be our choice of LLM for the speaker diarization task which to our knowledge has not been done yet using LLMs. For cases when there are very limited ground truth data, using zero-shot prompting with an LLM can be a promising approach to achieve more accurate speaker diarization results without any finetuning.

3 METHODS

3.1 Preliminary Data Collection Experiment using a Social Robot

To evaluate our concept, we used the experimental data collected by Shamekhi and Bickmore[1], as a part of their research on the development of a group facilitation social robot. In the following section, we briefly discuss the experimental design that informs the dataset for our experiment.

In this research, dyads of participants interacted with a meeting facilitation social robot (Furhat[2]), where the robot managed the multiparty conversation, enforced a meeting structure, ensured balanced participation on topics from both the participants, and facilitated a decision-making process where the dyads engaged in a simulated discussion for hiring different candidates for a particular role. In particular, the participants were asked to review and discuss a set of six fictional resumes for a sales manager position and select the best candidate for an interview. Resumes were designed so that the candidates had subjective strengths and weaknesses that promoted a discussion among the dyads.

The study used a structured hiring scenario and involved a greeting and introduction to the session by the robot. Following this, the participants were provided an opportunity to individually review the resumes of different candidates. Next, the two participants discussed every candidate with each other in front of the robot and eliminated unfavorable candidates. Finally, they selected and decided on the best candidate before ending the session. More details about the study can be found in their paper [1]. The average total length of each session was about 30 minutes, and the meetings were videotaped for future analysis.

3.2 Deriving the Ground Truth Speaker Labels

We carefully examined the video data from the study described in Section 3.1 and determined that the discussion between the dyads of participants on selecting the best candidate was an appropriate dataset for our exploratory diarization experiment. We transcribed all the video recordings (n = 20) using OpenAI's Whisper [52] and corrected any errors in speech recognition. From the meeting transcripts, we identified multiple multiparty dialog blocks between the two participants that were relevant to this study. These dialog blocks included at least one turn for each speaker, excluded any dialog from the social robot, and followed a single conversational flow without interruption.

For example:

> "I think he should have been more interactive with the statements, some more elaborated. Otherwise, his education is quite good. Even his experience is not that much, but his numbers are quite big. I think he should explain some of this. I think his overall structure for the resume is not that great since he started with skills and other abilities than his experience, even though he mentioned it."

The first and second authors of this work then manually identified and created speaker labels from the dialog blocks by watching the video recordings to establish our ground truth diarization data. The speaker labels for the above example are:

> "Manager 1: I think he should have been more interactive with the statements, some more elaborated. Otherwise, his education is quite good. Even his experience is not that much, but his numbers are quite big. I think he should explain some of this."

> "Manager 2: I think his overall structure for the resume is not that great since he started with skills and other abilities than his experience, even though he mentioned it."

In total, we had 74 dialog blocks across our dataset. The average number of speaker turns in each block was 6.27 (SD = 8.54). The

[1]https://openai.com/blog/chatgpt/
[2]https://furhatrobotics.com/

		Predicted Labels		
		Speaker 1	**Speaker 2**	**Total**
True Labels	**Speaker 1**	368	51	419
	Speaker 2	52	299	351
	Total	420	350	770

Table 1: Confusion matrix for sentence level annotations across the dataset

		Predicted Labels		
		Speaker 1	**Speaker 2**	**Total**
True Labels	**Speaker 1**	3661	353	4014
	Speaker 2	252	2944	3196
	Total	3913	3297	7210

Table 2: Confusion matrix for word level annotations across the dataset

average numbers of words and sentences in each block were 102.67 (SD = 74.64) and 9.55 (SD = 10.40), respectively.

3.3 Deriving Speaker Labels from ChatGPT

To derive the speaker labels using ChatGPT, we manually provided a prompt and the multiparty dialog block directly into the ChatGPT web interface [3]. The following is the prompt used to perform the diarization task:

> "Here is a conversation between two hiring managers about a candidate. The conversation potentially consists of multiple turns of dialog between the two managers and not just a simple back and forth. Without changing any of the words or sentence structure, split the prose as a dialog between the two managers by figuring out the end of turn based on context (and not just periods) and identifying which manager said what. Keep all the information the same. Do not paraphrase."

