Evaluation of Paralinguistic-Aware Spoken Dialogue Systems using Next-Utterance Classification

Kouki Miyazawa and Yoshinao Sato
Fairy Devices Inc.
{miyazawa,sato}@fairydevices.jp

Abstract

In spoken dialogues, paralinguistic cues frequently convey crucial information not captured by linguistic content alone. However, conventional spoken dialogue systems (SDSs), which typically comprise a cascade of an automatic speech recognition model and a large language model (LLM), lack the ability to recognize paralinguistic cues. Relying solely on transcribed text, paralinguistic-agnostic SDSs often cause dialogue breakdowns. To address this difficulty, integrating a paralinguistic recognition model into SDSs is essential. Therefore, this study focuses on evaluating such paralinguistic-aware SDSs. To this end, we propose a corpus-based evaluation method utilizing next-utterance classification as an automated alternative to human evaluation. Specifically, an LLM is tasked with predicting the dialogue act of the subsequent utterance given a transcribed dialogue history, comparing scenarios with and without paralinguistic attitude classes. Our experiments demonstrate that incorporating a paralinguistic attitude recognition model improves prediction performance, as measured by the mean reciprocal rank. Furthermore, we assessed the alignment of our proposed corpus-based evaluation method with subjective human evaluations. The results confirmed the validity of the proposed method as a reliable proxy for human evaluation, demonstrating its correlation with human rankings and agreement with human preferences in pairwise comparisons. Taken together, this work highlights that integrating paralinguistic recognition is a crucial step toward realizing more robust and natural spoken dialogue systems.

1 Introduction

In spoken dialogues, paralinguistic features, such as prosody, intonation, and voice quality, are essential for conveying information, such as emotion, attitude, and intent, not captured by linguistic content. In particular, the information conveyed by speech utterances with the same text can differ depending on paralinguistic cues (Hellbernd and Sammler, 2016). Human listeners can perceive paralinguistic cues and interpret the information they convey (Ishi et al., 2008)

Most recent spoken dialogue systems (SDSs) comprise a cascade of an automatic speech recognition (ASR) model, a text-based conversational model using a large language model (LLM), and a speech synthesis model. However, such systems cannot identify paralinguistic information. Paralinguistic-agnostic SDSs frequently fail to perceive the information expressed by speakers, causing dialogue breakdowns (Lin et al., 2024; Xue et al., 2024; Kang et al., 2025; Zhao et al., 2026).

Developing SDSs that comprehend paralinguistic cues requires assessing the contribution of paralinguistic recognition to the user experience. This study focuses on the evaluation of paralinguistic-aware SDSs. Integrating paralinguistic recognition models into ASR–LLM cascade systems provides them with paralinguistic capabilities. However, evaluating the effectiveness of a single module, namely a paralinguistic recognition model, within a complex SDS presents a significant challenge. A typical evaluation method involves users' subjective assessments, which are costly, thereby hindering the feasibility of large-scale evaluations. In addition, isolating a single factor of interest from the various factors that affect the subjective assessment of the overall behavior is challenging. In this study, we adopted a corpus-based offline evaluation method using next-utterance classification, which does not rely on users' subjective assessments. Specifically, we demonstrated the effectiveness of paralinguistic attitude recognition in SDSs by predicting the dialogue act (DA) of the next utterance from a transcribed dialogue history, both with and without paralinguistic attitude classes.

The major contributions of this study are as follows:

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 711–719

- We proposed an efficient evaluation framework that does not require users' subjective assessments for SDSs by employing a next-utterance DA prediction task.
- We confirmed that the proposed corpus-based evaluation method serves as a reliable proxy for subjective human evaluations.
- We demonstrated the effectiveness of incorporating a paralinguistic attitude recognition model to enhance an SDS by comparing paralinguistic-agnostic and paralinguistic-aware conditions.

The proposed evaluation framework is applicable to various human-computer interfaces that integrate LLMs with recognition models for non-linguistic modalities, including paralinguistic-aware SDSs.

