

Real-time Multimodal Addressee Detection in Multi-party Dialogue

Divesh Lala
Kyoto University
Kyoto, Japan
lala@sap.ist.i.kyoto-u.ac.jp

Taiga Mori
Graduate School of Informatics
Kyoto
Kyoto, Japan
mori.taiga.5h@sap.ist.i.kyoto-u.ac.jp

Koji Inoue
Graduate School of Informatics
Kyoto University
Kyoto, Japan
inoue@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
School of Informatics
Kyoto University
Kyoto, Japan
kawahara@i.kyoto-u.ac.jp

Abstract

For multi-party conversation, a critical task is addressee detection, the identification of the intended recipient of a conversational turn. The task of addressee detection is complicated by the fact that there may not be an explicit addressee. In human-robot and human-agent interaction, the agent should have the ability to reliably detect the addressee to ensure that it does not interrupt the flow of conversation. In this work we focus on the tasks related to multimodal addressee detection, by combining modalities from vision, audio and text using attention-based models. We compare the performance of various combinations of modalities for these tasks. We find that our models can significantly outperform the baseline and an LLM-based model. In the main task of addressee detection, the combination of multimodal features is effective. We also find that the text modality is better for predicting whether a turn has an explicit addressee, while vision modalities are better suited for identification of the explicit addressee. Our models are able to process information with low latency, making them suitable for real-time interaction scenarios.

CCS Concepts

• **Human-centered computing** → **Interaction design process and methods.**

Keywords

Addressee Detection; Conversational Turn-taking; Multi-party Dialogue

ACM Reference Format:

Divesh Lala, Koji Inoue, Taiga Mori, and Tatsuya Kawahara. 2026. Real-time Multimodal Addressee Detection in Multi-party Dialogue. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26)*, October 05–09, 2026, Napoli, Italy. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3776574.3831238>



This work is licensed under a Creative Commons Attribution 4.0 International License. *ICMI '26, Napoli, Italy*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2318-6/26/10

<https://doi.org/10.1145/3776574.3831238>

1 Introduction

For real-time situated conversation, autonomous agents and robots should be able to manage turn-taking tasks. These tasks involve knowing when the user has finished their turn so they can generate and respond in a timely manner. Recently there has been an increased focus on modeling this turn-taking behavior, resulting in models that can take the turn at near human-like speed, and in some cases even overlap human speech.

Conventionally, this research has been conducted in the context of dyadic conversation, where there are only two participants. However, everyday conversation also consists of multiple participants and designing agents which can manage these type of conversations should also be addressed. Turn-taking in multi-party conversation is arguably more complex because not only should participants be aware of when a turn has ended, but should also be able to recognize who should speak next. This phenomenon does not occur in dyadic conversation because there is only one possible candidate for the next speaker.

Furthermore, the turn itself may not explicitly indicate who can speak next. For example, if a participant in a multi-party conversation says “What do you all think?”, this indicates that any other participant may take the turn. Therefore, the addressee of the turn is not a specific person but the group as a whole. We term the prediction of whom the current speaker is directing their speech towards (including all the participants) as *addressee detection*.

A conversation agent engaging in multi-party dialogue requires an addressee detection model to maintain the rhythm of the conversation. It should be able to understand when a turn is directed at itself, so that it may respond quickly, or when it is directed at another participant, so that it does not interrupt the conversation. In order to do this, the agent requires an understanding of the multimodal behavior of the participants.

One approach to this could be to use multimodal large language models (MLLMs) which are able to analyze and predict the next speaker given audio, video and text information. MLLMs also possess the ability to do chain-of-thought reasoning to arrive at their prediction [35]. However, a drawback of this approach is the increase in latency. Decisions on whether or not to take the turn should be made very quickly, but using an MLLM can be costly in

terms of time compared to more traditional machine learning approaches. For human-agent interaction, a lightweight model should be considered as more beneficial even at a lower accuracy.

It is also essential to understand the extent to which modalities contribute towards the task of addressee detection. Conversation is of course a multimodal phenomenon which requires visual, audio and linguistic cues, and addressee detection is no exception to this. In this work we investigate the influence of particular modalities by comparing a range of unimodal and multimodal models. Furthermore, we look at several sub-tasks of addressee detection and the modalities that are necessary for an agent to achieve these tasks.

To address the above issues, this work undertakes a thorough analysis of addressee detection in a corpus of three-party conversations between humans in Japanese. We extract several modalities which can assist in determining who the speaker is addressing (if any) and then create both unimodal and multimodal models to compare them for various addressee detection tasks. We compare our models against an LLM to confirm the necessity of multimodal information. We also find that there are differences in the importance of modalities depending on the task.

The contributions of this work should be of interest to conversation researchers and practitioners. In this paper, we present the following outcomes:

- A description and annotation of the addressee detection task and several related sub-tasks
- A comparative evaluation of unimodal and multimodal models related to these sub-tasks
- A lightweight model that can be used for the purposes of multi-party human-agent interaction

2 Related Work

There has been much research conducted on turn-taking in conversation agents. This has often been from the perspective of dyadic interactions with the task being endpoint detection, or predicting when the user has finished their speaking turn [10, 13, 20, 30]. This includes recent full duplex models which allow the agent to take the turn before the user has finished speaking [7, 9, 19, 28]. Endpoint detection in multi-party dialogues has also been addressed in a similar manner using multimodal approaches [22, 23].

