Talk With Human-like Agents: Empathetic Dialogue Through Perceptible Acoustic Reception and Reaction

Haoqiu Yan[*1,3], **Yongxin Zhu**[*2,3], **Kai Zheng**[1,3],

Bing Liu[4], **Haoyu Cao**[4], **Deqiang Jiang**[4], **Linli Xu**[†1,3]

[1]School of Computer Science and Technology, University of Science and Technology of China
[2]School of Data Science, University of Science and Technology of China
[3]State Key Laboratory of Cognitive Intelligence, [4]Tencent Youtu Lab
{yanhq,zyx2016,dthdzk}@mail.ustc.edu.cn
{billbliu,rechycao,dqiangjiang}@tencent.com linlixu@ustc.edu.cn

Abstract

Large Language Model (LLM)-enhanced agents become increasingly prevalent in Human-AI communication, offering vast potential from entertainment to professional domains. However, current multi-modal dialogue systems overlook the acoustic information present in speech, which is crucial for understanding human communication nuances. This oversight can lead to misinterpretations of speakers' intentions, resulting in inconsistent or even contradictory responses within dialogues. To bridge this gap, in this paper, we propose PerceptiveAgent, an empathetic multi-modal dialogue system designed to discern deeper or more subtle meanings beyond the literal interpretations of words through the integration of speech modality perception. Employing LLMs as a cognitive core, PerceptiveAgent perceives acoustic information from input speech and generates empathetic responses based on speaking styles described in natural language. Experimental results indicate that PerceptiveAgent excels in contextual understanding by accurately discerning the speakers' true intentions in scenarios where the linguistic meaning is either contrary to or inconsistent with the speaker's true feelings, producing more nuanced and expressive spoken dialogues. Code is publicly available at: https://github.com/Haoqiu-Yan/PerceptiveAgent.

1 Introduction

Artificial Intelligence (AI) agents (Russell and Norvig, 2010; Negnevitsky, 2005) are entities designed to replicate human-like intelligence and functionalities, serving as the essential building blocks of AI systems. An ideal agent should be capable of perceiving its environment with sensors, making informed decisions, and then taking actions in response to users or scenarios. Recently,

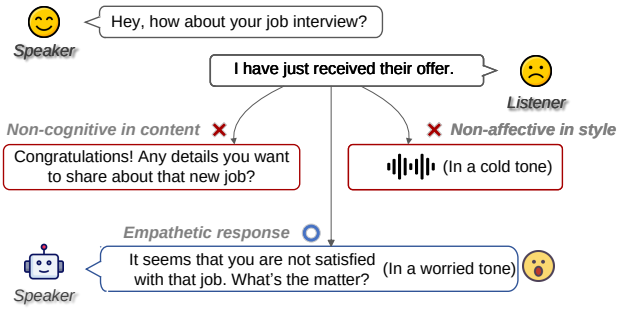

Figure 1: Examples illustrating the definition of empathy within dialogues.

Large Language Models (LLMs) (Wei et al., 2022; Shanahan, 2024; Taylor et al., 2022) have exhibited remarkable capabilities in diverse tasks, offering opportunities for building general AI agents that engage in human-like interactions, such as virtual assistants and intelligent robots. However, current text-only dialogue systems (Peng et al., 2023; Touvron et al., 2023) fall short in bridging the gap between experimental and realistic scenarios, where humans perceive and understand the world through diverse multi-modal information. Thus, the integration of acoustic information into dialogues has the potential to foster the development of more human-like agents, thereby enhancing the empathetic experience they offer.

Empathetic responses involve two essential aspects: cognitive and affective empathy (Cuff et al., 2016; Kim et al., 2021; Reis et al., 2011; Smith, 2006), which reflect an understanding of the human-talker's thoughts and feelings respectively. Specifically, cognitive empathy involves understanding the human-talker's thoughts, perspectives, and described events, enabling the agent to provide responses relevant to the dialogue topic (Sabour et al., 2022). Conversely, affective empathy entails responding based on observed emotional expressions in the dialogue history, contributing to the nat-

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15009–15022

uralness of synthesized speech (Cong et al., 2021; Guo et al., 2021; Nishimura et al., 2022). While recent works (Saito et al., 2023; Nguyen et al., 2022; Mitsui et al., 2023) leverage LLM's strong capabilities of contextual understanding and content generation to synthesize empathetic speeches, there remains a discrepancy between cognitive and affective empathy. This arises because cognitive content is preassigned before affective speech is deduced from latent representations of multi-modal dialogue history.

Recently, advancements in multi-modal content perception and generation have been achieved by various methods (Zhang et al., 2023; Huang et al., 2024; Chen et al., 2023; Wu et al., 2023), where audio is represented as either recognized text with an automatic speech recognition model or discrete features with a speech encoder. However, while linguistic information in speech is predominantly captured by both discrete acoustic units and textual representations, acoustic features tend to be disregarded. This oversight can lead to misinterpretations of the speaker's intentions, resulting in discrepant or even contradictory responses within the dialogue history. As illustrated in Figure 1, the left scenario fails to consider the perspective of the listener while the right one barely understands or empathizes with the speaker's feelings.