2 Related Work

Previous studies have demonstrated the effectiveness of DAs for dialogue state tracking in task-oriented spoken dialogue, which often contain disfluencies such as fillers, hesitations, and repetition (Gulzar et al., 2025). They utilized DAs for evaluating dialogue systems, rather than for improving dialogue management. Task-oriented dialogue systems were evaluated by assessing the accuracy of predicted dialogue states comprising slot-value pairs, termed joint goal accuracy in (Feng et al., 2023; Gulzar et al., 2025). In the current study, SDSs were evaluated by predicting the DA of the next utterance.

Recurrent neural networks were developed for next-utterance classification in (Khanpour et al., 2016). However, their model only predicted DAs and did not generate response sentences. Conversely, the target SDS in the current study can generate response text. Moreover, our evaluation incorporated paralinguistic recognition in the evaluation.

The ability of multimodal LLMs to predict next utterances in human dialogue was investigated in (Yang et al., 2026). In these previous studies, the prediction performance was evaluated in terms of lexical overlap, such as bilingual evaluation understudy (BLEU) and recall-oriented understudy for gisting evaluation (ROUGE). Conversely, we employed the next DA prediction task to evaluate the target SDSs because inappropriate DAs in system responses negatively affect dialogues, whereas tolerance toward generated response text is diverse and often ambiguous.

3 Method

In this study, we employed a corpus-based offline evaluation method with next-utterance classification, which does not rely on users' subjective assessments. Using this method, the quality of an SDS is measured by its capacity to accurately identify the appropriate DA of the subsequent system response given a dialogue history (Fig. 1). Pairs of a dialogue context and the correct DA for the next utterance are extracted from a speech dataset of two-party human-human dialogues. The speaker of the next utterance is considered the system, and the interlocutor as the user. In particular, the DAs of the actual utterances in the corpus are considered appropriate. Therefore, the quality of the SDS can be evaluated based on the performance of the subsequent DA prediction at each turn transition using the LLM employed in the target system.

Typically, multiple DAs may be considered appropriate, although only one is observed in the corpus. To consider multiple options for appropriate DAs, the LLM ranks the appropriate DAs. Consequently, the performance of next utterance classification is measured using the mean reciprocal rank (MRR) calculated as

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}, \tag{1}$$

where Q and rank_i denote a set of queries and the rank position of the correct DA among the predicted DAs for the i-th query, respectively.

The effectiveness of the paralinguistic recognition model was evaluated by comparing two conditions, namely paralinguistic-agnostic and paralinguistic-aware conditions. In the former, the DA of the subsequent utterance is predicted solely from the transcribed texts in the dialogue context. In the latter, paralinguistic information contained in past utterances is also available.

We used the DA taxonomy proposed in (Araki et al., 1999), as presented in Table 1, with some modifications. This tagging scheme was specifically designed for task-oriented dialogues. Although more complex DA frameworks, such as that defined by ISO standards (Bunt et al., 2012) exist, they are not necessary for the purpose of this study. Therefore, we simplified the original DA taxonomy into 11 classes for the prediction task, as shown in

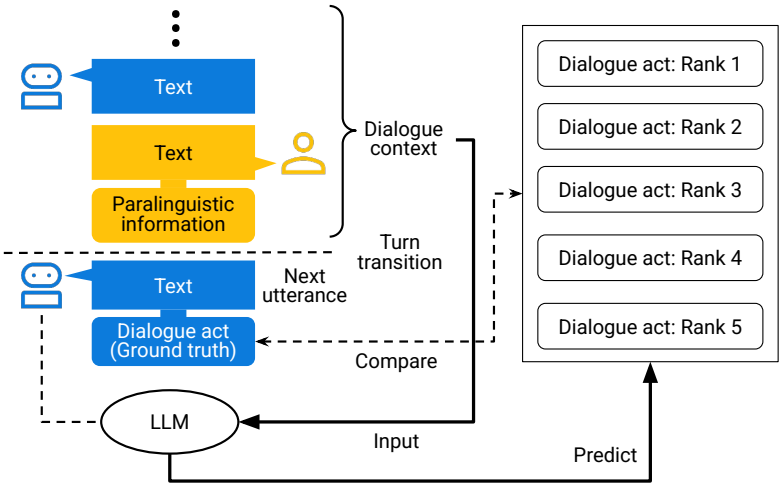

Figure 1: Next-utterance classification. The DA of the system response is predicted by the LLM based on the dialogue context, which comprises the transcribed texts and their associated paralinguistic information. The ground truth DA provided in the speech corpus and the top-5 predicted DAs are compared.