We then repeated this process for all the multiparty dialog blocks and derived speaker labels through ChatGPT for the entire dataset. The output from ChatGPT was in the same format as the ground truth with speaker labels (shown in 3.2) and the corresponding spoken text.

4 RESULTS

To evaluate the performance of the LLM-based approach in identifying speaker labels, we compared the output provided by ChatGPT against the ground truth data and computed four different measures of accuracy, as described in the following section. Since our speaker diarization task can be seen as text classification with text output, we looked to text similarity and text classification measures standardized in the NLP community [53, 54], to derive these metrics.

4.1 Exact Match

The Exact Match (EM) [55] is a metric often used for Question Answering (QA) tasks in NLP, measuring the precision that matches all the ground-truth answers precisely. In our case, we look at whether or not the entire input dialog block is precisely annotated with the correct speaker labels. The average EM across the dataset was 77%.

4.2 Sentence Level Annotation Accuracy

The sentence level annotation accuracy (SLAA) was computed as the ratio of the number of correct sentences attributed to a speaker to the total number of all sentences said by a speaker in a dialog exchange. The average SLAA across the dataset was 77%. By looking at sentence-level annotations through the lens of a binary classification problem (as a sentence having the label Speaker - 1 or Speaker - 2), we also computed the confusion matrix based on the annotations derived from ChatGPT as reported in Table 1. The f1 score derived using this approach was 0.88.

4.3 Word Level Annotation Accuracy

The word level annotation accuracy (WLAA) was computed as the ratio of the number of correct words attributed to a speaker to the total number of all words said by a speaker in a dialog exchange. The average WLAA across the dataset was 90%. By looking at the word level annotations through the lens of a binary classification problem (as a word having the label Speaker - 1 or Speaker - 2), we also computed the confusion matrix based on the annotations derived from ChatGPT as reported in Table 2. The f1 score derived using this approach was 0.92.

4.4 Jaccard Similarity

The Jaccard Similarity [53, 56] between the two groups of text was calculated as the ratio of the number of common words to the number of total unique words in both groups. Mathematically, it can be defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ where A is the ChatGPT annotated text and B is the ground truth annotated text. Our analysis shows that the average Jaccard Similarity for the two-speaker setting across our corpus is 0.82. The Jaccard Distance (computed as $1 - J(A, B)$) measures the dissimilarity between the two groups and is thus 0.18 for our case.

4.5 Prompt Engineering as a Guideline for using LLM for Speaker Diarization

Similar to other research on using LLMs for a variety of domains and tasks [43, 44, 47, 51], a core component of our experiment involved iterating over several prompt variants ("prompt engineering") to get the desired results. Our initial results indicated that ChatGPT not only performed diarization and identified speaker labels, but also summarized the speaker text when providing the output (around 60% of the time). After several iterations, we explicitly asked the tool to *"Keep all the information the same. Do not paraphrase."*, to avoid any summarization of text from its end. This kind of "prompting" to get the desired behavior out of an LLM aligns with the dimension of *Prompt programming as constraining behavior*, based on research by Reynolds and McDonnel [57]. This informs our primary guideline for using LLMs in diarization settings - an extensive study of prompt engineering specific to the task on hand is required, and we theorize that the performance of the model would vary widely depending on the prompt that is being used [58].

[3] https://chat.openai.com/chat

5 DISCUSSION

Our findings highlight a potential approach to using lexical cues to improve speaker diarization performance in multiparty conversations with a social robot. Given their extensive training on conversational dialog [59, 60], LLMs such as ChatGPT perform well on conversational tasks, possibly explaining the generally positive direction of our diarization findings. Our results are also highly dependent on the specific prompt we used, and prompt engineering, as described in Section 4.5 is required to extend the findings to other domains and settings. The Word Level Annotation Accuracy for the diarization task had the highest score, followed by the Jaccard Similarity, Sentence Level Annotation, and finally Exact Matching. Upon further analysis, we noticed that the majority of word-level annotation errors occurred at turn-taking boundaries. This aligns with the idea that language models like ChatGPT, while good at understanding general language patterns, may not have enough domain-specific experience to fully understand real-world interactions. This has also been noted in recent studies, such as those by Brown et al. [44], and Bisk et al.[61], which have highlighted the importance of grounding language models in real-world experiences. This can lead to LLMs being able to use learned patterns (i.e., after a certain number of sentences, after specific keywords in conversations) to determine when turn-taking normally should happen but lacks making predictions based on the complete context of real-world conversations.

