arXiv:2402.12786v2 [cs.CL] 30 May 2024

Advancing Large Language Models to Capture Varied Speaking Styles and Respond Properly in Spoken Conversations

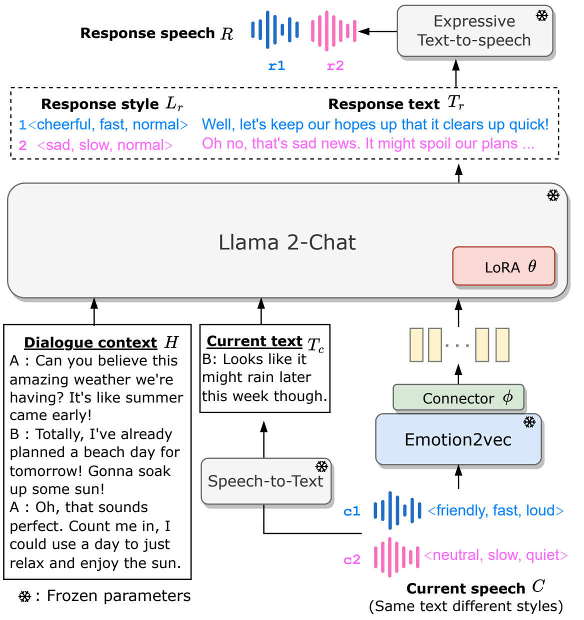
Guan-Ting Lin, Cheng-Han Chiang, Hung-yi Lee
National Taiwan University,
Taiwan
{f10942104, hungyilee}@ntu.edu.tw, dcml0714@gmail.com

Abstract

In spoken dialogue, even if two current turns are the same sentence, their responses might still differ when they are spoken in different styles. The spoken styles, containing paralinguistic and prosodic information, mark the most significant difference between text and speech modality. When using text-only LLMs to model spoken dialogue, text-only LLMs cannot give different responses based on the speaking style of the current turn. In this paper, we focus on enabling LLMs to *listen to* the speaking styles and respond properly. Our goal is to teach the LLM that "*even if the sentences are identical if they are spoken in different styles, their corresponding responses might be different*". Since there is no suitable dataset for achieving this goal, we collect a speech-to-speech dataset, **StyleTalk**, with the following desired characteristics: when two current speeches have the same content but are spoken in different styles, their responses will be different. To teach LLMs to understand and respond properly to the speaking styles, we propose the **Spoken-LLM** framework that can model the linguistic content and the speaking styles. We train Spoken-LLM using the StyleTalk dataset and devise a two-stage training pipeline to help the Spoken-LLM better learn the speaking styles. Based on extensive experiments, we show that Spoken-LLM outperforms text-only baselines and prior speech LLMs methods. [1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in dialogue generation, natural language understanding, and commonsense reasoning (Wei et al., 2022; OpenAI, 2023). While LLMs mostly focus on text modality, speech represents the most natural form of human communication in our daily lives. In this work, we aim to inject speech modality for modeling *spoken conversation* with Multi-modal LLMs (MM-LLMs). The main goal is to develop a humanizing agent capable of listening, understanding, and engaging in dialogue with humans, ultimately leading to higher user satisfaction.



1 <cheerful, fast, normal> Well, let's keep our hopes up that it clears up quick!
2 <sad, slow, normal> Oh no, that's sad news. It might spoil our plans ...

A : Can you believe this amazing weather we're having? It's like summer came early!
B : Totally, I've already planned a beach day for tomorrow! Gonna soak up some sun!
A : Oh, that sounds perfect. Count me in, I could use a day to just relax and enjoy the sun.

B: Looks like it might rain later this week though.

Figure 1: The overview framework of Spoken-LLM. (c1,r1) and (c2,r2) are the current and response speech sample pairs. c1 and c2 are fed into the model individually.

Speech signals contain linguistic aspects (words, phonetics, syntax, and semantics), paralinguistic elements (emotions and speaker characteristics), and prosodic factors (speaking style, emphasis, and attitude). In human conversation, while the dialogue primarily relies on the lexical aspect, the speaking styles convey rich information beyond text, and can even alter the semantics of the spoken sentences (Castro et al., 2019). Neglecting spoken styles can lead to misinterpretation of communica-

[1] Demo of the StyleTalk dataset and output of Spoken-LLM are at https://sites.google.com/view/spoken-llm/home. Code and dataset are available at https://github.com/DanielLin94144/StyleTalk.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6626–6642
August 11-16, 2024 ©2024 Association for Computational Linguistics

tion or unnatural human interaction. For example, as shown in Figure 1, the current speech with the same current text (*Looks like it might rain later this week though.*) but different speaking styles. The friendly speaking style leads to a cheerful response while speaking in a slow and neutral tone leans toward a sad and negative response.

Although there are recent studies on MM-LLMs for speech/audio and text, most of the existing studies focus on *content-centric* Spoken Language Modeling (SLM) (Lakhotia et al., 2021; Kharitonov et al., 2022), joint text and speech processing tasks (Rubenstein et al., 2023; Chou et al., 2023; Maiti et al., 2023; Nachmani et al., 2023; Zhang et al., 2023) or general audio perception and hearing ability (Tang et al., 2023; Gong et al., 2023a; Deshmukh et al., 2023). There is less attention on spoken dialogue with advanced methods and suitable datasets for modeling paralinguistics and speaking styles of spoken responses.

To model spoken dialogue with a generative language model, dGSLM (Nguyen et al., 2023) proposes a dual-tower SLM on discrete speech units to model two-channel spoken dialogue, but the generated spoken sentences lack semantic meaning. ParalinGPT (Lin et al., 2023b) organizes tasks in the sequence of current paralinguistic attribute prediction, response paralinguistic attribute prediction, and response text generation with autoregressive conditioning. However, it only uses the speech sentiment as speaking style, which might be primarily based on textual information, and how the speaking styles affect the spoken response is unclear. A concurrent work E-chat (Xue et al., 2023) enhances LLM to generate responses in different emotional contexts, but the training and evaluation data are entirely generated by GPT-3.5 without human supervision, equivalent to distillation and prompting of GPT-3.5. It can only generate response text, constraining its capacity to control response style or speech-to-speech modeling.

To overcome the current limitation, we collect a novel speech-to-speech conversational dataset named **StyleTalk**. This dataset is the first spoken conversation benchmark with *the same dialogue context and input sentence in different speaking styles, accompanied by corresponding expressive spoken responses* for speech-to-speech modeling. The dataset will be released upon the paper's acceptance.

Based upon the StyleTalk dataset, we propose a multi-modal two-stage training method named **Spoken-LLM** for spoken dialogue modeling. Spoken-LLM is a fusion of the widely-used open-sourced LLM (Llama 2-Chat (Touvron et al., 2023)) and a self-supervised speech emotion representation model (emotion2vec (Ma et al., 2023)). The proposed model can predict response speaking style and text, enabling the subsequent expressive Text-to-Speech (TTS) model to generate natural and diverse speech responses. We validate the performance through objective and subjective evaluations of spoken responses. With the same backbone model, the proposed method outperforms the text and speech LLM baseline in lexical/semantic similarity and response style F1 score. The human evaluation also indicates that the proposed method yields more reasonable and proper response speech than the text-only LLM baseline approach.