Table 2. The opening and closing classes appear only in specific contexts, while the remaining initiate and response classes are treated as exceptional cases. The hold class can occur in any context to delay the expression of other DAs; these classes are therefore discarded. In addition, DAs with ambiguous differences between them are merged into combined classes, namely suggest and request, proposal and invitation, and promise/offer and wish classes. This results in the final simplified 11-class DA taxonomy.

For the paralinguistic information, we used the four attitude classes investigated in (Miyazawa et al., 2025), as presented in Table 3. The user's attitude is crucial in generating the system response. Moreover, these attitudes are typically associated with the four main types of boundary pitch movement at the end of prosodic phrases (Igarashi and Koiso, 2012). Hence, speech utterances with identical text can convey different attitudes depending on paralinguistic cues. The paralinguistic attitude recognition resolves ambiguity in speech understanding that arises when the SDS relies solely on lexical information. Note that the paralinguistic attitudes used in this study are meaningful only at the last utterance of each transition construction unit. As these attitudes serve as dialogue control signals to elicit specific reactions from the interlocutor, they constitute a valid message exclusively at the end of a turn, rather than in the middle of a speaker's turn.

LLMs are not used as judges in the proposed method. In most LLM-as-a-judge frameworks, external LLMs are used to assess a target system. By contrast, in the proposed method, the LLM inside the target system is evaluated via next-utterance prediction using spoken dialogue corpora. Hence, the LLM operates as an actor, rather than a critic.

4 Data

The Chiba University Japanese Map Task Dialogue Corpus (MapTask) (The Japanese MapTask Corpus Project, 2007) was used as a spoken dialogue corpus. It comprises a collaborative task in which an instructor virtually guides a remote follower toward a destination through an audio channel. The two maps used by both parties differ, with information missing complementarily. Therefore, the participants are expected to resolve misunderstandings through spoken dialogue. The MapTask dataset comprises 128 dialogue sessions involving 64 participants, with a duration of 23 h. All 53,109 utterances are manually transcribed with timestamps in inter-pausal units.

5 Model

We employed the paralinguistic attitude recognition model proposed in (Miyazawa et al., 2025). It comprises a HuBERT backbone with a lightweight head containing fully connected layers; it is designed to classify an input speech utterance into one of four attitude classes and output the corresponding scores. Notably, this model infers the attitude of an input speech utterance solely from its acoustic features, without relying on transcribed text. In other words, paralinguistic recognition

Table 1: Dialogue acts modified and translated from (Araki et al., 1999).

Conventional	Opening	Opening part of the dialogue (e.g., greetings) not contributing to problem solving.
	Closing	Ending part of the dialogue (e.g., greetings, thanks) after problem solving.
Initiate	Suggestion	Request for action where the listener is not required to respond.
	Request	Request for action where the listener is required to respond.
	Proposal	Proposal for joint action where the listener is not required to respond.
	Invitation	Proposal for joint action where the listener is required to respond.
	Confirmation	Question where the speaker has a prediction about the listener's response based on context.
	Yes-No Question	Question asking for the correctness of information without speaker prediction.
	Wh-Question	Question asking for a specific value or expression without speaker prediction.
	Promise/Offer	Proposal of the speaker's own action.
	Wish	Statement describing the speaker's target state.
	Inform	Statement of the speaker's knowledge, opinion, or what they believe to be facts.
	Gratitude/Apology	Other statements such as expressions of gratitude or apology.
	Other Initiative	Acts such as dialogue management.
Response	Positive/Accept	Affirmative response to a Yes-No Question, or acceptance of a Request/Invitation.
	Negative/Reject	Negative response to a Yes-No Question, or refusal of a Request/Invitation.
	Answer	Response providing the value asked in a Wh-Question.
	Hold	Response delaying the answer while acknowledging the obligation to respond.
	Other Response	Response explicitly refusing to answer, etc.
Follow-up	Acknowledge	Utterance following a response, indicating that the purpose of the exchange has been achieved.

was performed independently of linguistic information. Therefore, in the proposed evaluation method, the linguistic and paralinguistic information fed into the LLM to predict the next-utterance dialogue were complementary.