Furthermore, some of the incorrect sentence-level speaker annotations included cases where conjunctions (such as 'and' and 'but') were present in the dialog block. In the case of *'and'*, the model incorrectly attaches the label of the preceding speaker to the text following *'and'*, whereas ground truth annotations show that a new speaker had started their turn with *'and'* by adding to the previous speaker's chain of thought. While grammatically 'and' represents a continuous flow of thought, conversationally humans use conjunctions to both receive and give turns when speaking [62]. We theorize that the incorrect results from ChatGPT might be due to the model relying on the rules of English grammar together with transitions in everyday conversations to identify the speaker labels. Similarly, in some cases that involved the use of the word *'but'*, the model associated the text following the word to another speaker, but in fact, the original speaker was just challenging their own line of thought. While we do not know how the LLM understands dialog, these instances shed some light on its operation, particularly around its emphasis on the rules of grammar and less on conversational norms. Finetuning pre-trained LLMs on diarization tasks can mitigate some of these known issues.

6 LIMITATIONS

Despite the positive results, it is important to note that there are several limitations in the current experiment. First, we did not evaluate our approach for diarization against other standardized datasets [63–67]. Furthermore, traditional diarization algorithms provide timestamps for the speaker labels, which were not performed in our experiment. We also did not perform one-shot or few-shot learning which could improve diarization performance [44]. An extensive study of prompt engineering could also lead to better results [57, 58, 68]. In addition, our diarization results in this study are limited to two speakers in a specific scenario of discussing resumes - further evaluation is required to understand how well LLMs can perform diarization in contexts involving more than two speakers and other conversational scenarios. Finally, we did not use the results from this study to actually drive the robot and evaluate it in a controlled laboratory study with users. However, our experiment was an initial step towards developing a fully-automated social robot system that can perform group facilitation, and the current work identifies and evaluates an approach to only incrementally automate parts of that system.

7 FUTURE WORK

There are several research paths that our exploratory work points to.

7.1 Building a Fully Automated Group Facilitation Social Robot:

Our immediate goal is to integrate diarization findings from this research in a fully functional multiparty social robot and evaluate the outcomes in a controlled laboratory study. The overarching theme behind this research direction is to build a system that can effectively facilitate and communicate within multiparty interactions through the use of automatic speaker diarization. The promise of LLMs on the conversational diarization task also opens doors for exploring their use for dialog management, turn-taking, and natural language understanding in human-agent/robot scenarios. Given the performance of LLMs on conversational tasks, LLMs can be used to extend and support dialog management strategies to manage conversation between humans and an agent/robot. For instance, current dialog management systems in human-agent interaction are still skewed in favor of supporting 1:1 conversations as opposed to multiparty conversations [69], and scaling dialog management in human-agent interactions to a multiparty setting is still an open problem. While interesting solutions such as treating each set of pairs as a separate 1:1 conversation have been proposed [35], an LLM finds several uses in this setting such as turn-taking management, identifying when the agent can merge into the conversation, and supporting Natural Language Understanding ("NLU") in appropriate settings. That said, the behavior of LLMs has yet to be completely understood, and there is ambiguity around their use in conversational systems. Further, more research is also necessary to investigate the potential consequences of inaccuracies produced by the diarization model during group facilitation sessions with social robots. Previous studies have explored "breakdown strategies" in the event of chatbot failures [70], and extending these findings to social robots presents an exciting avenue for further research.