Many works deal with addressee detection in multi-party conversation by focusing on the related problem of next speaker prediction. In this scenario, the model tries to determine the actual next speaker. Several approaches have been implemented using gaze, head and mouth information, as well as multimodal modeling [15–17, 24]. Recent works have used LLMs and multimodal LLMs to predict the next speaker [7, 12, 26], however the inherent latency of these models means they may have limited use in an online context.

Rather than predicting who will speak, addressee detection predicts who is being spoken to. This task has been the subject of prior multimodal research before widespread use of deep learning models [5, 21, 32, 33]. Current approaches have used more sophisticated models [25, 31], however these assume the addressee is an individual participant. In this work we will show that in multi-party conversation it is common for the addressee to be the group itself.

Recent works have implemented addressee detection in multi-party conversation robots through the use of LLM prompting [1,

3, 4]. In these cases, text and gaze information is sent to the LLM which decides who is being addressed. However, the reliance on LLMs could cause issues, in particular because the gaze behavior sent to the prompt cannot be fully captured and may ignore shifts during the turn.

This work differentiates itself from previous research by providing a thorough analysis of not only addressee detection but also related tasks that are of interest to conversation agents in multi-party scenarios. Unlike previous work we also try to detect situations where the addressee is all participants in the conversation, as this is a regular phenomenon in the real world. We also emphasize the need for models to have low latency, so we ensure that our models are able to function in real-time to allow for fast turn-taking. To our knowledge there is no existing work which approaches this task in a real-time, multimodal manner.

3 Data Collection

In this work we extract samples of conversation from TEIDAN, a situated multi-party dialogue corpus [14]. The corpus consists of three participants engaging in free conversation on a particular topic. The topics include subjects such as “What city should be the capital of Japan?” and “What would be the most important thing to take to a desert island?”. The participants converse around a circular table, forming an approximate equilateral triangle shape. Each trio of participants engages in three separate sessions of dialogue with a different topic, each with a duration of 5-10 minutes long. The data of 12 trios was collected, resulting in 36 sessions of data.

The corpus contains several streams of data. Video data was recorded of each participant by a web camera placed on the table in front of their seated position. All video recordings were synchronized. Audio data was collected from separate pin microphones for each participant and also from a circular microphone array which was placed in the center of the table. All conversations were transcribed manually, including non-lexical utterances such as backchannels and fillers. Figure 1 shows the layout of the dialogue environment.

3.1 Turn and addressee annotation

Conversation turns were annotated manually by one expert annotator and labeled as either A, B or C corresponding to the participant holding the speaking floor. While in general only one person at a time can be holding the floor, there were some cases of overlap when multiple participants talked over each other. A total of 3121 turns were annotated.

For every turn the addressee was also annotated. Possible labels were

- A, B or C, the addressee is the corresponding participant
- E (everyone), the turn is addressed to all participants
- W (whoever), the addressee is an unspecified participant
- N (none), no-one is addressed, such as self-talk
- U (unknown), the addressee cannot be determined
- O (other), the addressee is none of the above categories

Annotators watched the video of the conversation up until the turn had ended and decided the addressee label only on the previous context and not the actual next speaker. In other words, the addressee and next speaker are not necessarily the same.

