Exploring Emotional Nuances in Spoken Dialogue: Dataset Construction and Prediction of Emotional Dialogue Breakdown

Hyuga Nakaguro
Nara Institute of Science and Technology
RIKEN
nakaguro.hyuga.nl1@is.naist.jp

Koichiro Yoshino
Institute of Science Tokyo
RIKEN
Nara Institute of Science and Technology
koichiro.yoshino@riken.jp

Abstract

In spoken dialogue systems, even when the utterance text is identical, variations in speaking style or tone can convey different nuances. To respond appropriately in such situations, systems must be able to interpret paralinguistic information effectively. This study evaluates this capability using the **paraling-dial** dataset. This dataset was constructed by pairing a fixed utterance text with five distinct emotional expressions and gathering corresponding responses. Based on this resource, a task was established to detect the inconsistency between the utterance's emotion and the response's content, which was then used to evaluate existing dialogue models. Existing dialogue models showed insufficient performance on this task. To address this issue, we hypothesize that emotion should function not merely as an additional feature but as a control signal that dynamically modulates textual interpretation. Based on this idea, we propose a Feature-wise Linear Modulation (FiLM)-based model. Experiments show that the proposed model achieves 93.8% accuracy with gold emotion labels and 91.2% with predicted labels, demonstrating both the effectiveness and practicality of our approach. Furthermore, a comparison of control signals with different abstraction levels—emotion labels, emotion embeddings, and acoustic features—reveals that the highest-level abstraction (emotion labels) yields the best performance. This result suggests that, in multimodal tasks, the appropriate level of abstraction, rather than richer information, is crucial for designing effective control signals.

1 Introduction

The development of Large Language Models (LLMs), exemplified by ChatGPT, has enabled the building of conversational systems capable of engaging naturally with humans (OpenAI et al., 2023). While these systems generate coherent and contextually appropriate responses in text-based interactions, they still struggle to adequately account for the nonverbal intent inherent in spoken utterances—particularly the tone, prosody, and other nuances that convey emotional and attitudinal meaning (Schuller et al., 2013; Guyer et al., 2021). To accurately interpret the intent behind spoken utterances, it is essential to incorporate paralinguistic information.

Models capable of processing such information include speech language model (SLM) (Chu et al., 2024), which directly takes spoken audio as input, and HuBERT (Hsu et al., 2021), which extracts and utilizes acoustic representations embedded in speech signals. To assess whether these systems can effectively recognize and exploit the intent conveyed through paralinguistic cues, benchmark tasks are necessary to quantitatively evaluate this capability.

In the evaluation of dialogue systems, dialogue breakdown detection—assessing whether system responses are generated appropriately within a dialogue context—serves as a key benchmark (Higashinaka et al., 2016). While this task shares technical similarities with response ranking (selecting the most contextually compatible response), we specifically adopt the framework of dialogue breakdown to focus on the detection of fatal inconsistencies that disrupt conversational flow. However, existing studies on dialogue breakdown have primarily focused on content-level breakdowns expressed through text. Yet, human communication is not merely a sequence of propositional meanings; it is fundamentally shaped by the speaker's tone, stance, and attitude, which often convey information that diverges from or even contradicts the literal text. For instance, Walker et al. (2012) demonstrated the importance of dialogic properties for stance classification (Walker et al., 2012), while Riloff et al. (2013) characterized sarcasm as a contrast between positive sentiment and negative situations (Riloff et al., 2013). In contrast, this study defines dialogue

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 95–103
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

breakdowns arising from mismatches in paralinguistic information as “emotional dialogue breakdown”, and introduces a paralinguistic-level dialogue breakdown detection task that has not been addressed in previous text-based research. This happens when the quality of the conversation drops a lot because the emotion expressed in the system’s response is very different from the user’s emotion, even though the actual words in the reply make sense.