In this paper, we propose **PerceptiveAgent**, an empathetic multi-modal dialogue system that can discern deeper or more subtle meanings beyond the literal interpretations of words, based on speaking styles described in natural language. Specifically, PerceptiveAgent first comprehends the speaker's intentions accurately by a perceptive captioner model that captures acoustic features from each speech within dialogues. Subsequently, an LLM module acts as the cognitive core, producing the relevant response content with a caption describing how to articulate the response. A Multi-Speaker and Multi-Attribute Synthesizer (MSMA-Synthesizer) is then developed to synthesize nuanced and expressive speech.

Our contributions include the following:

- We pioneer the construction of a speech captioner model to perceive and express acoustic information through natural language.
- We develop an empathetic multi-modal dialogue system capable of identifying the speaker's true intentions through audio modality perception and generating empathetic speech.
- Experiments demonstrate that PerceptiveAgent can accurately discern the true intentions in scenarios where the literal interpretations of words are either contrary to or inconsistent with the speaker's true feelings.

2 Related Work

2.1 Multi-modal Dialogue Systems

Recent advances in multi-modal dialogue systems have primarily focused on transforming speech into discrete latent representation. For instance, Zhang et al. (2023); Chen et al. (2023); Wu et al. (2023) utilize speech encoders to perceive speech and then synthesize responses according to discrete acoustic units derived from LLMs, showing intrinsic cross-modal conversational abilities. Besides, works including (Nguyen et al., 2022; Mitsui et al., 2023) autonomously generate two-channel spoken dialogues, simulating realistic interactions between agents, including vocal interactions, laughter, and turn-taking. However, while discrete acoustic units capture linguistic information effectively, prosodic features are mostly ignored. To address this limitation and preserve prosodic information as much as possible, we develop a multi-modal dialog system that perceives prosody through speech captioning and responds empathetically using an LLM and a speech synthesizer.

2.2 Cross-Modal Text Generation

Cross-modal text generation involves generating text conditioned on other modalities such as audio and vision (Li et al., 2022; Liu et al., 2024; Zhang et al., 2024), where the key challenge is to align multi-modal features with the text latent space. Recent approaches (Zhu et al., 2023; Chen et al., 2023) address this challenge by aligning off-the-shelf pre-trained LLMs with learnable visual encoders (Li et al., 2023; Zhao et al., 2023), transforming multi-modal representations as learnable query embeddings while keeping both pre-trained LLMs and visual encoders frozen. Similarly, for audio caption tasks, audio embeddings are mapped to a sequence of prefix vectors and then taken as the context input for caption generation (Kim et al., 2023; Schaumlöffel et al., 2023; Xu et al., 2024). However, to the best of our knowledge, we are

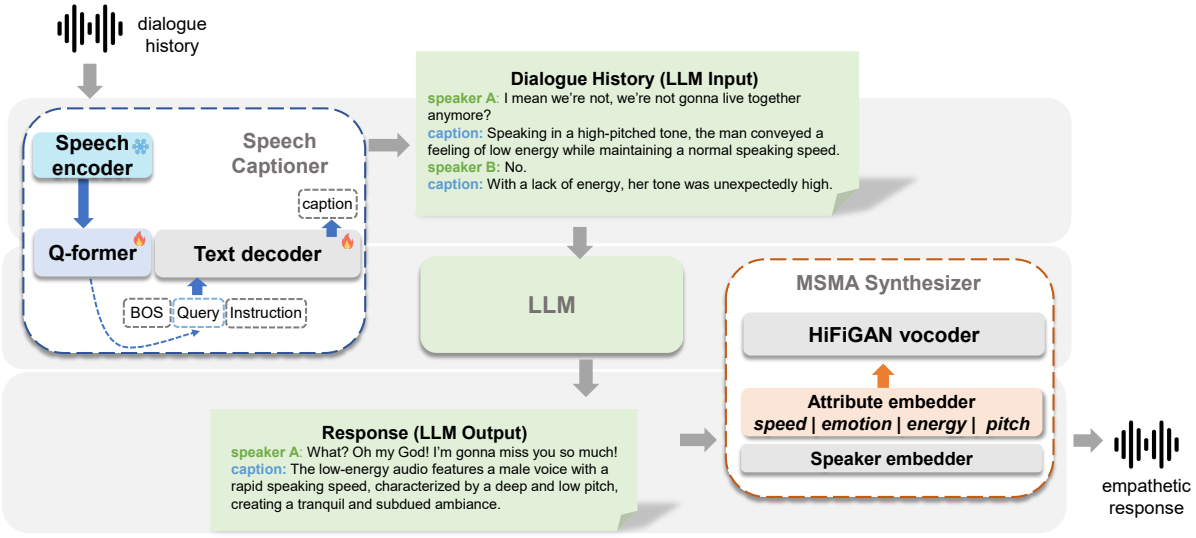

Figure 2: The overall architecture of PerceptiveAgent. Three components are interconnected: the speech captioner, the LLM and the MSMA-Synthesizer. The speech captioner serves as a multi-modal sensory system, perceiving acoustic information from the dialogue history, which is crucial for discerning the speakers' intentions. The LLM acts as the cognitive core, responsible for comprehending the speakers' thoughts and emotions. Conditioned on the response contents and multiple attributes provided by the LLM, the MSMA-Synthesizer generates expressive speech outputs.

the first to construct a speech captioner capable of perceiving acoustic information in dialogues.