Dataset	S2S	Expressive	Purpose	Diff styles&resp
IEMOCAP	✓	✓	recognition	✗
Switchboard	✓	✓	recognition	✗
MUStARD	✓	✓	recognition	✗
SEMAINE	✓	✓	recognition	✗
MELD	✓	✓	recognition	✗
MEISD	✓	✓	recognition	✗
MSP-improv	✓	✓	recognition	✗
SCQA	✗	✗	question answering	✗
NMSQA	✓	✗	question answering	✗
OpenSAQA*	✗	✓	question answering	✗
E-chat200*	✗	✓	dialogue generation	✗
StyleTalk	✓	✓	dialogue generation	✓

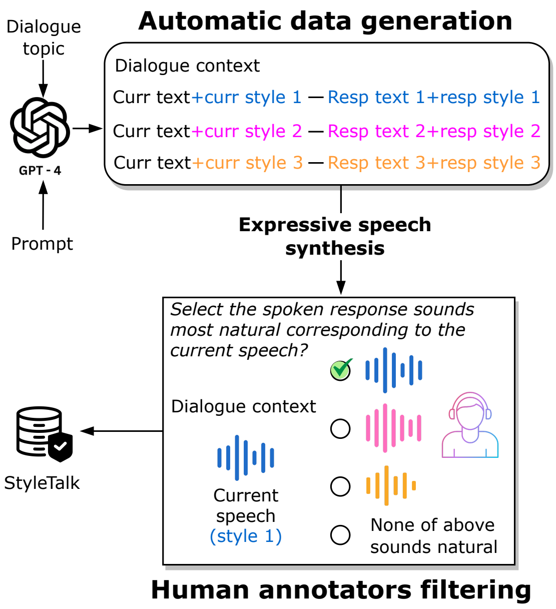
Table 1: The list of spoken conversation datasets. “S2S" means speech-to-speech, and “Diff styles&resp" stands for *the same sentence in different speaking styles and responses*. In the “Purpose" column, “recognition" refers to recognizing the speaking style attributes in the speech, “question answering" means the task is formulated as the (question, answer) pair, and “dialogue generation" is the general chatbot agent to response any kinds of input. The datasets noted with * are purely generated by LLM.

2 Dataset: StyleTalk

2.1 Overview

StyleTalk is a speech-to-speech conversation dataset. Each sample in the dataset comprises dialogue context (in text), current turn in speech (annotated with speaking style), and the response turn in speech (annotated with speaking style) (illustrated in Figure 1).

StyleTalk features the following characteristics: Multiple samples in StyleTalk share the same dialogue context, the text of the current input turn, but they have different responses speech since the speaking style of the current turn is different. To



Dialogue topic
Automatic data generation
Dialogue context
Curr text+curr style 1 — Resp text 1+resp style 1
Curr text+curr style 2 — Resp text 2+resp style 2
Curr text+curr style 3 — Resp text 3+resp style 3
GPT-4
Prompt
Expressive speech synthesis
Select the spoken response sounds most natural corresponding to the current speech?
Dialogue context
Current speech (style 1)
None of above sounds natural
StyleTalk
Human annotators filtering

Figure 2: Data collection pipeline of StyleTalk. The details of instruction and prompt template are in the Appendix.

the best of our knowledge, no existing corpora focus on such a characteristic.

By training on this dataset, we hope the LLM can learn to use the dialogue context and current turn, specifically, the speaking style, to predict the next turn. Given that speaking styles convey additional information beyond text, incorporating style modeling helps to disambiguate human intent and facilitates dialogue engagement.

2.2 Data Collection

In this section, we introduce how the dataset is collected. The data collection pipeline includes three stages: (1) using LLM for data generation text dialogue with style annotation, (2) using an expressive TTS model to synthesize speech from text dialogue, and (3) recruiting human annotators to filter the dataset. We illustrate the data collection pipeline in Figure 2.

2.2.1 LLM for Data Generation

Crafting a scenario with the same context and words but expressed in different speaking styles is a non-trivial task. Most dialogue corpora typically consist of one style, making it challenging to study the impact of various speaking styles on spoken responses.

Recently, LLMs have demonstrated human-level knowledge and powerful data generation capabilities when provided with well-designed prompts and instructions. In light of this, we propose leveraging GPT-4 (OpenAI, 2023) to generate spoken **dialogue set** consisting of a dialogue context, the same sentence presented in three different speaking styles, and three corresponding responses. To let LLM understand the speaking style information, the speaking style is represented in text by surrounded by special marker, for example, <*emotion, speed, volume*>. To increase the diversity of the dialogue, we prompt the GPT-4 with 17 common daily dialogue topics: school, work, family, health, entertainment, travel, food, sports, finance, technology, music, movies, books, games, beauty, shopping, and weather. Additionally, we use decoding with temperature sampling to ensure diversity in the dataset. The prompt template is shown in the appendix 6.

2.2.2 Expressive Speech Synthesis

To generate high-quality speech with style and prosody control, we utilize an industrial-grade Microsoft Azure Text-to-Speech (TTS) system[2]. For the speaking style, we employ **emotion** (neutral, cheerful, sad, friendly, unfriendly), **speeds** (slow, medium, fast), and **volumes** (quiet, medium, loud) for prosodic control. There are nine speakers, with four male and five female speakers.

2.2.3 Human Annotator Filtering

While LLMs can effectively follow instructions and generate reasonably coherent dialogue samples, LLMs are trained on textual data and lack exposure to human-human spoken dialogue. Additionally, the expressive TTS system may not achieve perfect naturalness and style-following in synthesizing speech under style conditions. The automatically generated data may exhibit unnatural characteristics for human speakers. Therefore, additional examination is necessary to check the quality of the speech data and the overall naturalness of spoken dialogue sample pairs.

To ensure data quality, we request human listeners to participate in a listening test conducted on the Amazon Mechanical Turk platform. An illustration of the listening test is provided in Figure 2. In this evaluation, participants are presented with a dialogue context text, the current spoken turn and three response spoken turns. They are then instructed to choose the most suitable response among the three options. Alternatively, if they perceive all three

[2] https://azure.microsoft.com/en-us/products/ai-services/text-to-speech

responses as unnatural, they can select the option "None of the above is natural." Participants need to be aware of the style of the current turn and *differentiate between the three response turns to identify the most natural one*. Through this evaluation, we aim to filter out sample pairs that are deemed unnatural or indistinguishable. Details are shown in Appendix A. We found out that only around 33% samples successfully passed the human filtering process. This suggests that LLM-generated spoken samples are either *not natural to human perception* or *the speaking style does not distinctly influence spoken responses*.

2.3 Data split

After manual filtering, we split the filtered data into training and evaluation sets with dialogue sets. “Sample" means a current and response speech pair. The detailed data statistics are shown in Appendix Table 7.

Train set: The training set is carefully curated through manual filtering, resulting in 1,878 dialogue sets and 1,986 samples.

Evaluation set: The evaluation set contains 486 dialogue sets and 981 samples, most of the dialogue sets have two to three speaking styles for the current text.

In addition to the train and evaluation set, a fully LLM-generated **unfiltered set** is introduced for data augmentation since the size of the training set is limited. The unfiltered set consists of 5,777 dialogue sets and 16,472 samples. It is crucial to note that this data is not subject to human supervision, and as such, the samples may not align perfectly with human standards.

3 Spoken-LLM framework

3.1 Overview

The framework of Spoken-LLM is illustrated in Figure 1. The main components include the large language models, speaking style encoder, speech-to-text conversion, and expressive TTS system. D_s and D_t denote the dimension of the speech encoder's output and LLM's input space, respectively.

We formulate the task as follows: given a multi-turn spoken dialog with dialogue context H in text T_h, a current turn C comprising speech S_c and text T_c. The prediction response speech R includes response style L_r and response text T_r. Note that we use the ground truth transcripts T_c of the current turn, since addressing speech recognition errors is not the focus of this work. The discussion of using ASR prediction is in section C.

3.2 Large Language Model

This study adopts the open-sourced Llama 2-Chat 7B model, derived from the fine-tuned version of Llama 2 (Touvron et al., 2023), exhibiting optimized dialogue generation capabilities. Throughout the training process, the Llama 2-Chat model remains frozen, and we introduce the trainable LoRA adapter (Hu et al., 2021) for parameter-efficient fine-tuning.

3.3 Speech Style Encoder

Among the self-supervised speech models (Yang et al., 2021; Lin et al., 2023a), emotion2vec (Ma et al., 2023) achieves state-of-the-art performance on diverse paralinguistics-related tasks. Precisely, it extends the data2vec 2.0 (Baevski et al., 2023) with both utterance-level and frame-level loss using emotional speech data, and extra chunk token embeddings are used to capture utterance-wise information.