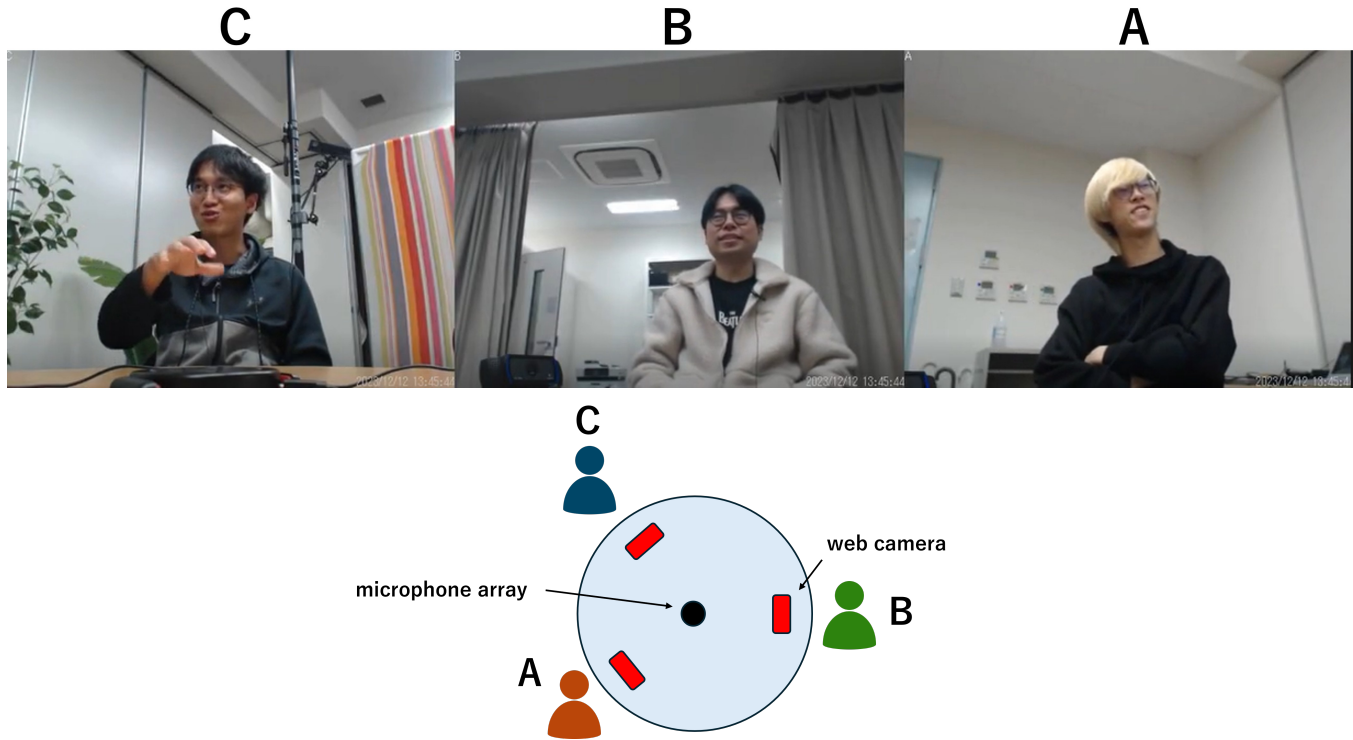


Figure 1: Screen capture of a multiparty conversation in the TEIDAN corpus with participants labeled.

General guidelines were provided to annotators to assist them in determining the addressee. For example:

- Does the speaker use the name of a person?
- Does the speaker use “we” or “us” in their utterance?
- Is the speaker responding to a question?
- Where is the speaker looking when they speak?

Speakers providing their own opinions were annotated as “everyone”, since they are directed towards the group as a whole. However, if this is a counter-argument to an opinion which has already been expressed (by uttering “but...”), the addressee is considered to be the person whose opinion is being challenged.

An addressee can also change within a turn. In this case, the annotator was asked to consider only the addressee at the end of the turn. For example in the following dialogue:

- A: “I think Osaka is good because the roads and railway are easy to understand.”
 B: “Like the Midosuji line?”
 A: “Exactly. Tokyo is really complicated, but Osaka is well laid out.”

When speaker A responds to B’s statement by saying “Exactly”, then the addressee is B, which is a direct answer to B’s original utterance. However the rest of the turn is expressing their opinion while looking at both B and C, therefore the addressee label is everyone.

For this addressee annotation, three annotators independently labeled each turn. We use the majority label as the final annotation.

In cases where all three annotators labeled the addressee differently, the final label was annotated as unknown. Inter-annotator agreement for this task using Fleiss’ kappa κ was 0.52, indicating moderate agreement. We note that labeling for this task is still somewhat subjective, particularly compared to identifying the next speaker.

3.2 Analysis of addressee labels

Figure 2 shows the distribution of addressee labels. The most frequent label is not a specific participant but the label for everyone (E). This indicates that addressee detection should not just predict a specific participant in the conversation, but also be able to predict whether a turn is directed towards the group as a whole.

We also find a portion of annotations labeled as unknown. This shows the difficulty of the task, where three annotators could not come to an agreement on the final label. For the purposes of this work, we mainly focus on the participant labels (A, B and C) and the everyone label (E) which account for over 90% of the labels.

We also compared the addressee labels to that of the actual next speaker. We only considered cases where the addressee label was an explicit participant (A, B or C) and the next speaker was not the same as the current turn’s speaker. It was found that out of 1,497 turns, the addressee label was different than the person who actually took the turn in 515 (34.4%) of them. This clearly shows that addressee detection and next speaker prediction are not equivalent phenomena. An example of a non-addressee taking the turn is as follows:

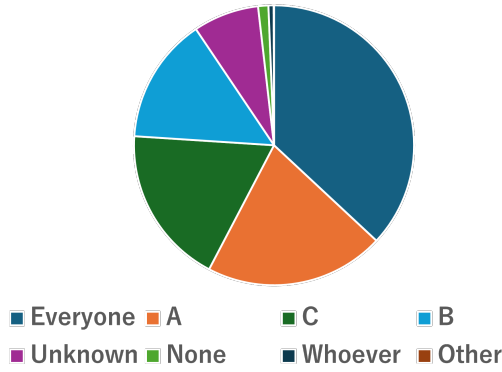


Figure 2: Distribution of addressee labels annotated in the TEIDAN corpus.

B speaking to C: “Ah, you can’t compete against Toyota.”

A: “Yeah. Toyota is really strong”

In multi-party conversation we argue that this phenomena is not uncommon, due to the underlying complexity of the involvement of several people. Dialogue agents do not have to accurately predict the next speaker, but should determine who *can* be the next speaker, which is closely related to the addressee.

4 Dialogue Tasks

The overall goal of this work is to predict the addressee of the turn which has just concluded. We make two assumptions. The first assumption is that the current turn speaker is known. The second assumption is that the end of the turn has been detected. This detection may be achieved by a separate turn-taking model, which should be investigated in future work.

In the corpus, the participants A, B and C are fixed and do not change. However, for training models which predict a particular participant, we need to be able to choose from labels which are not dependent on the explicit identity of the current speaker but also allow us to specify the non-speakers. To address this issue, if the current speaker is A then we do not predict B or C, but rather “left” and “right”, representing the relative position of the addressee candidates. We can then use this method for any speaker.