Gemini 2.5 Flash (Comanici et al., 2025) was used to predict the DAs in our experiments. As the transcripts were provided in the MapTask corpus, no ASR model was used, eliminating speech recognition errors.

6 Experiments

We evaluated SDSs using the proposed corpus-based evaluation method.

6.1 Data Preparation

The utterances for which DAs were to be predicted were selected from the MapTask corpus as follows:

- Utterances following turn transition points were selected.
- Backchannels and fragments that did not function as DAs were excluded.

To identify backchannels and fragments, we used three LLMs: Claude Sonnet 4.6 (Anthropic, 2026), Gemini 2.5 Flash (Comanici et al., 2025), and GPT-5 mini (Singh et al., 2025). Utterances for which any of the above LLMs classified as a backchannel or fragment based on the transcription of the

Table 2: Simplified dialogue acts used in this study.

Suggestion/Request
Proposal/Invitation
Confirmation
Yes-No Question
Wh-Question
Promise/Offer/Wish
Inform
Gratitude/Apology
Positive/Accept
Negative/Reject
Answer

dialogue session were excluded. In the prompts presented to the LLMs, backchannels and fragments were specified as short reactions of listeners that do not take the conversational floor and incomplete utterances, such as fillers, hesitations, and meaningless vocalizations, respectively.

The MapTask corpus did not provide ground truth for the DAs. Rather, the silver standard (i.e., enough reliable annotations generated automatically by pre-trained models (Rebholz-Schuhmann et al., 2010; Giorgi and Bader, 2018)) of the simplified DAs was determined using the above LLMs. Here, utterances that cannot be classified into any of these classes were categorized as the other act class. The inter-annotator agreement, measured by Fleiss's κ, was 0.485. This moderate agreement is primarily because a single utterance can inherently convey multiple DAs, as discussed in (Araki et al., 1999), whereas the LLMs were restricted to selecting only a single DA, and thus does not imply any flaw in the annotation process. To avoid ambiguous DA labels, the silver standard was determined by unanimous agreement among the three LLMs; utterances for which a consensus was not reached in determining their DAs were excluded. Consequently, a dataset comprising 7,532 utterances with DA labels was obtained for our experiments.

To verify the validity of the silver standard, three experts annotated DAs to a subset comprising 100 utterances. The subset was randomly chosen while ensuring balance in the silver standard DAs. Each worker classified DAs of all the above utterances. The accuracy of the silver standard was 0.90 when considering the expert classification as the ground truth. Therefore, our dataset comprised utterances with unambiguous DAs, and the DA labels were reliable. Note that this dataset was used only for evaluation, but not for training.

The paralinguistic information, comprising paralinguistic attitude classes and their corresponding scores, was obtained using the paralinguistic attitude recognition model described in Section 5. Note that the paralinguistic information was annotated only for the utterances preceding turn transition points.

6.2 Experimental Setup

We performed the evaluation in the following conditions:

Paralinguistic-agnostic The dialogue context comprised the transcribed texts of past utterances.

Paralinguistic-perturbed The dialogue context comprised the transcribed texts of past utterances, along with the paralinguistic attitudes of past utterances that were randomly generated.

Paralinguistic-aware The dialogue context comprised the transcribed texts of past utterances, along with paralinguistic attitudes of past utterances that were inferred using the paralinguistic attitude recognition model.

The dialogue history was limited to 60 s. The LLM was used to rank the five appropriate DAs in order of likelihood. The performance was evaluated using the MRR with the top-5 results (MRR@5), along with top-1 accuracy.