To effectively capture the nuances of dialogue conveyed through paralinguistic information, we constructed a dataset by attaching audio data with different emotional tones to the same user utterances. This dataset, named **paraling-dial**, serves as the foundation for evaluating emotional dialogue breakdown. We additionally constructed a benchmark by shuffling assigned emotion labels and response sentences to simulate emotional inconsistencies. Using this benchmark, we conducted experiments with existing SLMs (Chu et al., 2024) to detect dialogue breakdowns, including emotional dialogue breakdown. However, the results only marginally exceeded chance level, indicating that current SLM frameworks struggle to effectively handle emotional dialogue breakdown. This limitation implies that the challenge lies not merely in model performance, but in how paralinguistic and textual information are integrated.

Based on these findings, we developed an emotional dialogue breakdown detector utilizing a model based on Feature-wise Linear Modulation (FiLM) (Perez et al., 2018), in which emotion labels serve as modulation signals for textual interpretation. Experimental results demonstrated that the proposed method achieved better performance on the emotional dialogue breakdown detection task. These findings offer important insights into the types of information and model architectures that should be considered in future SLM research to more effectively integrate paralinguistic and linguistic cues. Furthermore, while this study focuses on the detection of inconsistencies, our framework provides a foundation for developing low-cost automatic evaluation metrics for generative systems. This addresses a critical need in the current era of LLMs and SLMs, where the rapid increase in generated content has made traditional manual evaluation prohibitively expensive and difficult to scale.

2 Related Work

2.1 Dialogue Evaluation in Text

Automatic evaluation of dialogue system performance has traditionally been conducted by comparing generated responses to reference responses, inspired by evaluation methods in machine translation. Metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) are representative examples of reference-based evaluation measures. However, the correlation between reference-based metrics and human judgments of dialogue quality is not always high. Consequently, other automatic evaluation methods have been developed to assess factors such as fluency and naturalness (Liu et al., 2016).

Research on dialogue breakdown has focused less on fine-grained aspects of individual utterances and more on binary evaluations of whether an entire system response is consistent with its dialogue context. Analyses from this perspective emphasize the importance of whether the system response appropriately captures the intent and context of the dialogue. Moreover, such evaluation frameworks allow for the easy creation of Mismatch examples of breakdowns, for instance, by shuffling or swapping utterances. While response ranking—identifying the most appropriate response from a set of candidates—is often used in retrieval-based systems, it primarily focuses on optimizing response selection. In contrast, dialogue breakdown detection serves a critical role in quality assurance and anomaly detection for interactive systems. By framing our task as emotional dialogue breakdown detection rather than simple multi-modal response ranking, we aim to identify specific instances where emotional mismatches lead to a total failure of the conversational experience, a perspective increasingly important for the safety and reliability of generative speech models.

2.2 Dialogue Evaluation in Speech and Paralinguistic Information

Speech language models (SLMs) are capable of capturing the rich information embedded in speech, and models such as HuBERT (Hsu et al., 2021), as well as large-scale pre-trained speech models, have demonstrated strong performance across a variety of speech-based tasks. However, it remains unclear whether these models can adequately account for nuances and tones present in spoken dialogue, particularly the paralinguistic information that conveys

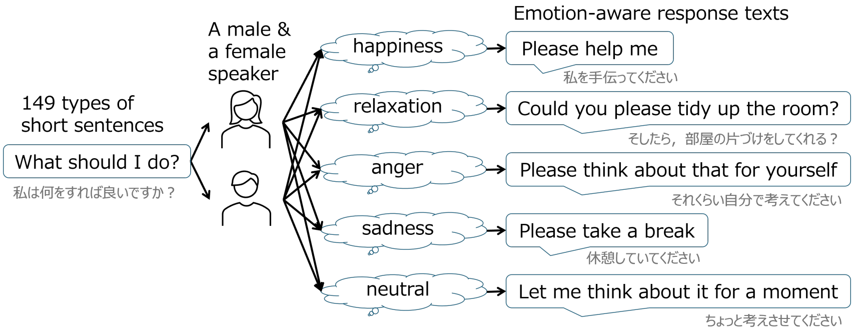
Emotion-aware response texts
A male &
a female
speaker
149 types of
short sentences
What should I do?
私は何をすれば良いですか？
happiness
relaxation
anger
sadness
neutral
Please help me
私を手伝ってください
Could you please tidy up the room?
そしたら，部屋の片づけをしてくれる？
Please think about that for yourself
それくらい自分で考えてください
Please take a break
休憩していてください
Let me think about it for a moment
ちょっと考えさせてください

Figure 1: Structure of the paraling-dial dataset. For the same utterance text, multiple speech recordings with different emotional expressions are paired with their corresponding responses.

subtle aspects of intent and emotion.