We choose emotion2vec as the speech encoder to extract universal paralinguistic and prosody embeddings. Two approaches are used for feature extraction. (1) *Utterance-level averaging embedding (utt)*: which involves a simple averaging of frame-wise representations to create an utterance-level embedding. The embedding is in $1 \times D_s$ dimension. (2) *Chunk embedding*: emotion2vec learns 10 extra chunk token embeddings to capture both fine-grained and global speech information. Chunk embeddings are in $10 \times D_s$ dimension.

A lightweight *Connector* module with layer normalization and a linear model is utilized to project the speech embeddings into the dimension of the language model's input space (from D_s to D_t). Only the parameters of the connector are updated, while the emotion2vec model remains frozen. The number of trainable parameters for utterance and chunk embeddings are the same.

3.4 Spoken Dialogue Modeling

1st-stage: style alignment: The first-stage training is used to align the speech embedding with LLM input space. To achieve this, the frozen LLM has to predict the current input style. Only the connector ϕ is trained. The training objective is to minimize the cross-entropy loss for classifying L_c:

$$\mathcal{P}(L_c|C, I_1; \phi), \tag{1}$$

where $C = \{T_c, S_c\}$. I_1 is the task instruction shown in Appendix F. Since this training stage requires a reasonable amount of data to have better alignment, we use the current speech from the unfiltered set for training.

2nd-stage: spoken response modeling: After the LLM can understand the speech embedding, the LLM is optimized to predict the response style and response text by training the LoRA adapter θ and speech connector ϕ. The second-stage training objective is the causal language modeling cross-entropy loss to predict the response style L_r then response text T_r:

$$\mathcal{P}(R|H, C, I_2; \theta, \phi) = \mathcal{P}(L_r|H, C, I_2; \theta, \phi) \mathcal{P}(T_r|H, C, L_r, I_2; \theta, \phi) \quad (2)$$

where $H = T_h$ and I_2 is the task instruction shown in Appendix F. The speaking style label L_r is integrated into the text through special bracket markers with the format <*emotion, speed, volume*>. T_h and T_c are fed into LLM subword embedding, and S_c is passed through the speech encoder plus the connector. We concatenate the resulting continuous embeddings as the input prompt for LLM.

Warmup pre-training: Given the limited size of the human-annotated training set, we propose leveraging the unfiltered set for model warmup pre-training. This allows the model to grasp general knowledge and understand the structure of the dialogue modeling task. Subsequently, we fine-tune the model on the training set to align with human perception, utilizing a smaller learning rate for stable training. This warmup training strategy is designed to mitigate overfitting on the small training data while maintaining good performance.

3.5 Inference

Once the model has completed training, when presented with a dialogue context and current speech input, the initial step involves converting the speech into text through either ground truth text in the oracle setup or an Automatic Speech Recognition (ASR) model in the ASR setup. Then, the Spoken-LLM generates response style and response text sequentially. The representation of the response style is surrounded by special bracket tokens, designed to enable the decoding of both the response style and text. Leveraging the capability of an expressive TTS model to control the response speaking style, we can synthesize the generated response back into speech. This synthesis takes into account the identified response style, and the generated response text, resulting in a synthesized speech output that is not only coherent but also aligns with the desired style and content.

4 Experiments

4.1 Baseline method

All baseline methods are fine-tuned on the same amount of training data and warmup pre-training, with the identical LLM backbone and speech style encoder for ParalinGPT.

Text-LLM (text-only): The initial simple baseline is built by simply fine-tuning text-to-text LLM on StyleTalk. This serves as a performance reference to evaluate the model's capability without knowing any explicit speaking style information. Since the model cannot predict the response style, A randomly selected response style is assigned for this method to synthesize expressive speech.

Text-LLM (cascaded): One can represent the style information in text to enable the model to better predict the response style and text. This approach, referred to as the cascaded pipeline method, involves cascading a style recognition model[3] with the text LLM. The Text-LLM (upper bound) method is the cascaded text-LLM with ground truth style labels.

ParalinGPT (Lin et al., 2023b): The *serialized multitasking* approach proposed by ParalinGPT is a sequential conditioning mechanism, unifying current style prediction, response style prediction, and response text generation within an autoregressive chain. The main difference between ParalinGPT and Spoken-LLM is that Spoken-LLM performs two-stage training (style alignment for current speech then focus on the response speech), but ParalinGPT directly models them in an autoregressive chain, which might be prone to error propagation if the incorrect current style prediction or focusing too much on the current style.

4.2 Evaluation Metrics

Objective evaluation: For automatic evaluation of response text, we adopt the widely-used text generation metric, including lexical-level score (BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005)), and

[3] We use the Spoken-LLM-*chunk* 1st-stage model as the style recognition model, which achieves 86.8, 99.2, 64.0 f1 scores on current emotion, speed, volume prediction, respectively.

Method	Response text				Response style		
	BLEU	**ROUGE$_l$**	**METEOR**	**BERT$_{f1}$**	**F1$_{emotion}$**	**F1$_{speed}$**	**F1$_{volume}$**
Text-LLM (text-only)	3.1	16.2	17.4	75.3	17.5	37.1	41.9
Text-LLM (cascaded)	3.2	17.3	19.1	76.0	37.5	52.9	65.6
Text-LLM (upper bound)	4.0	17.9	19.6	76.3	40.2	53.5	65.8
ParalinGPT-*utt*	3.1	16.8	18.5	75.9	32.3	51.9	64.8
ParalinGPT-*chunk*	3.1	16.5	18.2	75.8	34.0	54.8	**65.8**
Spoken-LLM-*utt*	2.8	16.6	**20.2**	75.8	47.4	61.5	56.5
Spoken-LLM-*chunk*	**4.0**	**17.8**	19.4	**76.3**	**49.6**	**62.1**	61.1

Table 2: Main results comparing text-LLM, ParalinGPT, and Spoken-LLM. *utt* and *chunk* refer to utterance-wise and chunk-wise speech embedding from emotion2vec. The Text-LLM (upper bound) method is the cascaded text-LLM with ground truth style labels.

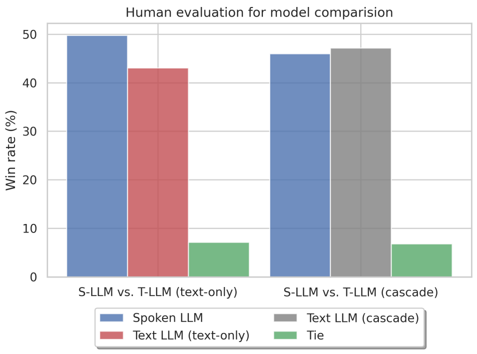


Figure 3: Human evaluation result comparing Spoken-LLM-chunk with Text-LLM (text-only) and Text-LLM (cascaded).

semantic-level (BERT Score (Zhang et al., 2019))[4]. For response style evaluation, since the style attributes are categorical, we calculate the Weighted F1 score for speaking emotion, speed, and volume.

Subjective evaluation: We perform the human evaluation on a set of 200 samples using an A/B test for model comparison. Three human evaluators are assigned to each sample, and they are instructed to rate the model based on both the generated text and speech. The details of subjection evaluation are in Appendix H.

4.3 Main Results

Spoken-LLM outperforms speech and text baselines: Table 2 shows the result on objective evaluation. Firstly, for the text-LLM baseline on response text metrics, we observe that adding current speech style information yields significantly better performance than the text-only method, indicating that recognizing the style information is beneficial to predict textual response. Next, we compare the Text-LLM and speech ParlinGPT baseline. ParalinGPT consistently outperforms the Text-LLM method on the response text metrics. However, on the response style, the text-LLM (cascaded) is better than ParalinGPT-*utt*. In contrast, our proposed Spoken-LLM methods perform slightly better than ParalinGPT on response text, with significantly superior performance on response style. Specifically, the Spoken-LLM-*chunk* achieves 49.6 F1 score on response emotion with 62.1 F1 score on response speaking speed.