The primary task is a 3-class problem which tries to predict either one of the non-speakers (left or right) or everyone (E) as the addressee. We ignore unknown and other annotations and set this as the main addressee detection task (**MAIN**). However, related to this main task are other sub-tasks which also demand investigation. These sub-tasks are slightly different, but may also be important in other contexts related to dialogue and turn-taking.

One sub-task is to construct a binary classifier which predicts whether the addressee is an explicit participant or not. In other words, the positive class is when the addressee label is either A, B or C and the negative class is any other label. We term this as an explicit addressee detection task (**EXP**).

Another sub-task is to predict which of the participants is the addressee, given that the addressee is a participant (A, B, or C).

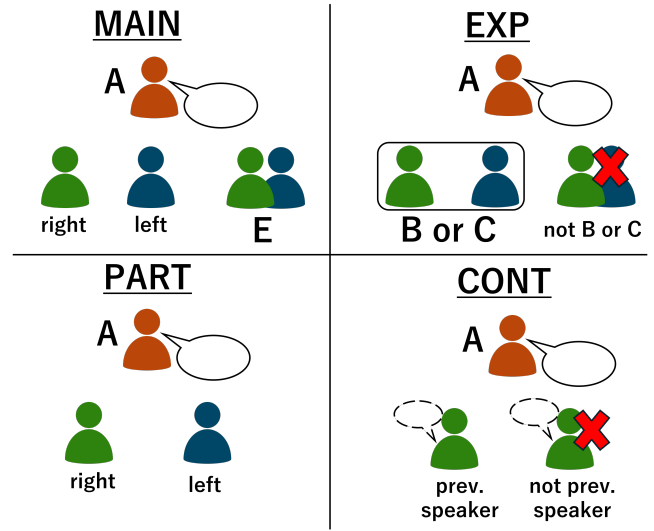


Figure 3: Overview of the tasks related to addressee detection in this paper. Assume that A is the turn speaker and has just completed their turn.

We term this as a participant detection task (**PART**), and in this corpus the model is also a binary classifier since there are three participants. We note that this task can be strongly influenced by the physical positions of the dialogue participants, which are fixed in the TEIDAN corpus.

Finally, we define a sub-task of continuing the interaction between a pair of participants. Assume that A was the speaker in the previous turn and B is the speaker in the current turn. We consider that interaction between the two has continued if the addressee of B’s turn is A. If the addressee is anything other than A, then their dyadic interaction has stopped. We frame this task as a continuation task (**CONT**) which predicts a continuing interaction.

Figure 3 summarizes the main task and three sub-tasks. A goal of this work is to understand how different modalities contribute to predicting each of the four tasks. For example, we hypothesize that the **PART** sub-task is best handled by vision modalities, since in this scenario the gaze and head orientation of the participants provides information that discriminates between explicit addressees.

5 Model architecture

We will now create several models which will focus on each of the tasks described in the previous section. We describe the process of feature extraction and the architecture of each of these models.

5.1 Feature extraction

First we extract all the turns in the corpus and use the addressee information as labels for all of the tasks. Some videos contained corrupted frames so these were removed for training. Note that the labels themselves will be slightly different depending on the task, as shown in Table 1.

For each turn we create the following feature sets:

- Eye gaze angles (yaw and pitch) for each participant

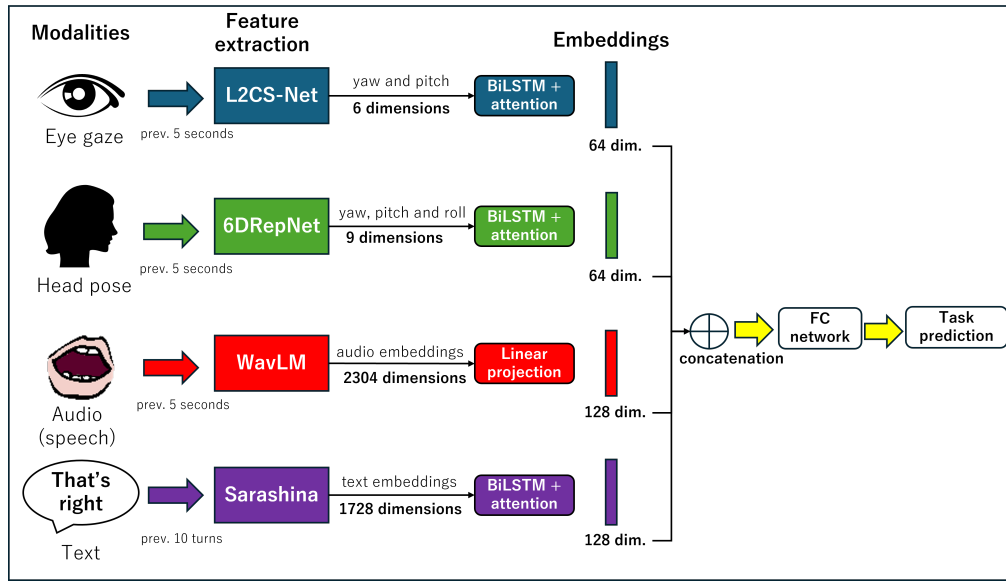


Figure 4: Multimodal architecture used in this work. For unimodal models the embeddings are input directly into a fully connected network for classification.