Traditionally, research on the nuances conveyed by paralinguistic cues in spoken utterances has largely relied on emotion labels. More recently, approaches such as WavReward (Ji et al., 2025) have attempted to evaluate dialogue quality by comprehensively considering both cognitive and emotional aspects of speech.

In this study, we extend the ideas from text-based dialogue breakdown detection to speech. Specifically, we propose an emotional dialogue breakdown detection task to evaluate whether systems can detect mismatches between the emotional nuances of an utterance and the corresponding system response.

2.3 Multimodal Information Integration

The limitations of existing systems stem from the framework used for multimodal information integration. Representative approaches include early fusion, late fusion, and joint embedding, yet none of these are necessarily well-suited for detecting emotional dialogue breakdown (D'mello and Kory, 2015).

- Early Fusion: Simply concatenates audio features and text embeddings, which can lead to imbalances in information contribution.
- Late Fusion: Processes each modality independently, which can result in the loss of subtle features.
- Joint Embedding: Assumes a symmetric correspondence between modalities, making it difficult to directly capture relationships in which emotion guides text interpretation.

In contrast, the feature-wise linear modulation (FiLM) architecture (Perez et al., 2018) adopted in this study uses paralinguistic information derived from speech as a control signal, dynamically applying scaling and shifting to the intermediate layers of textual representations. This design allows emotional nuances to directly influence the interpretation of text. In other words, FiLM differs from conventional feature concatenation by enabling integration through modulation of textual semantics rather than simple combination of features.

2.4 Emotional Dataset

Existing emotional speech dialogue datasets, such as IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2018), tend to vary the textual content of utterances along with changes in emotion. As a result, it is difficult to independently evaluate the influence of paralinguistic information on response selection. To rigorously analyze the effect of paralinguistic cues, it is essential to have a dataset in which the textual content remains fixed, while the nuances of speech—such as emotion labels—vary, leading to differences in the appropriateness of system responses. Constructing such a dataset is crucial for establishing a foundation for emotional dialogue breakdown detection tasks.

3 Emotional Dialogue Breakdown Detection Task

In this study, we define an emotional dialogue breakdown detection task in spoken dialogue and show that existing SLMs do not necessarily perform effectively on this task. We utilize the paraling-dial dataset, in which identical utterances are paired with different emotion labels and corresponding response sentences, to formulate this task.

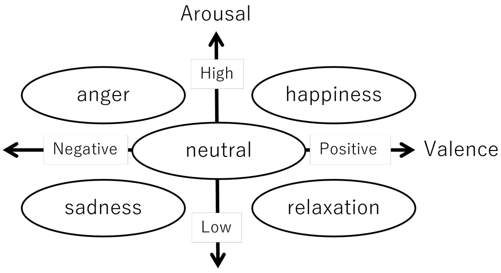

Figure 2: Relative positions of emotions presented to the speakers.

3.1 paraling-dial Dataset

The basic structure of the paraling-dial dataset constructed in this study is illustrated in Figure 1. For each single textual utterance, speakers produced recordings expressing five different emotions: happiness, relaxation, anger, sadness, and neutral. For each audio recording, a corresponding response sentence was manually annotated. Thus, for each textual utterance, the dataset contains as many utterance-response pairs as there are emotions, with each pair reflecting both the speech and the emotion-specific response.

The dataset was constructed in four steps:

1. Collection of utterances
2. Emotion assignment
3. Collection of emotion-aware responses
4. Recording of speech

Step 1: Collection of utterances 149 short dialogue-like sentences were collected from Aozora Bunko to serve as user utterances. Aozora Bunko was selected because it provides a vast and diverse range of literary expressions and dialogue styles, and as a copyright-free resource, it facilitates the open use and distribution of the resulting dataset.