Chunk vs. utterance-level embedding: We compare the granularity of speech embedding on both ParalinGPT and Spoken-LLM methods. Results show that the use of chunk embedding achieves better performance on response style prediction. As for response text, the Spoken-LLM benefits significantly from chunk embedding while ParalinGPT performs similarly. In general, chunk embedding extracts richer style-related information than average-pooling embedding, which is more helpful in modeling response speech.

4.4 Subjective evaluation

We perform the human listening evaluation to compare the generated samples of two methods. Specifically, we compare the proposed Spoken-LLM-*chunk* with Text-LLM (text-only) and Text-LLM (cascaded) baseline. As shown in Figure 3, Spoken-LLM wins over the Text-LLM (text-only) method by a large margin, demonstrating that it is important to consider the speaking style information to respond properly. On the other hand, human listeners slightly prefer more on Text-LLM (cascaded) than Spoken-LLM. This result can be explained in two ways: 1) From the objective evaluation of response text, the performance of Spoken-LLM and Text-LLM (cascade) is similar, so the human listeners might not differentiate the content difference, and 2) for the response style, it is possible to respond with more than one response style but still

[4] Score calculated by Hugging face Evaluate package

Training data	Response text				Response style		
	BLEU	ROUGE$_l$	METEOR	BERT$_{f1}$	F1$_{emotion}$	F1$_{speed}$	F1$_{volume}$
train	2.9	15.7	17.8	75.5	45.5	61.6	**61.7**
unfiltered	3.4	17.0	18.7	75.7	44.1	61.2	56.5
unfiltered→train	**4.0**	**17.8**	19.4	**76.3**	**49.6**	**62.1**	61.1

Table 3: Different training strategy and data usages on Spoken LLM-*chunk* method. “→" indicates the two-stage warmup training pipeline.

Method	self-BLEU
Ground truth	8.2
Text-LLM (text-only)	100.0
Text-LLM (cascaded)	11.2
ParalinGPT-*chunk*	11.3
Spoken-LLM-*chunk*	**10.9**

Table 4: The dialogue set-level self-BLEU score for different methods on the evaluation set.

sounds reasonably natural. Therefore, the current and response speaking style is not a one-to-one but a one-to-many relation. Future efforts should consider modeling with more than one response style for better performance and evaluation.

5 Analyses

5.1 Same sentence in different speaking styles induce diverse responses

Since the proposed StyleTalk evaluation set provides sets of the same dialogue context and current input content with two or three distinct speaking styles, we can analyze how diverse are the responses for each input speaking style. To measure the response text diversity, we adopt the self-BLEU (Zhu et al., 2018) score to measure the diversity of each dialogue set. Precisely, we average the BLEU score of two response sentences given two speaking styles as the **dialogue set-level self-BLEU score**. The lower self-BLEU score indicates the generated text is more diverse according to different speaking styles. The results are shown in Table 4. We observe that Spoken-LLM generates the most diverse response content compared to Text-LLM (cascaded) and ParalinGPT. In contrast, the Text-LLM (text-only) baseline generates the same content regardless of different speaking styles, yielding 100% self-BLEU score.

5.2 Warmup pre-training and data quality

Table 3 discusses different training strategies. Firstly, when we only utilize LLM-generated unfiltered data for training, despite the data amount being abundant compared to the train set, the performance of the response style is worse than the train set. Meanwhile, we observe that training on the unfiltered set can achieve better performance on response text, probably because the data amount of the train set is too small and prone to overfitting. We reveal that pre-training on the unfiltered set and fine-tuning on the train set (unfiltered→train) can boost the performance significantly, which enables the model to learn the task and the common language usage first and then align the human standard with the train set.

5.3 Qualitative example

We show the qualitative example in Table 5 with different models' outputs. This example shows that the Text-LLM (text-only) baseline predicts a more neutral sentiment response text, while the text-LLM (cascaded) and Spoken-LLM model generate text with a more aggressive and engaging tone, and the predicted response speaking styles are similar to the ground truth. More quantitative analyses are in Appendix D (style transition) and E (Diversity of current and responding style).

6 Related works

Speech-text Multimodal LLM: The progress in speech Self-supervised Learning (SSL) (Mohamed et al., 2022; Yang et al., 2021; Tsai et al., 2022; Lin et al., 2023a) and neural audio codec (Zeghidour et al., 2021; Défossez et al., 2022; Wu et al., 2023b; Yang et al., 2023; Kumar et al., 2023) enable extracting discrete speech units, drawing attention to generative spoken language modeling. Specifically, the discrete speech units are treated as a special language for unit language modeling (Borsos et al., 2023; Lakhotia et al., 2021; Kharitonov et al., 2022; Nguyen et al., 2023; Hassid et al., 2023), further enabling multiple speech processing tasks in single multimodal LLM (Rubenstein et al., 2023; Zhang et al., 2023; Chou et al., 2023; Maiti et al., 2023; Wang et al., 2023; Wu et al., 2023a;

Dialogue context	A : I’ve finally hit my goal of running five miles every day! B : That’s amazing, hard work really does pay off, doesn’t it? A : Absolutely, and I’ve never felt more alive!
Current text	B: I really need to get back into exercising too.
Not knowing current style	**Text-LLM (text-only)**: Totally understand, let’s start small together, maybe a walk or something?
Current style in <neutral, normal, quiet>	**Text-LLM (cascaded)**: <friendly, normal, normal>Definitely! Let’s find a day when we can go together; I could use the company. **Spoken-LLM**: <cheerful, normal, normal>That’s the spirit! Let me know if you want any tips or company on your runs! **Ground truth**: <friendly, normal, normal>Hey, we could be workout buddies if you want, I can help you get started!
Current style in <unfriendly, fast, loud>	**Text-LLM (cascaded)**: <neutral, normal, normal>Hey, no pressure though, take things at your own pace, you know? **Spoken-LLM**: <unfriendly, normal, normal>Whoa, no need to stress out about it, we can start with something small if you want. **Ground truth**: <unfriendly, normal, normal>Whoa there, no need to get upset, maybe we can find a way to ease you into a routine?

Table 5: A qualitative example. The model outputs the response of speaker A’s turn.

Pan et al., 2023; Nachmani et al., 2023). However, these works mostly leverage content information in speech, due to the speech unit clustering and the use of speech-transcript pairs for modality alignment. The multimodal LLM itself does not learn to model speaking style, and the models are mostly trained with single-turn utterances.

The other line of work aims for the universal speech and audio understanding model to have general audio perception and hearing ability, either leveraging off-the-shelf expert models (Huang et al., 2023; Shen et al., 2023), or with a single MM-LLM (Chu et al., 2023; Tang et al., 2023; Gong et al., 2023b,a; Deshmukh et al., 2023). Those methods are mainly trained to perform comprehensive speech and audio understanding tasks and then fine-tuned for audio-based instruction-following data generated by off-the-shelf LLM like GPT-3.5. However, they are *limited to only generating text responses without considering responding styles*, and the *data quality from LLM is unknown*. In contrast, we focus on modeling speaking style in speech-to-speech conversation with a manual-filtered dataset.

Speaking Style in Spoken Dialogue: Speaking style is important for *speech understanding* and *response generation* in spoken dialogue. The understanding of speaking styles in spoken dialogue is crucial for extracting style attributes such as emotion, sentiment, sarcasm, and more. Representative corpora for studying speaking styles in spoken conversations include IEMOCAP (Busso et al., 2008), SEMAINE (McKeown et al., 2010), MUStARD (Castro et al., 2019), Switchboard-sentiment (Chen et al., 2020), MELD (Poria et al., 2019), MEISD (Firdaus et al., 2020), and MSP-improv (Busso et al., 2016). These datasets are primarily constructed based on label annotations from real speech conversations (e.g., TV series) or acted spoken conversations.