2.3 Expressive Text-to-Speech Synthesis

Given a transcript, text-to-speech (TTS) models achieve voice variability by conditioning on a zero-shot speech prompt or a text prompt of the desired style. For instance, zero-shot TTS systems reproduce the speaker characteristics and acoustic environments of a speech prompt through in-context learning (Wu et al., 2022; Wang et al., 2023; Shen et al., 2023; Le et al., 2023) . However, these systems lack independent control over speaking styles, including prosody, emotion, and acoustic environment. To address this, text prompts have been introduced for more natural and general speech synthesis. Approaches like (Guo et al., 2023; Leng et al., 2023; Shimizu et al., 2023; Ji et al., 2023) express speaking styles in natural language, while methods such as (Polyak et al., 2021; Nguyen et al., 2023) utilize explicit labels to generate diverse speech that matches the prompt. We follow the latter direction and construct a speech synthesis model with multiple speaking style labels.

3 Methods

As a multimodal dialog system, PerceptiveAgent is capable of audio modality perception and empathetic speech generation, which is achieved through the incorporation of prosodic information expressed in natural language. To capture prosodic features from speech inputs, we propose a novel speech caption model that aligns audio features with the latent space of a pre-trained language model. To enhance empathy and diversity of the simulated speech communication, a multi-speaker and multi-attribute vocoder is developed. This vocoder synthesizes speech by conditioning on both response contents and captions of speaking styles, resulting in more engaging and realistic dialogues.

3.1 Speech Captioner

The speech caption model is designed to capture prosodic information and transcribe it as textual descriptions. It operates by encoding speech inputs by the speech encoder in ImageBind (Girdhar et al., 2023), followed by description generation by the pre-trained GPT-2 decoder (Radford et al., 2019). To bridge the gap between the speech encoder and the text decoder, we introduce a Querying Transformer (Q-former) pre-trained in BuboGPT (Zhao et al., 2023). This model is connected with a linear projection layer, which is subsequently followed by a text decoder. To effectively fine-tune this model, we integrate the following two fine-tuning strategies, while keeping the speech encoder frozen throughout the training procedure.

3.1.1 Multi-modal Embedding Alignment

Prefix tuning is utilized to align the output of the Q-former with the latent space of the text decoder. A query vector with fixed dimensions is generated by the Q-former. These embeddings interact with each other through self-attention layers and with frozen audio features through cross-attention layers. To bridge the gap with the word embedding space, query embeddings are used as prefix vectors and attended to by the text decoder. This bottleneck architecture serves to compel the queries to extract the acoustic information that is most relevant to the textual descriptions.

3.1.2 Instruction Tuning

To bridge the gap between the next-word prediction objective of the pre-trained decoder and the objective of acquiring multi-modal information conditioned on prefix sequences, instruction tuning is employed to train the speech captioner. We first construct an instructional dataset, where each instance comprises three elements: a query vector, an instruction, and a caption. The instruction is described as a natural language text sequence that specifies the task, serving to constrain the model's outputs to align with desired response characteristics or domain knowledge. This provides a channel for humans to intervene with the model's behaviors. Varied instructions are gathered using GPT-3.5-Turbo in this work. Additionally, the caption represents the desired output following the instruction, while the query vector is derived from acoustic representations. Throughout the training procedure, the parameters of the speech encoder are fixed, while the Q-former and text decoder remain trainable. During each inference process, instructions are randomly selected and incorporated into the generated sequence to enhance diversity and simulate human cognitive processes more effectively, thereby yielding more varied outputs.

3.2 PerceptiveAgent

Figure 2 illustrates the overall framework of PerceptiveAgent, a multi-modal dialogue system comprising three interconnected stages: Intention Discerning by the speech captioner, Comprehension through Sensory Integration by the LLM and Expressive Speech Synthesis by the MSMA-Synthesizer. PerceptiveAgent exhibits two key characteristics: (1) It leverages natural language to perceive and express acoustic information, and (2) It employs an LLM as the cognitive core in the system, to comprehend multi-modal contextual history and deliver audio responses.

3.2.1 Caption for Intention Discerning

In the initial stage, a speech caption model is employed to interpret acoustic information from audio inputs. Each speech within the dialogue history is encoded into latent features by a frozen speech encoder. These features are then compressed into a query vector with fixed dimensions, sharing the same latent space as the word embedding of a text decoder. Conditioned on this query sequence and instruction prompt, a textual caption describing the speaking styles for each speech is deduced by the text decoder.