Step 2: Emotion assignment Based on Russell's circumplex model of affect (Russell, 1980), five representative emotions—happiness, relaxation, anger, sadness, and neutral—were assigned to each utterance. These emotions were specifically chosen to span the four quadrants of the Valence-Arousal space, ensuring a comprehensive evaluation of paralinguistic cues. Within this framework, we distinguish between "Neutral" and "Relaxation": while Neutral represents a baseline state with medium valence and arousal, Relaxation is characterized by positive valence combined with low arousal, representing a calm and pleasant state that requires distinct linguistic and paralinguistic handling.

Step 3: Collection of emotion-aware responses Response sentences corresponding to each emotional variation were collected. These responses were created by a single trained annotator. To explore possible variations of responses for utterances with given paralinguistic cues, ChatGPT was used to generate candidate responses, which were then refined by the annotator to produce the final emotion-specific response variations. Annotators were instructed to produce as diverse responses as possible for each emotion label while remaining natural, in order to ensure that different paralinguistic cues would yield distinct responses. Examples of responses created for the same utterance under different emotions are shown in Table 1.

Step 4: Recording of speech Six speakers (three male and three female, all with professional experience in voice acting or theatrical performance) recorded each utterance with each assigned emotion, emphasizing the emotional differences. Recordings were conducted in a soundproof room using a directional microphone [1]. To ensure the quality and validity of the acted emotions, the first author supervised all recording sessions, providing real-time feedback and verifying that each utterance correctly reflected the intended emotional state. Speakers were shown Figure 2 and instructed to express the relative differences between emotions according to the positions depicted in the figure. In total, the dataset comprises 149 utterances × 5 emotions × 6 speakers = 4,470 utterance-response pairs. The total recording time was approximately 284.89 minutes, with an average duration of about 3.92 seconds per utterance.

3.2 Analysis of Acoustic Validity

To verify whether the speech recordings in paraling-dial acoustically reflect the intended emotions, we analyzed the distributions of fundamental frequency (F0) and root mean square (RMS) energy for each emotion label. F0 values above 700 Hz were excluded as likely pitch-tracking errors, since such frequencies are beyond the physiological range of human phonation. As shown in Figure 3, distinct trends can be observed for different emotions. For example, happiness is distributed over

[1] Sony ECM-674 Electret Condenser Microphone

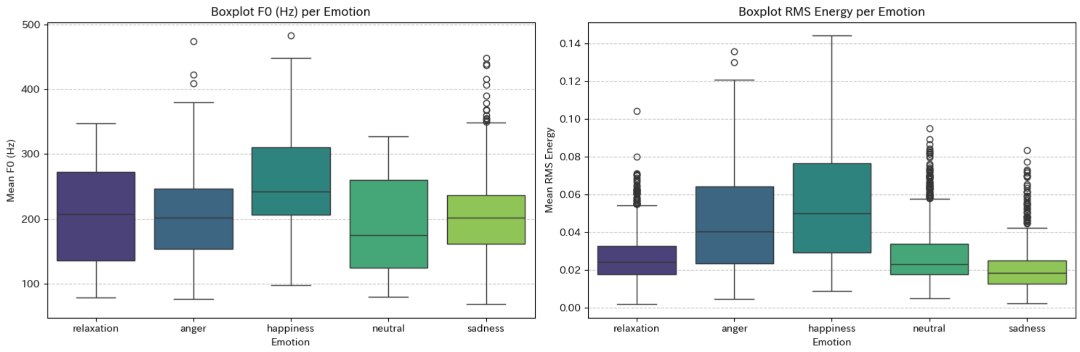

Figure 3: Analysis of paraling-dial audio. Distributions of fundamental frequency (F0, left) and root mean square energy (RMS, right) for each emotion.