Table 1: Label candidates for each task

Task	Label candidates (addressee is...)
EXP	in $\{A, B, C\}$, not in $\{A, B, C\}$
PART	left of speaker, right of speaker
CONT	previous speaker, not previous speaker
MAIN	left of speaker, right of speaker, everyone

- Head pose angles (yaw, pitch and roll) for each participant
- Audio embeddings for each participant
- Text embeddings for the utterances of each participant

Eye gaze angles are calculated from the video of each participant using a pre-trained L2CS-Net model [2]. We sample the output angles at a rate of 5 Hz to ensure that it can function in real-time. The window for prediction is the five seconds preceding the end of the turn, creating a six-feature sequence with a length of 25 per sample. We make the assumption that the end of the turn is where most of the information regarding the addressee can be found. Preliminary experiments with longer context windows or using only the duration of the current turn showed no additional benefit.

Head pose angles are similarly calculated using a pre-trained 6DRepNet model [11]. Data is sampled at a rate of 30 Hz. The window for prediction is also five seconds, creating a nine-feature sequence with a length of 150 per sample.

Although it is possible to use the audio from the individual pin microphones, we found that they picked up speech from other participants which led to severe cross-talk. For this reason, we extract audio from each participant using the signal from the microphone array. We apply a speech separation algorithm to the microphone array signal which enhances audio from a particular direction and

suppresses it from others [18]. This can also function in real-time, making it suitable for this work. The audio embeddings themselves are calculated using the WavLM Transformer-based model [8]. We apply the model to the previous five seconds of audio for each participant. Each embedding is averaged to produce a final embedding for each participant of 768 dimensions.

Text information is converted to embeddings using the SBIntuitions Sarashina sentence transformer model for Japanese¹, which produces a vector of 1728 dimensions for each utterance. We add speaker-specific information to the utterance before feeding it to the model (e.g. “Speaker A: What do you think about that”). We use the sequence of utterances with a context up to the last 10 turns of the dialogue. All participant utterances are used in this sequence.

5.2 Architecture

For the individual modalities of gaze, head pose and text, we use three bi-directional LSTMs [29] to capture sequential information. After the Bi-LSTM encoding, a fully connected network is used for the final classification. The hidden layer of the Bi-LSTMs is 64 dimensions for gaze and head pose and 128 dimensions for text. For the audio model, we only have the final embeddings to capture the sequence, so we concatenate all three of these to produce a 2304-dimension vector which is fed into a linear projection to reduce the embedding to 128 dimensions.

Multimodal models are constructed by concatenating the final hidden state of the Bi-LSTMs, and the audio model’s linear projection. These concatenated embeddings are then used as input to a fully connected network which is adapted to the specific task. We can create models with any combination of modalities using this architecture by including and excluding the embeddings of each modality.

¹<https://huggingface.co/sbintuitions/sarashina-embedding-v1-1b>

For both individual and multi-modal models we also implemented an attention mechanism for sequential data. Through initial experimentation we found that self-attention [34] produced weaker models, probably due to the relatively low number of training samples which caused overfitting. However, we found that simple additive attention [6] may potentially improve the model while using fewer parameters. We integrated this attention mechanism to all the Bi-LSTM models as a means of comparison. Figure 4 provides an overview of the architecture of the models.

6 Model training and results

For each of the four tasks we trained a total of 21 models. These models are:

- Individual modality models without attention (4 total)
- Individual modality models using Bi-LSTMs with attention (3 total)
- All pair-wise combinations of modalities with and without attention (12 total)
- Multimodal models consisting of all features with and without attention (2 total)

All models were trained for 20 epochs using the Adam optimizer with a learning rate of 0.001. Due to the low sample size we could train a model in approximately 3-4 minutes with a laptop GPU.

We evaluated the models by using 12-fold cross-validation over 36 dialogue sessions. Each fold consisted of three sessions with the same trio of participants. This cross-validation provides a more robust measure of performance as opposed to a random split, because it ensures that a particular trio is not used in the training set and each sample is evaluated exactly once.

We also use two baselines as comparison models. The first is a weighted random baseline which makes a random prediction based on the distribution of the labels. The second is the result of an LLM, GPT 5.1 [27]. The text of the conversation is input into this model and the LLM is asked to output predictions for each of the tasks. For example, we use the below prompt for the **EXP** task:

"You are an expert in analyzing conversations. You will receive the full conversation context up to and including the final turn. Analyze only the final turn in the provided context. The conversation participants are A, B, and C. Do not judge earlier turns. Base your answer on the final turn and its conversational context. Return JSON only. Return exactly two keys: answer and reason. The answer value must follow the task-specific choices. The reason value should briefly explain the evidence from the final turn and context.

Task 1: Decide whether the final turn is explicitly addressed to exactly one person. Return 1 if the final turn explicitly addresses exactly one listener. Return 0 otherwise. Return 0 when the utterance is addressed to the whole group, to nobody in particular, or to someone other than A, B, or C."

[transcript of conversation leading up to turn end]

Similar prompts were used for the other tasks.

Table 2: Best performing model for each task

Task (sample #)	Best model	Prec.	Rec.	F1
EXP (3061)	Head + text w/attn.	0.670	0.666	0.667
PART (1655)	Head + audio w/attn.	0.892	0.892	0.892
CONT (3024)	Audio + text	0.697	0.695	0.696
MAIN (2774)	All (attn.)	0.680	0.680	0.680

6.1 Results

We now report the performance of the models. Table 2 summarizes the best model from cross-validation. Note that this does not imply statistical significance compared to the other models. We find that the sub-task of choosing which participant is the addressee has the best performance, with the highest F1 score around 0.89. On the other hand, the highest F1 scores for the other tasks are very similar, around 0.67-0.70. We now present a more detailed evaluation by comparing models for each task. We detail the performance of each model on the sub-tasks in Figure 5. In these comparisons, we use the average of the macro F1 score over all the folds and their 95% confidence intervals. Note that this is a different metric than in Table 2 which shows a single F-score over all the samples.