3.2.2 Comprehension through Sensory Integration

Subsequently, an LLM module acting as the cognitive core is integrated into the system, where GPT-3.5-Turbo is employed. The transcribed textual content for each audio is merged with the previously generated caption before being fed into the LLM. Prompts in Appendix A and B are designed to effectively leverage the LLM's contextual understanding abilities. Upon recognizing speakers' intentions by assimilating both the contextual caption and content, the LLM deduces the relevant dialogue content and generates a caption describing how to articulate the derived content.

3.2.3 Expressive Speech Synthesis

Finally, empathetic audio responses are synthesized by the MSMA-Synthesizer, a Multi-Speaker and Multi-Attribute vocoder that is conditioned on the generated dialogue contents and captions. This vocoder is a modification of (Nguyen et al., 2023) to facilitate fine control over speech expressiveness. In addition to taking discrete speech units, speaker and style (emotion) as inputs, our vocoder introduces multiple prosodic attributes, including pitch, speed and energy. To synthesize each inference, the LLM's outputs of dialogue contents and captions are transformed into discrete units or attribute labels respectively, before being fed into the vocoder. Specifically, a text-to-unit (T2U) model is utilized to convert response contents into acoustic units with a Transformer machine translation structure (Vaswani et al., 2017). Emotional and prosodic labels are recognized from response captions by sentence classifiers, accomplished with GPT-3.5-Turbo in this work, while the speaker label is randomly selected.

The architecture of the vocoder comprises a speaker embedder, an attribute embedder and a HIFIGAN vocoder. The speaker embedder uses look-up tables to embed speaker identities, while a set of controllable attributes including speed, emotion, energy and pitch are embedded by the attribute embedder. To synthesize expressive speech, discrete units are initially embedded and up-sampled through a series of blocks consisting of transposed convolution and a residual block with dilated layers. Prior to duration prediction, this up-sampled sequence is concatenated with the speed embedding. The speaker embedding and style embedding are subsequently concatenated to each frame in the up-sampled sequence, which is transformed to a mel-spectrogram by the HiFiGAN generator.

4 Experiments

4.1 Experimental Setup

Datasets. We train our speech captioner on the TextrolSpeech (Ji et al., 2023) dataset, which consists of 236,220 pairs of captions and the corresponding speech samples. The captions in this dataset describe speaking styles in terms of five factors: gender, emotion, pitch, speed and energy.

For the MSMA-Synthesizer, we reproduce a vocoder proposed in (Nguyen et al., 2023) using the EXPRESSO, LJSpeech (Ito and Johnson, 2017) and VCTK (Yamagishi et al., 2019) datasets. The EXPRESSO dataset is subsequently labeled by the speech captioner and GPT-3.5-Turbo to recognize attributes of pitch, speed and energy for each speech. We then utilize this labeled EXPRESSO dataset and the reproduced vocoder to train the MSMA-Synthesizer. We refer to the reading and conversation sections of EXPRESSO as Exp-R and Exp-I respectively. Additionally, a T2U model is trained on the same datasets with the MSMA-Synthesizer to maintain consistency in unit distribution.

To evaluate the overall performance of our system, we utilize a speech dialogue dataset from MELD (Poria et al., 2019). This dataset provides emotion labels for each sentence, which serve as ground truth labels for both response content and speech evaluation. The speeches in this dataset are recorded in realistic scenes with interruptions and environmental noise. In our evaluation, we only consider conversations with two speakers.

We utilize English datasets throughout the entire training process. As a consequence, PerceptiveAgent currently supports only the English language. However, it is noteworthy that PerceptiveAgent can be readily expanded to accommodate multiple languages. Only the MSMA-Synthesizer module requires modification, as the language-agnostic nature of the speech captioner allows it to generate captions from various languages. Meanwhile, existing methods can recognize semantic contents and translate them into English.

Configurations. We utilize the speech encoder in ImageBind (Girdhar et al., 2023), the pre-trained Q-former in BuboGPT (Zhao et al., 2023), and the pre-trained GPT-2 (Radford et al., 2019) to implement the speech captioner. Finetuning is conducted for 43,000 steps with a batch size of 16. For decoding, we use Top-k sampling with k=10 and set the minimum and maximum sequence lengths to 20 and 50, respectively. We reproduce the vocoder for 400,000 steps with a batch size of 32 and learning rate of 0.0004, and train the MSMA-Synthesizer for 200,000 steps with a batch size of 32 and learning rate of 0.0004. The T2U model is structured as a sequence-to-sequence transformer with 4 encoder layers, 4 decoder layers, and 4 attention heads, with a dropout of 0.1. We utilize HuBERT (Hsu et al., 2021) with 2000 clusters to acquire units as targets[1], provided by the textlesslib toolbox (Kharitonov et al., 2022). Decoding is performed using Top-k sampling with k=10. All experiments are conducted on 4 NVIDIA GeForce RTX 4090 GPUs.

4.2 Evaluation

Speech-GPT3.5. We implement Speech-GPT3.5, a dialogue system focusing solely on linguistic information as a baseline. According to the textual history content recognized from the speech input, this system comprehends dialogue context with GPT-3.5-Turbo. After generating the response content, the audio response is synthesized by an off-the-shelf TTS (text-to-speech) model provided by OpenAI[2].