Table 1: Example of responses in paraling-dial

Sentences	Emotions	Responses
What should I do? 私は何をすれば良いですか?	happiness	Please help me. 私を手伝ってください.
	relaxation	Could you please tidy up the room? そしたら，部屋の片づけをしてくれる?
	anger	Please think about that for yourself. それくらい自分で考えてください.
	sadness	Please take a break. 休憩していてください.
	neutral	Let me think about it for a moment. ちょっと考えさせてください.
Can you do it? やっていただけますか?	happiness	Yes, I will. 了解です.
	relaxation	Okay, I’ll do it. 分かりました
	anger	I’m sorry. I’ll do it now. すみません.今すぐにやります.
	sadness	I’ll do it for you, so cheer up.代わりにやっておくから元気出して.
	neutral	Yes, sir. 承知しました.
I have a favor to ask you.お願いしたいことがあります.	happiness	What up? なになに?
	relaxation	What is it?なんでしょう?
	anger	Is there a problem? 何か問題ありましたか?
	sadness	I don’t know if I can do it. 私にできるかな.
	neutral	What is it? なんでしょう?

a higher F0 range, while sadness is concentrated in a lower RMS energy range, indicating clear separation of acoustic features across emotion. On the other hand, relaxation and neutral show similar distributions, suggesting that these emotions are relatively close in the emotional space. These results confirm that paraling-dial possesses sufficient acoustic validity for analyzing paralinguistic information and evaluating models.

3.3 Construction of the Emotional Breakdown Dataset

The paraling-dial dataset consists of multiple audio recordings of the same utterance, each labeled with a different emotion, along with their corresponding response sentences. By swapping the response sentences with those corresponding to a different emotion, it is possible to simulate dialogue situations exhibiting emotional dialogue breakdown. Specifically, we created the following two classes of utterance-response pairs. The task of emotional dialogue breakdown detection is to predict Matched/Mismatched for a given pair of utterance audio and response sentence.

- Matched: A pair consisting of an utterance spoken with a specific emotion and its corresponding correct response.
- Mismatched: A pair consisting of an utterance spoken with a specific emotion and a response corresponding to a different emotion, intentionally mismatched.

To construct the Mismatched pairs, we grouped all utterance-response pairs by their common utterance text. Within each group, we kept the audio recordings (utterance + emotion) fixed and shuffled the response sentences among the five different emotion labels. This procedure ensures that while the textual content of the response remains logically and contextually consistent with the user's utterance text, a paralinguistic mismatch is introduced between the emotion conveyed in the speech and the intent of the response.

Using this constructed dataset, it is possible to evaluate a model's ability to handle emotional consistency under conditions where the textual content is kept constant, while manipulating only paralinguistic factors (i.e., emotion). This setup prevents the model from relying on simple text-based context matching and forces it to integrate paralinguistic information to achieve correct classification.

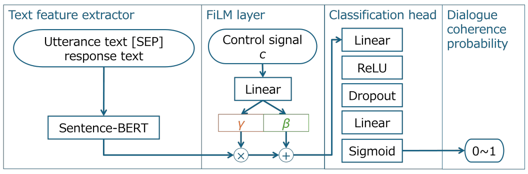

Figure 4: Proposed model architecture.

4 Emotion-based Modulation of Text Interpretation

Detecting emotional dialogue breakdown, as defined in this study, requires methods that can appropriately handle the nuances and tones conveyed by paralinguistic information in spoken utterances. While existing SLM training approaches have attempted to incorporate such information, it is not guaranteed that these cues are adequately utilized. Based on the hypothesis that emotion serves as a control signal that dynamically modulates the interpretation of text, we propose a new model for emotional dialogue breakdown detection that employs Feature-wise Linear Modulation (FiLM) (Perez et al., 2018).

4.1 FiLM-based Modulation Model

To instantiate the hypothesis that emotion can act as a control signal modulating text interpretation, we employ Feature-wise Linear Modulation (FiLM) in this study. The proposed model consists of three main components: a text feature extractor, a FiLM layer, and a classification head (Figure 4).