The linguistic and paralinguistic information of each utterance was combined in the prompt provided to the LLM as follows:

```
[text] (paralinguistic attitude: {
  'agreement': [score_for_agreement],
  'disagreement': [
      score_for_disagreement],
  'question': [score_for_question],
  'stalling': [score_for_stalling]})
```

where the square brackets represent place holders. In the prediction task, the simplified DAs were used.

6.3 Results

The results are presented in Table 4. The paralinguistic-aware system substantially outperformed the other systems. These results demonstrate the effectiveness of paralinguistic information for SDSs in predicting appropriate DAs.

Table 3: Paralinguistic attitudes proposed in (Miyazawa et al., 2025).

Class	**Definition**	**Expected reaction**
Agreement	in favor, accept to continue	perform the approved action, moving on to the next
Disagreement	against, dissatisfied, request to stop	cancelling the rejected action, asking for instructions
Question	not understand, confirm facts, listen back	answering the question, rephrasing the previous utterance
Stalling	thinking, worried, request to wait	waiting for instructions, providing additional information

To rigorously assess performance improvements, we conducted permutation tests with 10,000 iterations and the Holm-Bonferroni correction at a significance level of 0.05 to compare the paralinguistic-aware system with the paralinguistic-agnostic and perturbed systems. For effect sizes, we report the mean difference. As a result, the paralinguistic-aware system significantly outperformed the agnostic baseline in MRR ($\Delta\mathrm{MRR}{=}0.029$, $p{=}0.0004$) and accuracy ($\Delta\mathrm{Accuracy}{=}0.032$, $p{=}0.0004$). Furthermore, the paralinguistic-aware system achieved statistically significant gains compared to the perturbed baseline in MRR ($\Delta\mathrm{MRR}{=}0.016$, $p{=}0.0004$) and accuracy ($\Delta\mathrm{Accuracy}{=}0.019$, $p{=}0.0004$).

Interestingly, the paralinguistic-perturbed system was slightly superior to the paralinguistic-agnostic system. This suggests the importance of considering paralinguistic information to predict DAs even when scores are unreliable. This slight improvement may be attributed to a priming effect; the explicit inclusion of attitude labels in the prompt could have guided the LLM to implicitly focus on the speaker's conversational intent, even when accompanied by unreliable scores. Furthermore, the system achieved higher performance using paralinguistic information inferred by the paralinguistic attitude recognition model, as expected. Hence, our experimental results show the importance of a more accurate paralinguistic model for SDSs.

7 Subjective Assessment

To verify the validity of the proposed evaluation method in terms of human subjective evaluation, we measured its correlation with human evaluations of the next utterance DAs and agreement with human preference in pair-wise comparison of system responses.

Table 4: Performance of the next DA prediction. MRR@5 and accuracy denote the MRR within the top five results and top-1 accuracy, respectively. Values are reported as mean ± 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals (CIs).

Paralinguistic	**MRR@5**	**Accuracy**
Agnostic	0.634 ± 0.008	0.463 ± 0.011
Perturbed	0.647 ± 0.008	0.476 ± 0.011
Aware	$\mathbf{0.663} \pm 0.008$	$\mathbf{0.496} \pm 0.011$

To evaluate the correlation between the appropriateness of next-utterance DAs predicted using the model and humans, we selected a small subset of our dataset with human ranking. First, 100 utterances were randomly selected while maintaining the balance of the silver standard DAs. Then, three experts evaluated all the simplified DA classes as an act of the next utterance based on the past dialogue history on a scale of one to four (greater is better). Here, both text and audio of the dialogue context were accessible for the evaluators. For each utterance, the evaluation scores assessed by humans were considered as a ranking of DAs with ties allowed. Consequently, we obtained $3 \times 100 \times 11$ values ranging from one to four.

Table 5: Correlation between model and human rankings. Values are reported as mean ± 95% BCa bootstrap CIs.