Metrics. The performance of PerceptiveAgent is evaluated in terms of two fundamental aspects: 1) *cognitive empathy* demonstrates the ability to consider the perspective of speakers, reflected in the content of the response; and 2) *affective empathy* exhibits the ability to emotionally understand and share the speaker's feelings, reflected in the

[1] https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt

[2] https://platform.openai.com/docs/guides/text-to-speech

	BERTScore	Accuracy
Speech-GPT3.5	53.03±10.20	0.74
PerceptiveAgent	**54.36±9.25**	**21.89**
-w/o captions	-	16.53

Table 1: Performance evaluation of PerceptiveAgent. BERTScore (%) measures the quality of cognitive empathy in linguistic contents, while accuracy (%) assesses the quality of affective empathy in acoustic responses.

prosody of the generated audio response. Cognitive and affective empathy are assessed by evaluating the quality of generated textual responses and audio responses, respectively.

To evaluate the quality of dialogue text generation, we employ the BERTScore automatic evaluation metric proposed by Zhang et al. (2020), which computes a similarity score for each token in the candidate sentence with each token in the reference sentence. To evaluate the expressiveness of audio generation, we employ an expressive style classifier proposed by Nguyen et al. (2023) to recognize emotion labels for both generated and true speeches. Classification accuracy is used to measure the performance.

Besides, we evaluate the perception ability of the speech captioner on the validation and test datasets, which are split from the TextrolSpeech dataset. We approach this model as a multi-attribute classification task. Upon generating captions from speeches, the predicted labels for attributes including gender, emotion, pitch, speed and energy are determined by a sentence classifier, GPT-3.5-Turbo, while the true labels are provided in the TextrolSpeech dataset. Weighted metrics including precision, recall and F1-score are used to quantify the disparity between the predicted and true labels.

Moreover, the expressiveness of the speech synthesizer is assessed on the validation and test datasets split from the EXPRESSO dataset. We use the same expressive style classifier employed in affective empathy evaluation, to measure the preservation of emotion in the resynthesized speech. For evaluating the preservation of prosody, we compute the F0 Frame Error (FFE), which measures the percentage of frames with a deviation of more than 20% in pitch value between the input and resynthesized output.

4.3 Result Analysis

4.3.1 PerceptiveAgent

Table 1 presents the overall performance of PerceptiveAgent on cognitive empathy and affective empathy, evaluated on the generated content and audio, respectively. BERTScore measures the semantic similarity between the generated and real response contents, while accuracy assesses the similarity and diversity of emotions between the generated and real speeches. Overall, compared to Speech-GPT3.5, PerceptiveAgent demonstrates a strong ability in generating empathetic responses with a closer alignment to the dialogue context in terms of linguistic content and a higher expressiveness in acoustic information. Specifically, PerceptiveAgent achieves a slightly higher BERTScore than Speech-GPT3.5, primarily because our model can generate content that more accurately captures the speaker's intentions and contains more emotionally intense words. Additionally, PerceptiveAgent notably outperforms Speech-GPT3.5 in terms of accuracy, as the latter doesn't incorporate any emotion prompts during speech generation, thus maintaining a limited variety of prosody. Despite this, the accuracy of PerceptiveAgent still remains at a relatively moderate level. This is because the generated responses, while contextually appropriate, may not entirely align with the real responses in terms of semantics and emotions.

4.3.2 Speech Captioner

Table 2 evaluates the speech captioner's generalization performance on both the validation and test sets. Overall, it is evident that that the model achieves the highest F1-score for gender, followed by pitch and emotion. This underscores the model's proficiency in accurately discerning these attributes from input speech. Besides, both gender and emotion exhibit closely aligned precision and recall metrics, affirming the model's predictive prowess for these attributes. Meanwhile, there exists a notable disparity between precision and recall when predicting energy, indicating variable performance and a tendency towards confident predictions. Conversely, the model's performance in predicting speed is unsatisfactory, which can be attributed to the imbalanced distribution of speed in the training dataset, with over 60% of samples labeled as "neutral".

We also discuss how errors in speech processing are affected by demographics of the speakers. Ta-

Attribute	Validation			Test		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Gender	99.3	97.5	98.4	99.3	98.6	99.0
Emotion	85.8	85.4	85.1	87.3	87.1	86.8
Pitch	85.6	76.8	80.4	79.6	72.1	75.3
Energy	72.4	57.4	63.1	77.7	65.3	69.9
Speed	47.2	36.7	41.3	48.5	41.5	44.7

Table 2: Performance evaluation of the speech captioner. Precision, recall and F1-score (%) are utilized to measure its generalization ability on both the validation and test sets. Predicted labels are obtained through semantic classification on the generated captions, while the true labels are derived from the TextroSpeech dataset.