Another research direction involves *spoken conversation in the form of spoken question answering*. Datasets in this category include NMSQA (Lin et al., 2022), SCQA (You et al., 2022), OpenSAQA (Gong et al., 2023a), and E-chat200 (Xue et al., 2023), where the data sample is presented as a tuple of (question, answer). Specifically, for style-related questions and answers, OpenSAQA employs GPT-3.5 to generate textual questions based on the speech content and metadata style information, while E-chat considers text with emotion labels as the question for GPT-3.5 to generate the responding answer as gold answers.

In all existing datasets, only one style is attached to the speech, and one corresponding response speech exists for each conversational context. Thus, prohibiting researchers from investigating the impact of different styles given the same context and the same words. Additionally, the SQA data in OpenSAQA and E-chat are fully generated by GPT-3.5 and not carefully checked by humans, resembling distillation and prompting of GPT-3.5, which is concerning whether the sample follows a human standard as spoken conversation. Our work provides the spoken dialogue data with the same context, the same current text with different speaking styles, and the corresponding response speech with human annotator filtering.

7 Conclusion

This paper focuses on enhancing LLM by modeling how the same sentence spoken with different speaking styles causes different responses in speech, in the spoken conversation scenario. Due to the absence of a suitable dataset, we first collect the speech-to-speech StyleTalk dataset that contains the same dialogue context the same sentence spoken in different styles, and the corresponding different response speech. Next, we propose Spoken-LLM, a two-stage multi-modal training framework to capture different speaking styles and respond properly. The proposed method yields better performance than the text and speech baseline on objective metrics and performs better than text-only LLM on subjective evaluation. We encourage the research community to use the released StyleTalk for joint speaking style and language modeling.

Limitation

Data scale: The current StyleTalk training set consists of only around 2K samples, which may lead to training instability and overfitting. Utilizing a larger-scale dataset could alleviate these issues and eliminate the need for a pre-training stage on unfiltered LLM-generated data.

Real speech with diverse and mixed styles: The speech data in StyleTalk is synthesized from the Azure TTS system with style control. However, incorporating spontaneous speech with even more diverse styles is preferable. Moreover, the current one-hot emotion simplified the problem since speech emotion may by expressed multi-label distribution.

Direct speech-to-speech modeling: The Spoken-LLM generates predefined style attributes to feed into the expressive TTS system. Future work on directly modeling responding speech has the potential to eliminate the need for explicit style labels.

Toward human-like spoken dialogue: Real-world human communication includes backchannel, laughter, and turn-taking behaviors, which is beyond the turn-based spoken dialogue system (Nguyen et al., 2023; Mitsui et al., 2023). Future endeavors to explore speaking style with those behaviors can make the spoken dialogue model closer to human conversation.

Acknowledgement

We want to express our gratitude to the reviewers for their detailed feedback and actionable suggestions, which helped us strengthen our paper. Guan-Ting Lin is supported by a joint NTU GICE and Nvidia Ph.D. scholarship program. Additionally, we thank the National Center for High-performance Computing (NCHC) of the National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

References

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. 2023. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pages 1416–1429. PMLR.

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.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629.

Eric Chen, Zhiyun Lu, Hao Xu, Liangliang Cao, Yu Zhang, and James Fan. 2020. A large scale speech sentiment corpus. In *Proc. LREC*, pages 6549–6555.

Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski, and Michael Auli. 2023. Toward joint language modeling for speech units and text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6582–6593.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *arXiv preprint arXiv:2305.11834*.

Kevin Everson, Yile Gu, Huck Yang, Prashanth Gurunath Shivakumar, Guan-Ting Lin, Jari Kolehmainen, Ivan Bulyko, Ankur Gandhe, Shalini Ghosh, Wael Hamza, et al. 2024. Towards asr robust spoken language understanding through in-context learning with word confusion networks. *arXiv preprint arXiv:2401.02921*.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Meisd: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th international conference on computational linguistics*, pages 4441–4453.

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023a. Joint audio and speech understanding. *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023b. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2023. Textually pretrained speech language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Mutian He and Philip N. Garner. 2023. Can ChatGPT Detect Intent? Evaluating Large Language Models for Spoken Language Understanding. In *Proc. INTERSPEECH 2023*, pages 1109–1113.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, et al. 2022. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved RVQGAN. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

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

Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu wen Yang, Hsuan-Jui Chen, Shuyan Annie Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Lin shan Lee. 2022. DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering. In *Proc. Interspeech 2022*, pages 5165–5169.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G Ward. 2023a. On the utility of self-supervised models for prosody-related tasks. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111. IEEE.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-yi Lee, and Ivan Bulyko. 2023b. Paralinguistics-enhanced large language modeling of spoken dialogue. *arXiv preprint arXiv:2312.15316*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2023. Voxtlm: unified decoder-only models for consolidating speech recognition/synthesis and speech/text continuation tasks. *arXiv preprint arXiv:2309.07937*.

Gary McKeown, Michel F Valstar, Roderick Cowie, and Maja Pantic. 2010. The semaine corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*, pages 1079–1084. IEEE.

Kentaro Mitsui, Yukiya Hono, and Kei Sawada. 2023. Towards human-like spoken dialogue generation between ai agents from written dialogue. *arXiv preprint arXiv:2310.01088*.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Eliya Nachmani, Alon Levkovitch, Julian Salazar, Chulayutsh Asawaroengchai, Soroosh Mariooryad, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2023. Lms with a voice: Spoken language modeling beyond speech tokens. *arXiv preprint arXiv:2305.15255*.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

OpenAI. 2023. Gpt-4 technical report.

Jing Pan, Jian Wu, Yashesh Gaur, Sunit Sivasankaran, Zhuo Chen, Shujie Liu, and Jinyu Li. 2023. Cosmic: Data efficient instruction-tuning for speech in-context learning. *arXiv preprint arXiv:2311.02248*.

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.

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 Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, Jiatong Shi, et al. 2022. Superb-sg: Enhanced speech processing universal performance benchmark for semantic and generative capabilities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8479–8492.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023. Viola: Unified codec language models for speech recognition, synthesis, and translation. *arXiv preprint arXiv:2305.16107*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023a. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Yi-Chiao Wu, Israel D Gebru, Dejan Marković, and Alexander Richard. 2023b. Audiodec: An open-source streaming high-fidelity neural audio codec. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Qian Chen, and Lei Xie. 2023. E-chat: Emotion-sensitive spoken dialogue system with large language models. *arXiv preprint arXiv:2401.00475*.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu, and Yuexian Zou. 2022. End-to-end spoken conversational question answering: Task, dataset and model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1219–1232.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021.

Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Appendix

A Details of human annotators filtering

We assign three listeners for each test. All listeners are based in the United States with HIT approval rate higher than 95%, given that the corpus is in American English. Only the pairs that receive a majority vote and do not have anyone choosing "None of the above is natural" are retained in our corpus. Each test contains 20 samples for evaluation. The example of the annotation interface is shown in Figure 7. We pay the annotators 3 USD for each test. On average, based on the time of playing audio (if played twice for each sample) and reading the content, it takes 10 minutes on one test, so the hourly wage is around 18 USD.

B Implementation details

The model is trained using a two-stage approach with distinct learning rates. The learning rate is 1e-3 and 2e-4 during the 1-stage and 2-stage, respectively. The batch size is 128, and LoRA (r=8) is utilized for efficient fine-tuning of the LLM. To facilitate stable training, a warmup learning rate strategy is with 100 initial steps then linear decay. We use 10% of the training samples as the validation data to assess model performance during training. Model checkpoint is selected based on the validation set performance. In the inference stage, a temperature of 0.7 was applied to control the randomness of generated outputs, and top-p sampling with a probability threshold of 0.95 was used. The number of trainable parameters is 7.8M (0.11% for total parameters). All experiments are run with a single A40 48G GPU.