Results for the sub-task which predicts whether the addressee is an explicit participant (**EXP**) are shown in the top graph of Figure 5. We find that the text modality is the most important predictor here. Text alone and multimodal combinations which use text greatly outperform ones which do not have it. We find that the visual modalities (gaze and head pose) are worse than random chance. This can be intuitively explained by the nature of the task. Utterances by themselves provide better information on whether or not the addressee is an explicit participant. For example, the use of names and "you" would indicate this. On the other hand, gaze and head information by itself does not provide enough information as to whether someone is speaking to a person or to the group.

The opposite phenomena occurs in the **PART** sub-task, which predicts an explicit participant only. When there is an explicit participant, gaze and head information are strongly associated with increased performance of the model. Even unimodal models provide a macro F1-score greater than 0.8. The model has most likely learned that looking in the direction of a person is an indicator that they are the ones being addressed. The addition of audio also appears to slightly improve the models. Conversely, text information alone is a poor predictor of an explicit participant.

In the speaker continuation task (**CONT**), audio appears to be a strong modality. All multimodal models which contain audio outperform models which do not have it, and even audio alone is relatively strong. This could be due to the use of backchannels by the previous speaker who is continuing their line of conversation with the current speaker. For example, the audio embeddings may capture the verbal interactions between the previous and current speaker which produces a signal that their dialogue will continue.

Finally, for the main task (**MAIN**) we find that in general multimodality improves the performance of the models. The gaze + text models and full multimodal models are the best, although not significantly better than the other multimodal models. It should also

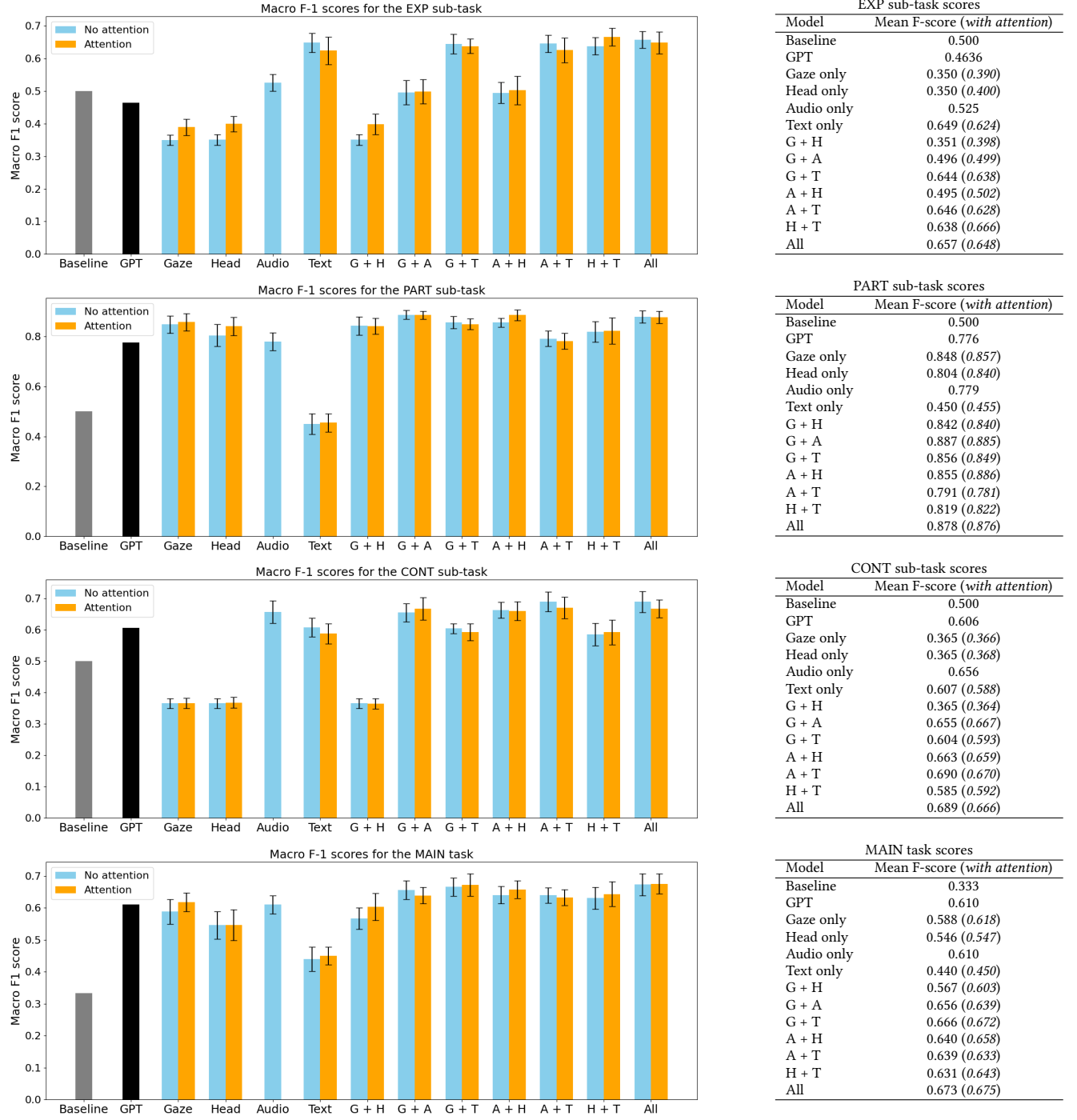


Figure 5: Average macro F-1 scores for each of the tasks, with 95% confidence intervals. Tables show the precise scores.

be noted that although text alone is relatively weak, the addition of other modalities greatly improves its performance.