Attribute	Male			Female		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Emotion	84.3	85.4	84.2	87.4	85.5	86.0
Pitch	88.2	82.8	85.3	84.8	71.0	75.9
Energy	74.4	60.0	65.0	71.2	54.9	60.9
Speed	46.4	43.1	44.6	48.0	30.6	37.3

Table 3: Comparison of the speech captioner's performance across genders.

Method	Accuracy		FFE
	Exp-R	Exp-I	Exp
GT	91.9	75.1	-
EXPRESSO	**87.9**	67.0	**0.17±0.12**
MSMA	83.8	**70.8**	0.39±0.16

Table 4: Preservation evaluation of MSMA-Synthesizer. Accuracy (%) is evaluated on EXPRESSO read (Exp-R) and conversation (Exp-I) dataset. F0 Frame Error (FFE) is calculated on EXPRESSO (Exp). GT represents the results of automatic metrics calculated on real audio. EXPRESSO and MSMA refer to the synthesizers in EXPRESSO and PerceptiveAgent respectively.

ble 3 compares the performance of the speech captioner across genders, which represents the most prevalent factor. The F1-score on male speech surpasses that on female speech in terms of pitch, energy and speed, despite the comparable sample sizes for male and female groups (8634 VS. 8983). This demonstrates a variation in the model's performance depending on the gender of the speakers.

4.3.3 MSMA-Synthesizer

Table 4 assesses the MSMA-Synthesizer's ability to preserve emotion and prosody features on the test set, where the EXPRESSO Synthesizer is reproduced by us. The "GT" method represents the results of automatic metrics calculated on real audio. Clearly, the MSMA-Synthesizer achieves higher accuracy on the read dataset compared to EXPRESSO. This suggests that an integration of multiple attributes into speech synthesis can more effectively enable the model to synthesize emotionally expressive audio in dialogue scenarios, meeting the requirements of our system. However, there is a decrease in accuracy on the Exp-R dataset, which is relevant to the less apparent variation in prosody with emotional transitions. Additionally, in terms of FFE, it can be observed that incorporating multiple attributes into the MSMA-Synthesizer may lead to some degradation in speech synthesis quality. However this degradation remains within an acceptable range.

4.3.4 Ablation Study

Effectiveness of Captions. The last line in Table 1 demonstrates the effectiveness of captions in PerceptiveAgent. The system without captions synthesizes speech using randomly selected labels for all four speaking attributes (pitch, speed, energy, and emotion), while maintaining the same response contents as the PerceptiveAgent. It is evident that the PerceptiveAgent outperforms the system without captions, highlighting the effectiveness of captions in generating speech with affective empathy.

Effectiveness of Style Factors. To discern the discrete impact of distinct speaking style factors, we conduct an ablation experiment by systematically

Method	Accuracy		FFE
	Exp-R	Exp-I	Exp
GT	91.9	75.1	-
EXPRESSO	**87.9**	67.0	**0.17±0.12**
MSMA	83.8	**70.8**	0.39±0.16
-style	82.2	69.0	0.40±0.16
-speed	31.8	9.2	0.44±0.13
-energy	31.0	9.1	0.44±0.13
-gender	30.8	8.7	0.44±0.13
-pitch	30.7	7.4	0.43±0.13

Table 5: Performance of the MSMA-Synthesizer conditioned on single speaking style factors.

varying each factor while maintaining the others at their default values. Table 5 presents that the model with style remained achieves the highest accuracy and the lowest FFE, while the models with the other factors exhibit similar performance. This underscores the predominant contribution of style to the effectiveness of expressive speech synthesis.

5 Case Study

Figure 3 presents two cases comparing the response quality between Speech-GPT3.5 and PerceptiveAgent. It demonstrates that by explicitly incorporating acoustic information through captions, the LLM can more accurately comprehend the speaker's intentions and generate more accurate and contextually appropriate responses. The first and second examples illustrate scenarios where the speaker's intention either contradicts or aligns with the linguistic contents, respectively.

The first example in Figure 3 (a) depicts an unplanned meeting conversation between two friends. Analyzing solely from the textual contents, it is suggested that the speaker B is extremely excited and delighted about this conversation. However, a closer examination of the key words of "lower vocal" and "subbed energy" in speaker B's caption reveals an evasive attitude towards the situation. Consequently, when confronted with speaker A's question, "Were you here waiting for me?", it can be inferred that speaker B is not inclined to engage in extensive conversation. The absence of nuanced captions poses a challenge for Speech-GPT3.5, leading to a misinterpretation and generating a response that implies a strong desire to continue the conversation. In contrast, PerceptiveAgent provides a response in accordance with the underlying meaning. Therefore, despite potential inconsistencies between linguistic contents and speaker intentions disrupting the accuracy of dialogue context understanding, PerceptiveAgent, with the aid of captions, can effectively capture the speaker's intent by correctly discerning the acoustic information of speech.

In the second example in Figure 3 (b), the speaker A receives a paper from his mother and intends to share it with his friends. It can be inferred that he is highly excited at the moment, as evidenced by the key words "treble tone" and "energetically" in the caption. Recognizing speaker A's excited mood, the response generated by PerceptiveAgent mirrors the same enthusiasm and curiosity, aligning well with the ground truth. However, Speech-GPT3.5 fails to perceive speaker A's excitement and merely raises the question in a bland manner. Thus, in scenarios where the textual contents coincides with the speaker's intent, our model can also provide responses that correspond to the context of the conversation.