7.2 Leveraging Large Language Models and Lexical Cues for Performing Speaker Diarization:

Previous research on using lexical cues from ASR for diarization has almost exclusively focused on identifying the speaker role or speaker identity [19, 42] as steps toward the larger diarization task. Furthermore, systems that directly infer speaker labels have

required training end-to-end neural network models [40, 41] or be used in conjunction with other signals (acoustic/video) with deep learning approaches [71–74]. Thus, the potential of using an LLM to perform diarization, as shown in this work, can help reduce the time and resources needed for training new models and/or provide better initial embeddings to optimize the training process.

LLMs for diarization remain a relatively unexplored paradigm and given their ability to perform this task without any gradient updates or additional training, can be used to aid in the diarization process. We hypothesize that lexical cues picked up by LLMs, in conjunction with other multimodal cues such as acoustic and video signals, can greatly improve the performance of current diarization algorithms. Furthermore, by developing language models trained specifically on the diarization task, we can further improve the performance of these algorithms in identifying speaker labels from lexical input. By way of studying prompt engineering, we can leverage current state-of-the-art models such as GPT-3 [44], FLAN-T5 [75], and OPT [76], in addition to ChatGPT, to directly use them for diarization tasks. In addition, we would like to test different models based on few-shot learning and one-shot learning to see if there are substantial improvements in accuracy based on the learning approach. Consequently, evaluating the models on data with more than 2 speakers and different conversational scenarios can show whether or not LLMs are truly robust for the speaker diarization task. Finally, to establish LLMs as a reliable method for performing speaker diarization, further research is necessary to compare their performance against standard diarization datasets, such as presented in [63–67].

8 CONCLUSION

In this work, we have shown promise in the use of LLMs, such as ChatGPT, to perform speaker diarization in a multiparty interaction with a social robot. Our results provide a case for using lexical input from ASR transcripts for performing speaker diarization. This interdisciplinary work contributes to the fields of HCI and NLP, by presenting a novel way to perform speaker diarization for implementation in social robots & embodied conversational agents to better facilitate multiparty interactions. We look forward to a future where language models are used more extensively in human-agent/robot interactions.

ACKNOWLEDGMENTS

This work was supported, in part, by the National Cancer Institute (NIH grant number R01CA273208).

REFERENCES

[1] Ameneh Shamekhi and Timothy Bickmore. A multimodal robot-driven meeting facilitation system for group decision-making sessions. In *2019 International Conference on Multimodal Interaction*, pages 279–290, 2019.

[2] Takayuki Kanda, Michihiro Shimada, and Satoshi Koizumi. Children learning with a social robot. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 351–358. IEEE, 2012.

[3] Violeta Rosanda and Andreja Istenic Starcic. The robot in the classroom: a review of a robot role. In *International Symposium on Emerging Technologies for Education*, pages 347–357. Springer, 2020.

[4] Qaysar Salih Mahdi, Idris Hadi Saleh, Ghani Hashim, and Ganesh Babu Loganathan. Evaluation of robot professor technology in teaching and business. *Information Technology in Industry*, 9(1):1182–1194, 2021.

[5] Yunus Terzioğlu, Prasanth Murali, Everlyne Kimani, and Timothy Bickmore. Sharing the spotlight: Co-presenting with a humanoid robot. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 551–560. IEEE, 2022.

[6] Ha Trinh, Reza Asadi, Darren Edge, and T Bickmore. Robocop: A robotic coach for oral presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):1–24, 2017.

[7] Sigurdur Orn Adalgeirsson and Cynthia Breazeal. Mebot: A robotic platform for socially embodied telepresence. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 15–22. IEEE, 2010.

[8] Dina Utami and Timothy Bickmore. Collaborative user responses in multiparty interaction with a couples counselor robot. In *Proceedings of the Conference on Human Robot Interaction (HRI)*, 2019.

[9] Timothy Bickmore and Justine Cassell. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 396–403, 2001.

[10] Aasish Pappu, Ming Sun, Seshadri Sridharan, and Alex Rudnicky. Situated multiparty interaction between humans and agents. In *International Conference on Human-Computer Interaction*, pages 107–116. Springer, 2013.