C ASR prediction as input

In this particular setup, we use the Whisper base ASR (Radford et al., 2023) model to transcribe the current speech into text, which is then input into the trained model for inference. The Word Error Rate (WER) on the current turn speech within the evaluation set is 3.21%. In this setup, we test with the text-LLM (cascaded) and Spoken-LLM-*chunk* models. In Table 6, compared to using the ground truth transcripts, we observe slight performance degradation in response text and style for both the Spoken-LLM-*chunk* and text-LLM (cascaded). It's important to note that addressing ASR error propagation on LLM is beyond the scope of this paper. However, several previous efforts have delved into investigating methods to mitigate such issues (He and Garner, 2023; Everson et al., 2024), which may be one of the further directions especially when the more expressive and spontaneous speech as current input speech.

D Style transition

In this section, we delve into an analysis of the correlation between input and output emotions. While the dataset comprises diverse samples with varying dialogue contexts and inputs, human responses exhibit discernible patterns associated with specific styles. Notably, individuals are inclined to respond with particular styles given a certain current style, and conversely, they are less likely to adopt certain styles in their responses. For instance, in cases where the input style is cheerful, the corresponding response style is more inclined towards positivity, such as cheerful and friendly emotions, as opposed to styles such as unfriendly or sad.

In Figure 4, we present a visual representation of the output style distribution corresponding to different input styles. The visualization reveals that for each input style, certain response styles are markedly more prevalent than others, underscoring the nuanced relationship between input and output emotions.

E Diversity of current and responding styles

In exploring human responses across varied styles, individuals may employ more assertive or passive

Method	Response text				Response style		
	BLEU	ROUGE$_l$	METEOR	BERT$_{f1}$	F1$_{emotion}$	F1$_{speed}$	F1$_{volume}$
Text-LLM (cascaded)	3.1 (-0.1)	16.9 (-0.4)	18.5 (-0.6)	75.9 (-0.1)	37.0 (-0.5)	52.5 (-0.4)	63.7 (-2.1)
Spoken-LLM-*chunk*	3.3 (-0.7)	17.1 (-0.9)	19.0 (-0.4)	75.9 (-0.4)	47.5 (-2.1)	60.3 (-1.8)	57.4 (-3.7)

Table 6: The results of using whisper base ASR model's prediction as input current text on text-LLM (cascaded) and Spoken-LLM-*chunk*.

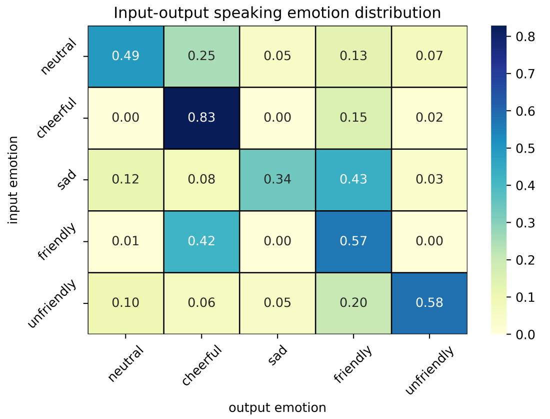


Figure 4: The output emotion distribution given input emotion. Each row is the probability distribution for an input-output pair.

speaking approaches, resulting in potential differences in content. In this context, we delve into an examination of how the interplay between input and output styles influences response diversity. Specifically, we seek to determine whether the transition between styles in certain scenarios leads to responses characterized by increased diversity or a tendency to adopt simpler, more predictable patterns. This investigation sheds light on the intricate dynamics of style transitions and their impact on the richness and complexity of response text.

In Figure 5, we present the result of the top-5 and bottom-5 diverse pairs, organized according to their self-BLEU scores. The self-BLEU score represents the average BLEU score for each style transition pair, with lower scores indicating greater diversity in the responses. Notably, we observe that the top-5 diverse pairs frequently involve responses characterized by positive and excitement styles such as cheerfulness. Conversely, the bottom-5 non-diverse pairs are associated with empathy, particularly when the input style is sad. This analysis provides insights into the response diversity across various style transition scenarios, emphasizing notable patterns in the use of distinct emotional styles.

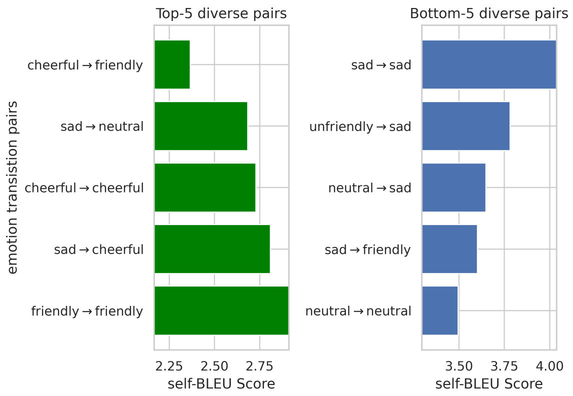


Figure 5: Top-5 and Bottom-5 diverse pairs in the train and evaluation set. The self-BLEU is normalized for each style transition pair to make a fair comparison. The pairs with fewer than 5 pairs are removed. The lower the self-BLEU score, the more diverse the lexical response given different dialogue contexts and input.

	eval	train	unfiltered
# dialogue set 1	16	1,770	0
# dialogue set 2	445	108	859
# dialogue set 3	25	0	4,918
# sample	981	1,986	16,472

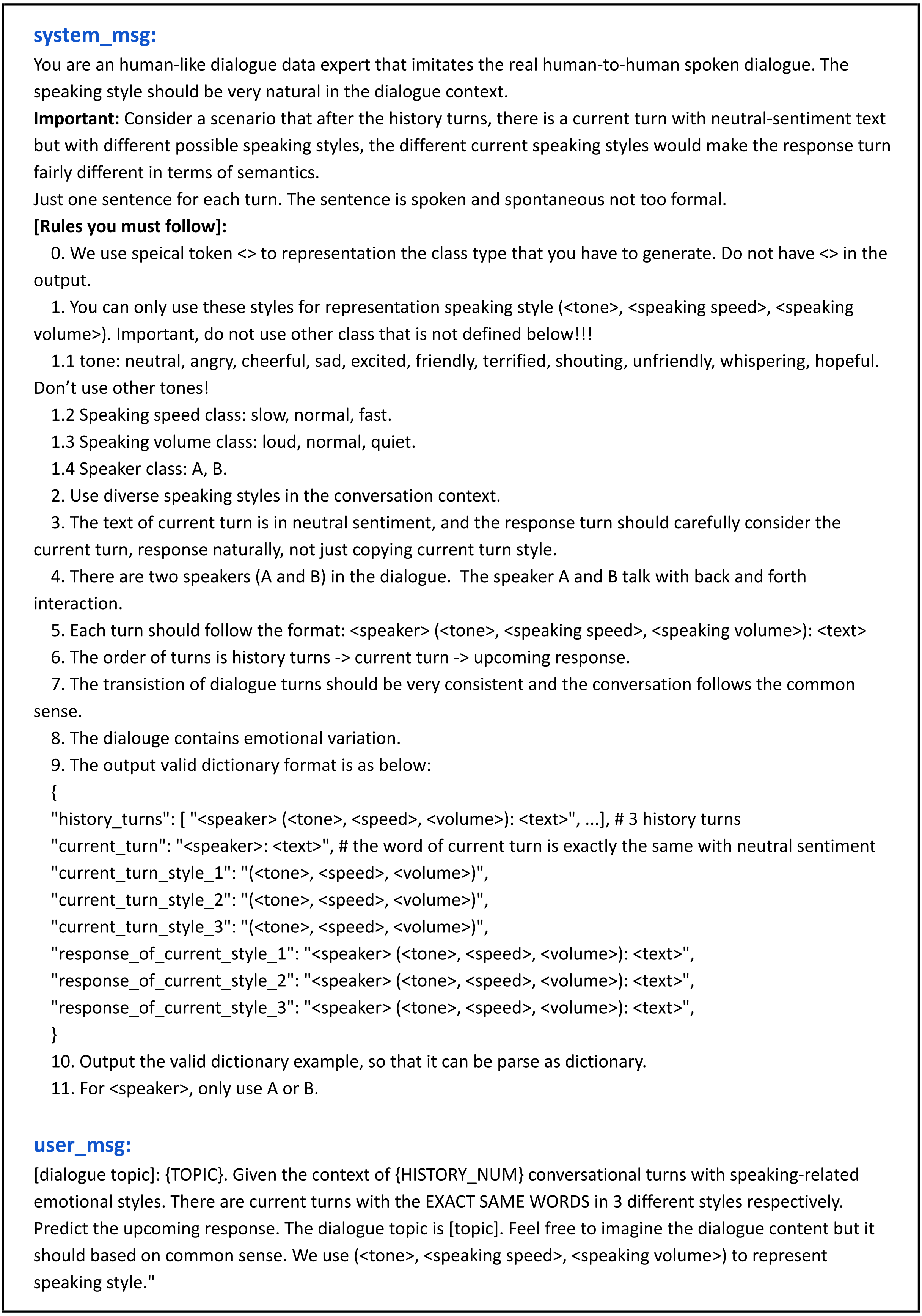
Table 7: Data statistics of StyleTalk. The # dialogue set 1, 2, and 3 mean the amount of different speaking styles for the current speech. # sample is the number of current and response speech pairs.