Paralinguistic	**NDCG**	**Kendall's τ**
Agnostic	0.829 ± 0.014	0.482 ± 0.028
Perturbed	0.844 ± 0.012	0.503 ± 0.026
Aware	$\mathbf{0.860} \pm 0.012$	$\mathbf{0.531} \pm 0.024$

The correlation was measured using a nor-

malized documented cumulative gain (NDCG) (Järvelin and Kekäläinen, 2002) and a Kendall's τ (Kendall, 1938). Table 5 presents the results. The paralinguistic-aware system achieved higher correlations with the human rankings. This result shows that paralinguistic recognition capabilities enhance the ability of SDSs to generate responses that appear more natural to users.

To rigorously assess the statistical significance of the correlation metrics, we conducted permutation tests with 10,000 iterations and applied the Holm-Bonferroni correction at a significance level of 0.05 to compare the paralinguistic-aware system against the paralinguistic-agnostic and paralinguistic-perturbed systems. To quantify the effect sizes, we employ the mean difference for NDCG and Cohen's q for Kendall's τ. As a result, the paralinguistic-aware system significantly outperformed the agnostic baseline in NDCG (ΔNDCG=0.030, p=0.0008) and Kendall's τ (Cohen's q=0.052, p=0.0008). Furthermore, the paralinguistic-aware system achieved statistically significant gains compared to the perturbed baseline in NDCG (ΔNDCG=0.015, p=0.0049) and Kendall's τ (Cohen's q=0.040, p=0.0095).

In addition, to investigate whether the differences in the evaluation results obtained using the proposed method are perceptible to humans, we performed a preference test with participants. First, we selected utterances where the DAs predicted by the paralinguistic-agnostic and paralinguistic-aware systems did not match. To validate the correctness of the DA prediction, we considered three conditions:

Agnostic-Only Correct The paralinguistic-agnostic system predicted the correct DA, whereas the paralinguistic-aware system predicted an incorrect DA.

Both-Incorrect Both systems failed to predict the correct DAs.

Aware-Only Correct The paralinguistic-agnostic system predicted an incorrect DA, whereas the paralinguistic-aware system predicted the correct DA.

For each condition, 702 utterances were randomly selected, totaling 2106 utterances. For each utterance, the past dialogue history was retained and the utterance of interest was masked as the next utterance to be predicted. Next, the two systems generated the texts for the next utterance based on the given dialogue context. For the paralinguistic-aware system, the context also included the paralinguistic attitudes of the utterances immediately preceding the turn transitions in the format provided in Section 6.2. Finally, 34 non-expert workers performed pairwise comparisons, judging the more natural-sounding text in each pair on a scale of one to four. Here, a higher score indicates that the paralinguistic-aware system generated a more appropriate response than the paralinguistic-agnostic system. Each subject assessed 120 stimuli. Consequently, 4080 trials were conducted in total.

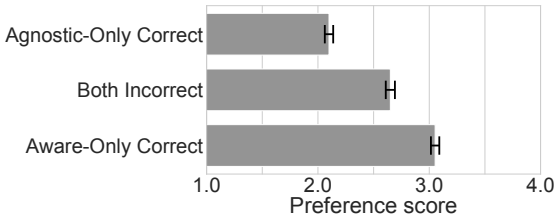

Figure 2: Human preference scores on a scale of one to four. A greater score means that the paralinguistic-aware system performed better than the paralinguistic-agnostic system. Error bars indicate standard deviations.

The results are presented in Fig. 2. Notably, a system that correctly predicted the next-utterance DA was more highly rated than a system that failed the prediction. This result indicates that failures to select an appropriate DA often trigger dialog breakdowns and negatively affect subjective evaluation, while the overall subjective evaluation depends on various other factors, such as response content, timing, and speech quality. Moreover, the paralinguistic-aware system obtained fairly better evaluations than the paralinguistic-agnostic system even when both systems predicted incorrect DAs. Therefore, improving the performance of the DA prediction will likely contribute to the overall subjective quality of the SDS. These results validate the proposed corpus-based evaluation method as a reliable proxy for subjective human evaluations.

8 Conclusion

This study proposed a corpus-based offline evaluation method for paralinguistic-aware SDSs to verify the effectiveness of utilizing paralinguistic information. Specifically, we evaluated the performance of next DA prediction using a LLM employed in the SDS, using dialogue contexts both with and without paralinguistic information. Our experimen-

tal results demonstrate that SDSs can select more appropriate responses when using paralinguistic information than when such information is neglected. Therefore, these results show that paralinguistic information should be incorporated to enable SDSs to communicate smoothly with humans. Moreover, the proposed corpus-based evaluation method was validated in terms of its alignment with subjective human assessments.