6 Conclusion

In this paper, we propose PerceptiveAgent, an empathetic multi-modal dialogue system capable of accurately discerning the speaker's intentions through the integration of perceptive speech captions and to respond with nuanced and expressive spoken dialogues. Specifically, PerceptiveAgent comprises three cascaded modules: a speech captioner for intention discernment, an LLM for comprehension through sensory integration, and an MSMA-Synthesizer for expressive speech synthesis. Initially, the system employs a perceptive captioner model to capture acoustic features from each speech within dialogues. Subsequently, an LLM module serves as the cognitive core, generating relevant response content with a caption conditioned on the comprehension of the speaker's intentions. An MSMA-Synthesizer is then developed to synthesize expressive speech. Experimental results indicate PerceptiveAgent's strong ability in empathetic response generation, closely aligning with the dialogue context in terms of linguistic contents and exhibiting high expressiveness in acoustic information. Additionally, a case study demonstrates PerceptiveAgent's capability to accurately identify the speaker's intentions in scenarios where the literal interpretations of words are either contrary to or inconsistent with the speaker's true feelings.

Input

Speaker A: Are you being British?!
Caption: In a hushed voice, she speaks rapidly. Speaking softly, she maintains a high-pitched voice.
Speaker B: No. Not anymore.
Caption: Employing a lower vocal quality and customary speaking speed, his expression radiates subdued energy.
Speaker A: Why are you shopping here? You don't live in this neighborhood. Were you here waiting for me?
Caption: Her subdued energy doesn't affect the high pitch of her voice at a regular speed.

Response

Ground Truth
Speaker B: Yeah, huh. I'm just uh, you know I'm just picking up some things for a party.
Caption: The boy's low energy is evident in his fast, low-pitched speech.
Speech-GPT3.5
Speaker B: Yes, I was actually. I needed to talk to you about something important.
PerceptiveAgent
Speaker B: I needed to pick up a few things, and I happened to be in the area. That's all.
Caption: Speaking with a low energy level, his slow speaking pace conveys a sense of calmness and control.

(a) Contradictory Example

Input

Speaker A: Hey guys check it out! My mom sent me the paper!
Caption: With a treble tone, she converses naturally and energetically. Her speaking rate is normal, despite her high-pitched voice.

Response

Ground Truth
Speaker B: Ooh, let's see it!
Caption: Speaking with a treble, she maintains a regular conversational pace. Speaking with normal energy, she maintains a treble tone.
Speech-GPT3.5
Speaker B: What does it say?
PerceptiveAgent
Speaker B: That's great! Let's take a look at it together.
Caption: A speaker with a low tone and normal speed, who is energetically speaking to others and taking interest in the conversation.

(b) Consistent Example

Figure 3: Cases comparing the response quality between Speech-GPT3.5 and PerceptiveAgent.

7 Limitations

Although PerceptiveAgent excels at providing empathetic responses in terms of both linguistic and acoustic contents, several limitations can be observed in this system: 1) **Dataset Limitation**: PerceptiveAgent's perception ability is currently constrained by the comprehensiveness of the training dataset in describing speech information. Presently, it is unable to discern speaker identity and background noise from speech; 2) **Time Delay Limitation**: PerceptiveAgent is a system cascaded by three interconnected components, which introduces accumulated delays to the response time, and 3) **Length Limitation**: The maximum token length in LLMs may limit the multi-turn dialogue.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 62276245).

References

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. X-LLM: bootstrapping advanced large language models by treating multi-modalities as foreign languages. *CoRR*, abs/2305.04160.

Jian Cong, Shan Yang, Na Hu, Guangzhi Li, Lei Xie, and Dan Su. 2021. Controllable context-aware conversational speech synthesis. In *Interspeech*, pages 4658–4662. ISCA.

Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: A review of the concept. *Emotion review*, 8(2):144–153.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190. IEEE.

Haohan Guo, Shaofei Zhang, Frank K. Soong, Lei He, and Lei Xie. 2021. Conversational end-to-end TTS for voice agents. In *IEEE Spoken Language Technology Workshop*, pages 403–409. IEEE.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23802–23804. AAAI Press.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2023. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. *CoRR*, abs/2308.14430.

Eugene Kharitonov, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Paden Tomasello, Ann Lee, Ali Elkahky, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2022. textless-lib: a library for textless spoken language processing. *CoRR*, abs/2202.07359.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240. Association for Computational Linguistics.

Minkyu Kim, Kim Sung-Bin, and Tae-Hyun Oh. 2023. Prefix tuning for automated audio captioning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*.

Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiang-Yang Li, Sheng Zhao, Tao Qin, and Jiang Bian. 2023. Prompttts 2: Describing and generating voices with text prompt. *CoRR*, abs/2309.02285.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Chaohu Liu, Kun Yin, Haoyu Cao, Xinghua Jiang, Xin Li, Yinsong Liu, Deqiang Jiang, Xing Sun, and Linli Xu. 2024. HRVDA: high-resolution visual document assistant. *CoRR*, abs/2404.06918.