In general, multimodal models outperformed the LLM across tasks. Perhaps surprisingly, the LLM was also outperformed by the text-only model in the **EXP** task and performed best in the

PART sub-task, even though this task is much more suited for visual features. These results suggest that a multimodal approach is needed. A multimodal LLM may have performed better on this task given video information, but would have been much slower and so is not suitable for our purposes.

We did not find evidence that the attention mechanism could improve the model. Initial experiments with self-attention degraded the model performance, and although additive attention did not have this effect, it also did not significantly improve the models.

We also analyzed the predicted labels of the **MAIN** task to better understand the errors. The confusion matrix of the All model with attention is shown in Figure 6. We find that the majority of errors are when the addressee is a participant and misclassified as everyone, as opposed to the other participant. This emphasizes the difficulty in detecting if an addressee is the entire group.

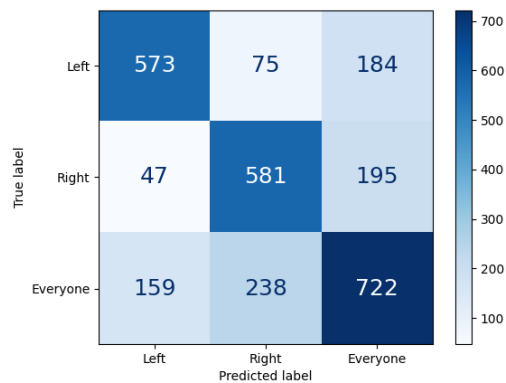


Figure 6: Confusion matrix for All with attention model on MAIN task.

6.2 Discussion

We have trained models for a variety of tasks related to addressee detection. We found that the performances of the models are related to the modalities used. Visual modalities can choose between explicit participants. Deciding if an addressee is an individual or the group itself requires text information. Audio is most useful for predicting whether the dialogue continues between a pair of participants. Our main task, to classify an addressee as one of the other participants or everyone, required multimodal information.

We also found that a multimodal approach could outperform the LLM, although the LLM was surprisingly effective, considering that it did not have access to any visual information. However, the text-only GPT model had an inference time of over one second, which is relatively slow for a real-time conversation system. A major benefit of our approach is that it is relatively lightweight.

This work has provided some insights which can be beneficial for research on multi-party turn-taking. Although other work in this area predicts the next speaker, we consider the intended addressee(s) of the speaker to be equally important. This allows us to implement a turn-taking model that could be compared with next speaker prediction in real-time conversation. Our model can potentially be adapted to continuously detect the addressee during the speaker's turn, making it useful for guiding listening behaviors in conversational robots. Our long-term goal is to create a conversational robot which can engage in multi-party free conversation

with humans with conversational dynamics such as turn-taking. This work is a step towards this goal.

6.3 Limitations and Future Work

In general, there is room for improvement in this task. Although we achieved results above the baseline, it is still not known if the main addressee detection model is robust enough for multi-party human-agent interaction. For example, we used ground-truth transcriptions for training the text models, whereas in a real-world conversation we must use results from automatic speech recognition which may contain errors.

The environment was also static, with participants in fixed positions. This means that gaze and head pose were influential for certain tasks such as predicting an individual addressee. We are investigating visual features which are independent of the positions of participants, such as the change in gaze angles.

Future work in this area involves the implementation and evaluation of addressee detection in a multi-party scenario. Our model can function only if the end of the turn is reliably predicted, and this task is also being investigated. The sub-tasks we addressed in this work are also of value to other conversation phenomena and the results can be used for comparisons in future research. We did not find attention mechanisms to be particularly effective, perhaps due to the relatively small sample size, so we are aiming to collect more data for this purpose.

7 Conclusion

In this work we tackled the issue of addressee detection in multi-party conversation. We describe how addressees were annotated for each turn in the TEIDAN corpus and showed that addressee detection is a distinct phenomenon from predicting the next speaker. Several classification tasks were formulated related to addressee detection. We used gaze, head pose, audio and text modalities in various combinations to train classification models for these tasks. It was found that different modalities were suitable for different tasks and that multimodality generally improved performance on them. We also further confirmed the necessity of multimodality by showing that these models could outperform an LLM. All models are lightweight and can function in real-time so we aim to integrate this into an embodied spoken dialogue system in future work.

Safe and Responsible Innovation Statement

This work could potentially have societal impacts due to the ability to analyze and classify human behavior using artificial intelligence. Our model involves the analysis of human behavior, so identities of participants in the conversation can be revealed by video, text and audio. We have taken care to ensure that we have authorization to use data that we have collected to train the models described in this paper. Our contribution should not be interpreted in any way as making a judgment on the behavior of participants in a conversation, nor should any of the models be used to evaluate a conversation in general.