F Instruction

I_1: Instruction: Classify speaking style of speech. The speaking style is represented in (emotion, speed, volume).
I_2: Instruction: Generate human-like response given context. speaking style is represented in (emotion, speed, volume).

G Prompting GPT-4 for Data Generation

We utilize `gpt-4-1106-preview` and the prompt template in Figure 6.



system_msg:

You are an human-like dialogue data expert that imitates the real human-to-human spoken dialogue. The speaking style should be very natural in the dialogue context.

Important: Consider a scenario that after the history turns, there is a current turn with neutral-sentiment text but with different possible speaking styles, the different current speaking styles would make the response turn fairly different in terms of semantics.

Just one sentence for each turn. The sentence is spoken and spontaneous not too formal.

[Rules you must follow]:

0. We use speical token <> to representation the class type that you have to generate. Do not have <> in the output.

1. You can only use these styles for representation speaking style (<tone>, <speaking speed>, <speaking volume>). Important, do not use other class that is not defined below!!!

1.1 tone: neutral, angry, cheerful, sad, excited, friendly, terrified, shouting, unfriendly, whispering, hopeful. Don’t use other tones!

1.2 Speaking speed class: slow, normal, fast.

1.3 Speaking volume class: loud, normal, quiet.

1.4 Speaker class: A, B.

2. Use diverse speaking styles in the conversation context.

3. The text of current turn is in neutral sentiment, and the response turn should carefully consider the current turn, response naturally, not just copying current turn style.

4. There are two speakers (A and B) in the dialogue. The speaker A and B talk with back and forth interaction.

5. Each turn should follow the format: <speaker> (<tone>, <speaking speed>, <speaking volume>): <text>

6. The order of turns is history turns -> current turn -> upcoming response.

7. The transition of dialogue turns should be very consistent and the conversation follows the common sense.

8. The dialouge contains emotional variation.

9. The output valid dictionary format is as below:

```
{
"history_turns": [ "<speaker> (<tone>, <speed>, <volume>): <text>", ...], # 3 history turns
"current_turn": "<speaker>: <text>", # the word of current turn is exactly the same with neutral sentiment
"current_turn_style_1": "(<tone>, <speed>, <volume>)",
"current_turn_style_2": "(<tone>, <speed>, <volume>)",
"current_turn_style_3": "(<tone>, <speed>, <volume>)",
"response_of_current_style_1": "<speaker> (<tone>, <speed>, <volume>): <text>",
"response_of_current_style_2": "<speaker> (<tone>, <speed>, <volume>): <text>",
"response_of_current_style_3": "<speaker> (<tone>, <speed>, <volume>): <text>",
}
```

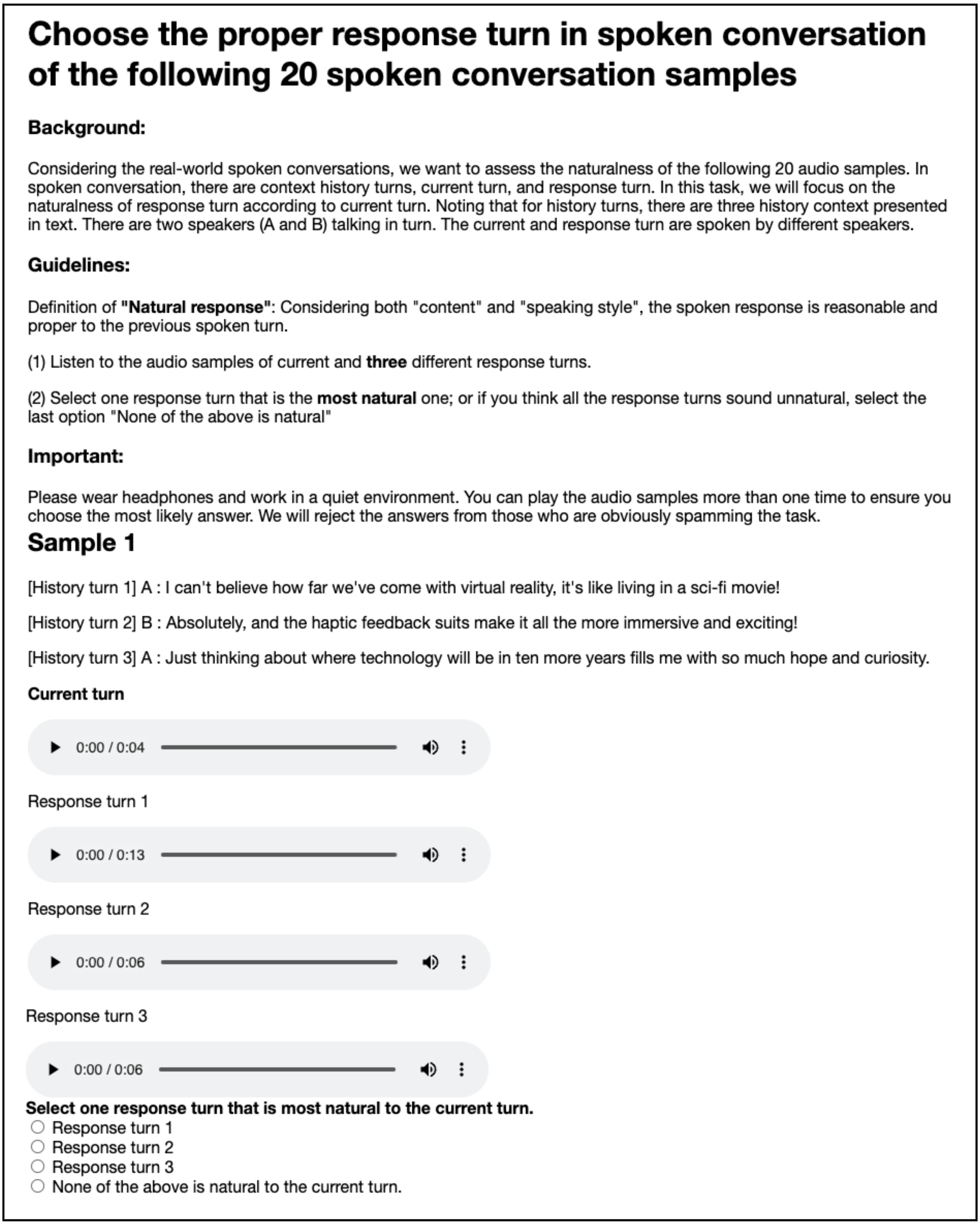
10. Output the valid dictionary example, so that it can be parse as dictionary.