Kentaro Mitsui, Yukiya Hono, and Kei Sawada. 2023. Towards human-like spoken dialogue generation between AI agents from written dialogue. *CoRR*, abs/2310.01088.

Michael Negnevitsky. 2005. *Artificial intelligence: a guide to intelligent systems*. Pearson education.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarandi, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. 2023. EXPRESSO: A benchmark and analysis of discrete expressive speech resynthesis. *CoRR*, abs/2308.05725.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2022. Generative spoken dialogue language modeling. *CoRR*, abs/2203.16502.

Yuto Nishimura, Yuki Saito, Shinnosuke Takamichi, Kentaro Tachibana, and Hiroshi Saruwatari. 2022. Acoustic modeling for end-to-end empathetic dialogue speech synthesis using linguistic and prosodic

contexts of dialogue history. In *Interspeech*, pages 3373–3377. ISCA.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. In *Interspeech*, pages 3615–3619. ISCA.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 527–536. Association for Computational Linguistics.

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

Harry T Reis, Michael R Maniaci, Peter A Caprariello, Paul W Eastwick, and Eli J Finkel. 2011. Familiarity does indeed promote attraction in live interaction. *Journal of personality and social psychology*, 101(3):557.

Stuart J Russell and Peter Norvig. 2010. *Artificial intelligence a modern approach*. London.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11229–11237. AAAI Press.

Yuki Saito, Shinnosuke Takamichi, Eiji Iimori, Kentaro Tachibana, and Hiroshi Saruwatari. 2023. Chatgpt-edss: Empathetic dialogue speech synthesis trained from chatgpt-derived context word embeddings. *CoRR*, abs/2305.13724.

Timothy Schaumlöffel, Martina G Vilas, and Gemma Roig. 2023. Peacs: Prefix encoding for auditory caption synthesis. In *Proceedings of the Detection and Classification of Acoustic. Scenes Events Challenge*, pages 1–3.

Murray Shanahan. 2024. Talking about large language models. *Commun. ACM*, 67(2):68–79.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *CoRR*, abs/2304.09116.

Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana. 2023. Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions. *CoRR*, abs/2309.08140.

Adam Smith. 2006. Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record*, 56(1):3–21.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*, abs/2211.09085.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *CoRR*, abs/2301.02111.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal LLM. *CoRR*, abs/2309.05519.

Yihan Wu, Xu Tan, Bohan Li, Lei He, Sheng Zhao, Ruihua Song, Tao Qin, and Tie-Yan Liu. 2022. Adaspeech 4: Adaptive text to speech in zero-shot scenarios. In *Interspeech*, pages 2568–2572. ISCA.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi

Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19323–19331. AAAI Press.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). https://doi.org/10.7488/ds/2645.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15757–15773. Association for Computational Linguistics.

Fang Zhang, Yongxin Zhu, Xiangxiang Wang, Huang Chen, Xing Sun, and Linli Xu. 2024. Visual hallucination elevates speech recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19542–19550. AAAI Press.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*. OpenReview.net.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *CoRR*, abs/2307.08581.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Prompt for Dialogue Generation with Captions

You are the last speaker in the following daily dialogue.
Each speaker is provided with a speaking style caption after the conversation, which contains gender, speaking speed, pitch, energy and emotion. You MUST give a response FIRST depending on the dialogue history, followed by a speaking style caption.
NOTE: It is important to recognize the speaker's intention from the speaking style before generating response. You MUST keep response as short as possible.

Here is an example:

Input:
Speaker A: My specimen is deposited into the container in the room. Janice! You're not... gone?
Speaking in a high pitch, the male speaker conveyed a touch of low energy during normal-paced conversation.

Speaker B: Oh! Sid is still in his room. So did you do it? Did you make your deposit?
Speaking swiftly, her voice remains soft and steady. With little energy, her voice reaches a high pitch.

Speaker A: Yeah! yeah... The hard part is over!
The man's words are barely audible, but he keeps up the standard pace.

Speaker B:

Output:
That's not the hard part honey! The hard part is what comes next, I mean aren't you worried about the results?
A woman speaks quickly and her high-pitched tone is a trademark of her dazed.

Input:
{dialogue history with caption}

Output:
{response content}
{response caption}

B Prompt for Dialogue Generation without Captions

You are the last speaker in the following daily dialogue.
You MUST give a response depending on the dialogue history. You MUST keep response as short as possible.

Here is an example:

Input:
Speaker A: My specimen is deposited into the container in the room. Janice! You're not... gone?

Speaker B: Oh! Sid is still in his room. So did you do it? Did you make your deposit?

Speaker A: Yeah! yeah... The hard part is over!

Speaker B:

Output:
That's not the hard part honey! The hard part is what comes next, I mean aren't you worried about the results?

Input:
{dialogue history without caption}

Output:
{response content}