The proposed evaluation framework can be applied to other paralinguistic-aware SDSs. By providing a scalable alternative to costly subjective assessments, we believe our approach will significantly accelerate the development of affective and multimodal dialogue systems.

Although this study demonstrated the effectiveness of the proposed evaluation method using a task-oriented dialogue dataset, its applicability to non-task-oriented dialogues remains a subject for future research. Moreover, the current DA taxonomy was specifically designed for task-oriented dialogues; future studies will explore DA taxonomies suitable for open-ended and chat-oriented conversations. Furthermore, paralinguistic information tends to convey more diverse emotional and conversational nuances beyond the paralinguistic attitude classes used in this study. These extensions will enable us to more comprehensively assess the impact of paralinguistic-aware SDSs across broader domains.

References

Anthropic. 2026. System card: Claude Opus 4.6.

Masahiro Araki, Toshihiko Itoh, Tomoko Kumagai, and Masato Ishizaki. 1999. Proposal of a standard utterance-unit tagging scheme (in japanese). *Journal of the Japanese Society for Artificial Intelligence*, 14(2):251–260.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards LLM-driven dialogue state tracking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 739–755.

John M Giorgi and Gary D Bader. 2018. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087—-4094.

Haris Gulzar, Monikka Busto, Akiko Masaki, Takeharu Eda, and Ryo Masumura. 2025. Leveraging LLMs for written to spoken style data transformation to enhance spoken dialog state tracking. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 1743–1747.

Nele Hellbernd and Daniela Sammler. 2016. Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, 88:70–86.

Yosuke Igarashi and Hanae Koiso. 2012. Pitch range control of japanese boundary pitch movements. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 1949–1952.

Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication*, 50(6):531–543.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422—446.

Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, and Ozlem Kalinli. 2025. Frozen large language models can perceive paralinguistic aspects of speech. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 4323–4327.

Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2012–2021, Osaka, Japan.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-Yi Lee, and Ivan Bulyko. 2024. Paralinguistics-enhanced large language modeling of spoken dialogue. In *Proceedings of the*

IEEE International Conference on Acoustics, Speech and Signal Processing, pages 10316–10320.

Kouki Miyazawa, Zhi Zhu, and Yoshinao Sato. 2025. Paralinguistic attitude recognition for spoken dialogue systems. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 137–142, Bilbao, Spain. Association for Computational Linguistics.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, and Udo Hahn. 2010. The CALBC silver standard corpus for biomedical named entities — a study in harmonizing the contributions from four independent named entity taggers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, page 568—573, Valletta, Malta. European Language Resources Association (ELRA).

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. OpenAI GPT-5 system card. *Preprint*, arXiv:2601.03267.

The Japanese MapTask Corpus Project. 2007. Chiba university japanese map task dialogue corpus (MapTask).

Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Mengzhe Chen, Qian Chen, and Lei Xie. 2024. E-Chat: Emotion-sensitive spoken dialogue system with large language models. In *Proceedings of the International Symposium on Chinese Spoken Language Processing*, pages 586–590.

Yueyi Yang, Haotian Liu, Fang Kang, Mengqi Zhang, Zheng Lian, Hao Tang, and Haoyu Chen. 2026. SayNext-Bench: Why do LLMs struggle with next-utterance prediction? *Preprint*, arXiv:2602.00327.

Zhixian Zhao, Shuiyuan Wang, Guojian Li, Hongfei Xue, Chengyou Wang, Shuai Wang, Longshuai Xiao, Zihan Zhang, Hui Bu, Xin Xu, Xinsheng Wang, Hexin Liu, Eng Siong Chng, Hung yi Lee, and Lei Xie. 2026. The ICASSP 2026 HumDial challenge: Benchmarking human-like spoken dialogue systems in the LLM era. *Preprint*, arXiv:2601.05564.