11. For <speaker>, only use A or B.

user_msg:

[dialogue topic]: {TOPIC}. Given the context of {HISTORY_NUM} conversational turns with speaking-related emotional styles. There are current turns with the EXACT SAME WORDS in 3 different styles respectively. Predict the upcoming response. The dialogue topic is [topic]. Feel free to imagine the dialogue content but it should based on common sense. We use (<tone>, <speaking speed>, <speaking volume>) to represent speaking style."

Figure 6: Prompt template. {TOPIC} and {HISTORY_NUM} are variables. The system message and user message are sent to GPT-4 (`gpt-4-1106-preview`) API.



Choose the proper response turn in spoken conversation of the following 20 spoken conversation samples

Background:

Considering the real-world spoken conversations, we want to assess the naturalness of the following 20 audio samples. In spoken conversation, there are context history turns, current turn, and response turn. In this task, we will focus on the naturalness of response turn according to current turn. Noting that for history turns, there are three history context presented in text. There are two speakers (A and B) talking in turn. The current and response turn are spoken by different speakers.

Guidelines:

Definition of **"Natural response"**: Considering both "content" and "speaking style", the spoken response is reasonable and proper to the previous spoken turn.

(1) Listen to the audio samples of current and **three** different response turns.

(2) Select one response turn that is the **most natural** one; or if you think all the response turns sound unnatural, select the last option "None of the above is natural"

Important:

Please wear headphones and work in a quiet environment. You can play the audio samples more than one time to ensure you choose the most likely answer. We will reject the answers from those who are obviously spamming the task.

Sample 1

[History turn 1] A : I can't believe how far we've come with virtual reality, it's like living in a sci-fi movie!

[History turn 2] B : Absolutely, and the haptic feedback suits make it all the more immersive and exciting!

[History turn 3] A : Just thinking about where technology will be in ten more years fills me with so much hope and curiosity.

Current turn

▶ 0:00 / 0:04

Response turn 1

▶ 0:00 / 0:13

Response turn 2

▶ 0:00 / 0:06

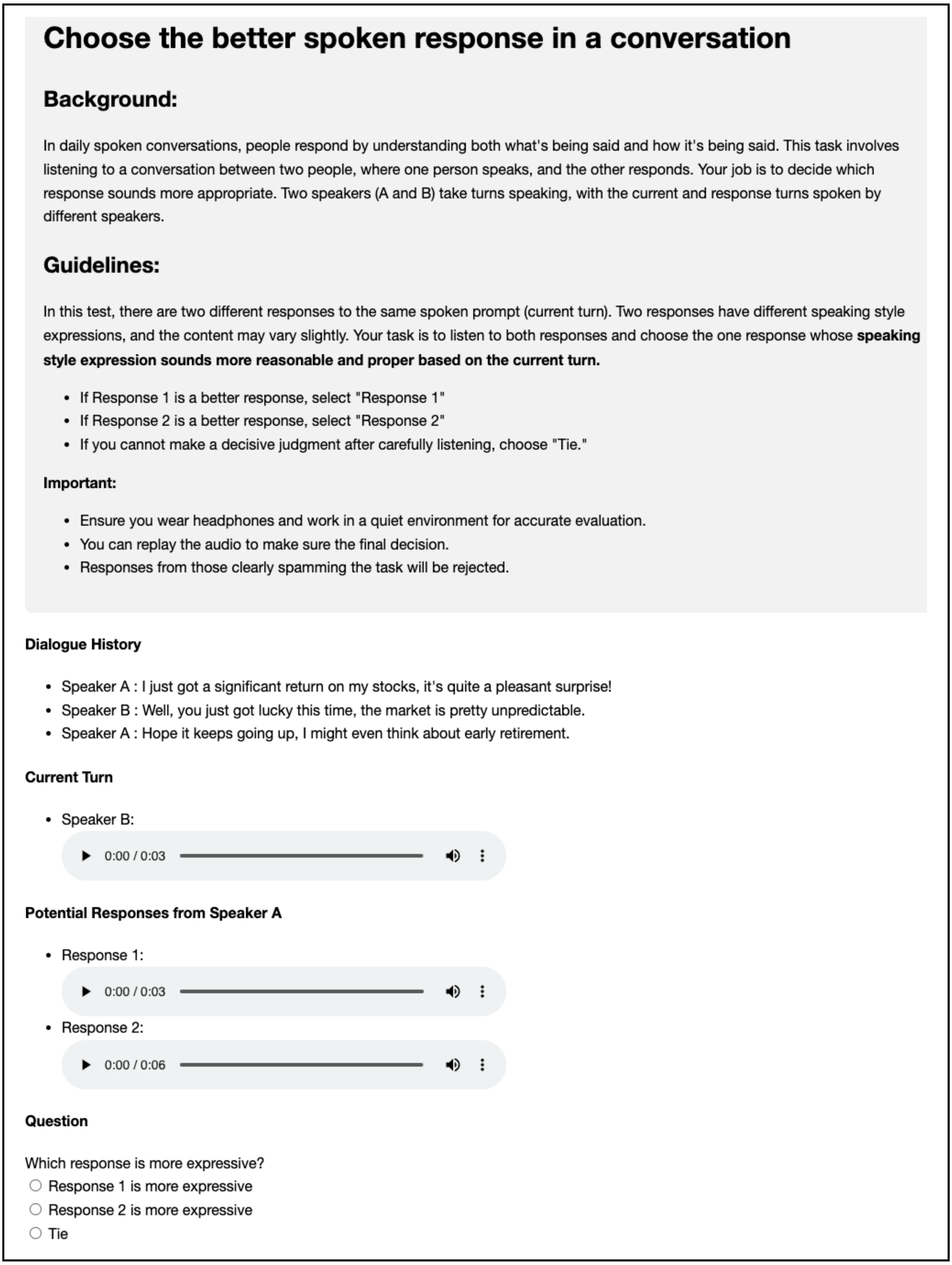
Response turn 3

▶ 0:00 / 0:06

Select one response turn that is most natural to the current turn.

○ Response turn 1
○ Response turn 2
○ Response turn 3
○ None of the above is natural to the current turn.

Figure 7: Human annotators filtering template.



Choose the better spoken response in a conversation

Background:

In daily spoken conversations, people respond by understanding both what's being said and how it's being said. This task involves listening to a conversation between two people, where one person speaks, and the other responds. Your job is to decide which response sounds more appropriate. Two speakers (A and B) take turns speaking, with the current and response turns spoken by different speakers.

Guidelines:

In this test, there are two different responses to the same spoken prompt (current turn). Two responses have different speaking style expressions, and the content may vary slightly. Your task is to listen to both responses and choose the one response whose **speaking style expression sounds more reasonable and proper based on the current turn.**

- If Response 1 is a better response, select "Response 1"
- If Response 2 is a better response, select "Response 2"
- If you cannot make a decisive judgment after carefully listening, choose "Tie."

Important:

- Ensure you wear headphones and work in a quiet environment for accurate evaluation.
- You can replay the audio to make sure the final decision.
- Responses from those clearly spamming the task will be rejected.

Dialogue History

- Speaker A : I just got a significant return on my stocks, it's quite a pleasant surprise!
- Speaker B : Well, you just got lucky this time, the market is pretty unpredictable.
- Speaker A : Hope it keeps going up, I might even think about early retirement.

Current Turn

- Speaker B:

Potential Responses from Speaker A

- Response 1:

- Response 2:

Question

Which response is more expressive?
○ Response 1 is more expressive
○ Response 2 is more expressive
○ Tie

Figure 8: Subjective evaluation template.

H Subjective evaluation

Each test contains 10 samples for evaluation. The example of subjective evaluation interface is shown in Figure 8. We pay the annotators 3 USD for each test. On average, based on the time of listening to audio (if played three times for each sample) and reading the content, it takes 10 minutes on one test. The hourly wage is around 18 USD.

I Dataset license

We released the StyleTalk dataset under the MIT license.