Acknowledgments

This work was supported by JST Moonshot R&D JPMJPS2011 and JST PRESTO JPMJPR24I4.

References

- [1] Giulio Antonio Abbo, Maria Jose Pinto-Bernal, Martijn Catrycke, and Tony Belpaeme. 2025. Fast Multi-Party Open-Ended Conversation with a Social Robot. *arXiv preprint arXiv:2503.15496* (2025).
- [2] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. 2023. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*. IEEE, 98–102.
- [3] Angus Adlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernandez Garcia, Nancie Gunson, Weronika Sieńska, Christian Dondrup, and Oliver Lemon. 2024. 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*. 62–70.
- [4] Angus Adlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernández García, Nancie Gunson, Weronika Sieńska, Marta Romeo, Christian Dondrup, and Oliver Lemon. 2024. A multi-party conversational social robot using LLMs. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 1273–1275.
- [5] Oleg Akhtiamov, Maxim Sidorov, Alexey A Karpov, and Wolfgang Minker. 2017. Speech and Text Analysis for Multimodal Addressee Detection in Human-Human-Computer Interaction. In *Interspeech*. 2521–2525.
- [6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [7] Kent K Chang, Mackenzie Hanh Cramer, Anna Ho, Ti Ti Nguyen, Yilin Yuan, and David Bamman. 2026. Multimodal conversation structure understanding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7437–7458.
- [8] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.
- [9] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037* (2024). <https://doi.org/10.48550/arXiv.2410.00037>
- [10] Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: A transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2981–2990.
- [11] Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 2024. Toward Robust and Unconstrained Full Range of Rotation Head Pose Estimation. *IEEE Transactions on Image Processing* 33 (2024), 2377–2387. doi:10.1109/TIP.2024.3378180
- [12] Lukas Hilgert and Jan Niehues. 2025. 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)*. 60–71.
- [13] Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Multilingual turn-taking prediction using voice activity projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 11873–11883.
- [14] Koji Inoue, Divesh Lala, Mikey Elmers, Keiko Ochi, and Tatsuya Kawahara. 2025. An LLM Benchmark for Addressee Recognition in Multi-modal Multi-party Dialogue. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*. 330–334.
- [15] Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Predicting next speaker based on head movement in multi-party meetings. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2319–2323.
- [16] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, Ryuichiro Higashinaka, and Junji Tomita. 2019. Prediction of who will be next speaker and when using mouth-opening pattern in multi-party conversation. *Multimodal Technologies and Interaction* 3, 4 (2019), 70.
- [17] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Transactions on Interactive Intelligent Systems (TIIS)* 6, 1 (2016), 1–31.
- [18] Yuto Ishikawa, Kohei Konaka, Tomohiko Nakamura, Norihiro Takamune, and Hiroshi Saruwatari. 2024. 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, 730–734.
- [19] Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577* (2024).
- [20] Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *2019 International Conference on Multimodal Interaction*. 226–234.
- [21] Heeyoung Lee, Andreas Stolcke, and Elizabeth Shriberg. 2013. Using out-of-domain data for lexical addressee detection in human-human-computer dialog. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 221–229.
- [22] Meng-Chen Lee and Zhigang Deng. 2024. Online multimodal end-of-turn prediction for three-party conversations. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 57–65.
- [23] Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal turn analysis and prediction for multi-party conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 436–444.
- [24] Usman Malik, Julien Saunier, Kotaro Funakoshi, and Alexandre Pauchet. 2020. Who speaks next? Turn change and next speaker prediction in multimodal multiparty interaction. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 349–354.
- [25] Carlo Mazzola, Francesco Rea, and Alessandra Sciutti. 2023. Real-time addressee estimation: Deployment of a deep-learning model on the icub robot. *arXiv preprint arXiv:2311.05334* (2023).
- [26] Taiga Mori, Koji Inoue, Divesh Lala, Keiko Ochi, and Tatsuya Kawahara. 2026. Analysing Next Speaker Prediction in Multi-Party Conversation Using Multimodal Large Language Models. In *Proceedings of the 16th International Workshop on Spoken Dialogue System Technology*. 83–94.
- [27] OpenAI. 2025. GPT-5.1: A smarter, more conversational ChatGPT. <https://openai.com/index/gpt-5-1/> Accessed: April 21st 2026.
- [28] Rajarshi Roy, Jonathan Raiman, Sang-gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. 2026. PersonaPlex: Voice and Role Control for Full Duplex Conversational Speech Models. *arXiv preprint arXiv:2602.06053* (2026). <https://doi.org/10.48550/arXiv.2602.06053>
- [29] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.
- [30] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178.
- [31] Fiseha B Tesema, Jason Gu, Wei Song, Hong Wu, Shiqiang Zhu, Zheyuan Lin, Min Huang, Wen Wang, and Rajesh Kumar. 2023. Addressee detection using facial and audio features in mixed human-human and human-robot settings: A deep learning framework. *IEEE Systems, Man, and Cybernetics Magazine* 9, 2 (2023), 25–38.
- [32] David Traum et al. 2009. A comparison of addressee detection methods for multiparty conversations. In *Proc. DiaHolmia 2009*. 99–106.
- [33] TJ Tsai, Andreas Stolcke, and Malcolm Slaney. 2015. A study of multimodal addressee detection in human-human-computer interaction. *IEEE Transactions on Multimedia* 17, 9 (2015), 1550–1561.
- [34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [35] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karayannis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923* (2023).