[11] Zerrin Yumak, Jianfeng Ren, Nadia Magnenat Thalmann, and Junsong Yuan. Modelling multi-party interactions among virtual characters, robots, and humans. *Presence: Teleoperators and Virtual Environments*, 23(2):172–190, 2014.

[12] Chinmaya Mishra and Gabriel Skantze. Knowing where to look: A planning-based architecture to automate the gaze behavior of social robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1201–1208. IEEE, 2022.

[13] Zerrin Yumak and Nadia Magnenat-Thalmann. Multimodal and multi-party social interactions. In *Context aware human-robot and human-agent interaction*, pages 275–298. Springer, 2016.

[14] Hannah Louise Bradwell, Gabriel E Aguiar Noury, Katie Jane Edwards, Rhona Winnington, Serge Thill, and Ray B Jones. Design recommendations for socially assistive robots for health and social care based on a large scale analysis of stakeholder positions: Social robot design recommendations. *Health Policy and Technology*, 10(3):100544, 2021.

[15] Zheng-Hua Tan, Nicolai Bæk Thomsen, Xiaodong Duan, Evgenios Vlachos, Sven Ewan Shepstone, Morten Højfeldt Rasmussen, and Jesper Lisby Højvang. isociobot: a multimodal interactive social robot. *International Journal of Social Robotics*, 10(1):5–19, 2018.

[16] SE Tranter, K Yu, DA Reynolds, G Evermann, DY Kim, and PC Woodland. An investigation into the interactions between speaker diarisation systems and automatic speech transcription. Technical report, Tech. Rep. CUED/F-INFENG/TR-464, Cambridge University Engineering Department, 2003.

[17] Jan Silovsky, Jindrich Zdansky, Jan Nouza, Petr Cerva, and Jan Prazak. Incorporation of the asr output in speaker segmentation and clustering within the task of speaker diarization of broadcast streams. In *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, pages 118–123. IEEE, 2012.

[18] Leonardo Canseco-Rodriguez, Lori Lamel, and Jean-Luc Gauvain. Speaker diarization from speech transcripts. In *Proc. ICSLP*, volume 4, pages 3–7, 2004.

[19] Nikolaos Flemotomos, Panayiotis Georgiou, and Shrikanth Narayanan. Linguistically aided speaker diarization using speaker role information. *arXiv preprint arXiv:1911.07994*, 2019.

[20] Tae Jin Park, Kyu J Han, Jing Huang, Xiaodong He, Bowen Zhou, Panayiotis Georgiou, and Shrikanth Narayanan. Speaker diarization with lexical information. *arXiv preprint arXiv:2004.06756*, 2020.

[21] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[22] Yoichi Matsuyama, Iwao Akiba, Akihiro Saito, and Tetsunori Kobayashi. A four-participant group facilitation framework for conversational robots. In *Proceedings of the SIGDIAL 2013 Conference*, pages 284–293, 2013.

[23] Yoichi Matsuyama, Iwao Akiba, Shinya Fujie, and Tetsunori Kobayashi. Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech & Language*, 33(1):1–24, 2015.

[24] Yoichi Matsuyama and Tetsunori Kobayashi. Towards a computational model of small group facilitation. In *2015 AAAI spring symposium series*, 2015.

[25] Hamish Tennent, Solace Shen, and Malte Jung. Micbot: a peripheral robotic object to shape conversational dynamics and team performance. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 133–142. IEEE, 2019.

[26] Ameneh Shamekhi, Q Vera Liao, Dakuo Wang, Rachel KE Bellamy, and Thomas Erickson. Face value? exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

[27] Maria Teresa Parreira, Sarah Gillet, Marynel Vázquez, and Iolanda Leite. Design implications for effective robot gaze behaviors in multiparty interactions. In *HRI*, pages 976–980, 2022.

[28] Justine Cassell and Kristinn R Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538, 1999.

[29] Crystal Chao and Andrea Lockerd Thomaz. Turn taking for human-robot interaction. In *2010 AAAI Fall Symposium Series*, 2010.