First, to extract textual information from the utterances, the input utterance text and the candidate response text are concatenated with a "[SEP]" token and fed into a pre-trained Sentence-BERT (Reimers, 2019). This produces a 768-dimensional feature vector x. Next, as a control signal, we use a 5-dimensional one-hot vector c representing the emotion labels (happiness, relaxation, anger, sadness, neutral). This vector serves as a high-level abstraction of the paralinguistic information associated with the text. In addition, we compare features extracted at various levels of abstraction related to paralinguistic information; details of the evaluation are described in Section 5. In the FiLM layer, the control signal c is fed into a fully connected layer to dynamically generate a scale parameter γ and a shift parameter β. The text feature vector x is modulated by the affine transformation defined in Eq. 1.

$$FiLM(x, c) = \gamma(c) \odot x + \beta(c) \qquad (1)$$

This operation allows the representation of x to adaptively change in the feature space depending on the emotional information c. Finally, in the classification head, the modulated 768-dimensional vector is processed through a network consisting of: Linear($768 \rightarrow 128$) $\rightarrow$ ReLU $\rightarrow$ Dropout(0.3) $\rightarrow$ Linear($128 \rightarrow 1$) $\rightarrow$ Sigmoid to output the probability that the dialogue is experiencing a breakdown.

5 Experimental Setup

5.1 Dataset

For the experiments, we used the emotional dialogue breakdown detection dataset constructed from the paraling-dial dataset described in Section 3. From the 4,470 utterance-response pairs, we generated Match examples, where the emotion and response matched, Mismatch examples, where the emotion and response were intentionally mismatched, in a 1:1 ratio. The dataset was split into training, validation, and test sets in an 8:1:1 ratio.

5.2 Control Signals for Comparative Models

In the proposed method, we use a high-level one-hot vector representing the emotion label as the control signal. To examine how the level of abstraction affects performance, we also constructed comparative models using control signals of varying abstraction levels:

- Low-level features (acoustic features): MFCC (Logan et al., 2000), MFCCΔ, MFCC$\Delta\Delta$, RMS energy, fundamental frequency (F0), mean spectral centroid, and standard deviation of the spectral centroid.
- Mid-level features (emotion embeddings): The acoustic features above were input to a two-layer neural network trained for emotion classification, and the resulting intermediate embedding vector was used as the control signal. The emotion classifier achieved 88.6% accuracy on the test set.
- High-level feature (one-hot vector): The acoustic features were input to a Random Forest classifier (Breiman, 2001) for emotion classification, and the predicted labels were converted into a one-hot vector to serve as the control signal. This classifier achieved 88.3% accuracy on the test set.

All FiLM models were trained using the common hyperparameters listed in Table 2.

Table 2: Common hyperparameters

Optimizer	Adam (Kingma, 2014)
Learning rate	0.001
Batch size	32
Epochs	100 (early stopping, patience=5)

6 Experimental Results and Discussion

In this section, we evaluate the effectiveness of the proposed method, analyze the impact of the abstraction level of the control signal on performance, and discuss the insights obtained from the results.

6.1 Effectiveness of the Proposed Method

First, we evaluated the emotional dialogue breakdown detection task using Qwen2-Audio. The model achieved an accuracy of 50.0%, suggesting that existing SLMs may not adequately capture the speech features required for the proposed emotional dialogue breakdown task. In other words, the paralinguistic information utilized by conventional SLMs appears insufficient for fully representing the nuances and tones of spoken utterances. Next, we examined the validity of our hypothesis that emotion can serve as a control signal dynamically modulating text interpretation by applying the FiLM-based model. When the ground-truth emotion labels were provided as high-level control signals, the FiLM model achieved 93.8% accuracy on the emotional dialogue breakdown task, confirming that the proposed architecture can detect emotional dialogue breakdown with high precision.

Furthermore, to simulate practical scenarios, we used the outputs of a separately trained emotion predictor (accuracy 88.3%) as the control signal. Under this setting, the FiLM model achieved 91.2% accuracy, demonstrating that the proposed approach is robust to errors in emotion prediction and can function effectively in real-world environments.

6.2 Performance Comparison by Abstraction Level of Control Signals

Next, we examined the impact of the abstraction level of the control signal on performance. Table 3 presents the accuracy achieved when using features of three abstraction levels: high-level (emotion label one-hot vector), mid-level (emotion embedding), and low-level (acoustic features).