[30] Andrea L Thomaz and Crystal Chao. Turn-taking based on information flow for fluent human-robot interaction. *AI Magazine*, 32(4):53–63, 2011.

[31] Dan Bohus and Eric Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–8, 2010.

[32] Dan Bohus and Eric Horvitz. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*, pages 98–109, 2011.

[33] Meriam Moujahid, Helen Hastie, and Oliver Lemon. Multi-party interaction with a robot receptionist. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 927–931. IEEE, 2022.

[34] Gabriel Skantze, Martin Johansson, and Jonas Beskow. Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 67–74, 2015.

[35] David Traum. Issues in multiparty dialogues. In *Workshop on Agent Communication Languages*, pages 201–211. Springer, 2004.

[36] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4052–4056. IEEE, 2014.

[37] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. End-to-end text-dependent speaker verification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119. IEEE, 2016.

[38] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. Speaker diarization with lstm. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 5239–5243. IEEE, 2018.

[39] Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, et al. The stc system for the chime-6 challenge. In *CHiME 2020 Workshop on Speech Processing in Everyday Environments*, 2020.

[40] Yusuke Fujita, Shinji Watanabe, Shota Horiguchi, Yawen Xue, Jing Shi, and Kenji Nagamatsu. Neural speaker diarization with speaker-wise chain rule. *arXiv preprint arXiv:2006.01796*, 2020.

[41] Naoyuki Kanda, Xuankai Chang, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. Investigation of end-to-end speaker-attributed asr for continuous multi-talker recordings. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 809–816. IEEE, 2021.

[42] Huanru Henry Mao, Shuyang Li, Julian McAuley, and Garrison Cottrell. Speech recognition and multi-speaker diarization of long conversations. *arXiv preprint arXiv:2005.08072*, 2020.

[43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[44] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[46] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021.

[47] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[48] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[49] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[50] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[51] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[52] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[53] Wael H Gomaa, Aly A Fahmy, et al. A survey of text similarity approaches. *international journal of Computer Applications*, 68(13):13–18, 2013.

[54] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*, 2020.

[55] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[56] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

[57] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[58] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

[59] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022.

[60] Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. Godel: Large-scale pre-training for goal-directed dialog. arXiv, June 2022.

[61] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

[62] Gareth Walker. Coordination and interpretation of vocal and visible resources:'trail-off'conjunctions. *Language and speech*, 55(1):141–163, 2012.

[63] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer, 2006.

[64] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE, 2003.

[65] Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249*, 2020.

[66] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. The fifth'chime'speech separation and recognition challenge: dataset, task and baselines. *arXiv preprint arXiv:1803.10609*, 2018.

[67] Lei Sun, Jun Du, Chao Jiang, Xueyang Zhang, Shan He, Bing Yin, and Chin-Hui Lee. Speaker diarization with enhancing speech for the first dihard challenge. In *Interspeech*, pages 2793–2797, 2018.

[68] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[69] Jelte van Waterschoot, Merijn Bruijnes, Jan Flokstra, Dennis Reidsma, Daniel Davison, Mariët Theune, and Dirk Heylen. Flipper 2.0: A pragmatic dialogue engine for embodied conversational agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 43–50, 2018.

[70] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[71] Yifan Ding, Yong Xu, Shi-Xiong Zhang, Yahuan Cong, and Liqiang Wang. Self-supervised learning for audio-visual speaker diarization, 2020.

[72] Pavel Campr, Marie Kunešová, Jan Vaněk, Jan Čech, and Josef Psutka. Audio-video speaker diarization for unsupervised speaker and face model creation. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, pages 465–472, Cham, 2014. Springer International Publishing.

[73] Athanasios Noulas, Gwenn Englebienne, and Ben J.A. Krose. Multimodal speaker diarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):79–93, 2012.

[74] Nikolaos Tsipas, Lazaros Vrysis, Konstantinos Konstantoudakis, and Charalampos Dimoulas. Semi-supervised audio-driven tv-news speaker diarization using deep

neural embeddings. *The Journal of the Acoustical Society of America*, 148(6):3751–3761, 2020.

[75] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[76] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.