These results indicate that low-level acoustic features contribute little to classification in the emotional dialogue breakdown detection task, whereas

Table 3: Accuracy for different abstraction levels of control signals

Abstraction Level	Accuracy
High-level (emotion level)	91.2%
Mid-level (emotion embedding)	72.0%
Low-level (acoustic features)	50.4%

higher-level features closer to the emotion label substantially improve accuracy. This suggests that, to appropriately capture the nuances and tone of utterances contained in paralinguistic information, it is essential to solve the task using features with an appropriate level of abstraction.

6.3 Discussion

The results presented above support the hypothesis of this study that emotion can be treated as a control signal dynamically modulating text interpretation. The FiLM architecture enables effective modeling of the asymmetric and complex interactions between emotion and textual features by dynamically modulating the text representation according to the emotional state.

Furthermore, the comparison of control signals at different levels of abstraction suggests that representations containing more low-level information are not necessarily optimal. Selecting an appropriate level of abstraction that aligns with the structure and objective of the task is crucial.

However, several limitations of this study should be noted. First, the paraling-dial dataset consists of acted speech rather than natural spontaneous dialogue. We intentionally opted for acted speech to create a highly controlled environment where only the emotional tone varies while the textual content remains strictly identical a condition that is extremely difficult to isolate in existing natural speech datasets. This allowed us to rigorously evaluate the impact of paralinguistic cues in isolation. Second, the current evaluation is based solely on objective metrics. While these metrics demonstrate the model's technical proficiency, conducting subjective human evaluations remains a crucial future task to confirm whether the detected "breakdowns" align with human perception of conversational naturalness. In the emotional dialogue breakdown detection task, representations closer to high-level emotion labels, such as one-hot vectors explicitly encoding discrete emotions, were found to be better suited for the task than lower-level acoustic or embedding features.

7 Conclusion

In this study, we aimed to capture the paralinguistic nuances of spoken utterances that should be considered in spoken dialogue systems. To this end, we constructed a benchmark dataset for emotional dialogue breakdown and developed a corresponding detection model. Specifically, we created the paraling-dial dataset, in which identical utterances were spoken with different emotion labels, and responses were assigned according to the emotional nuances of the utterances. Based on this dataset, we defined the emotional dialogue breakdown detection task and developed a FiLM-based detector that treats emotion as a control signal dynamically modulating text interpretation. The FiLM-based model achieved 93.8% accuracy using ground-truth emotion labels and 91.2% accuracy using predicted labels, confirming both the validity and practical applicability of our approach. Furthermore, we investigated the impact of the abstraction level of control signals. For this task, high-level features closer to discrete emotion labels were found to contribute most to classification accuracy, highlighting the importance of selecting task-relevant representations in multimodal learning. These findings suggest that, in future SLM training, it is crucial to leverage representations and signals that are highly relevant to the target task.

In future work, we plan to extend our framework to more natural, spontaneous speech environments. Furthermore, to move beyond the five discrete emotion categories used in this study, we aim to incorporate continuous emotion vectors or develop task-datasets that are not restricted to specific labels. Finally, integrating subjective user studies will be essential to further validate the practical utility of our emotional dialogue breakdown detector in real-world human-computer interaction.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 22H03654 and 22K17958. A part of this work is supported by JSPS KAKEN-HI grant number 23K24910.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of* *the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36.

Joshua J Guyer, Pablo Briñol, Thomas I Vaughan-Johnston, Leandre R Fabrigar, Lorena Moreno, and Richard E Petty. 2021. Paralinguistic features communicated through voice can affect appraisals of confidence and evaluative judgments. *Journal of nonverbal behavior*, 45(4):479–504.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Shengpeng Ji, Tianle Liang, Yangzhuo Li, Jialong Zuo, Minghui Fang, Jinzheng He, Yifu Chen, Zhengqing Liu, Ziyue Jiang, Xize Cheng, et al. 2025. Wavreward: Spoken dialogue models with generalist reward evaluators. *arXiv preprint arXiv:2505.09558*.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Beth Logan et al. 2000. Mel frequency cepstral coefficients for music modeling. In *Ismir*, volume 270, pages 1–11. Plymouth, MA.

OpenAI, :, Josh Achiam, et al. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 592–596.