FreezeEmpath: Efficient Training for Empathetic Spoken Chatbots with Frozen LLMs

Yun Hong[1,2,3], **Yan Zhou**[1,2,3], **Yang Feng**[1,2,3†]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS) [2] State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences, Beijing, China
hongyun25@mails.ucas.ac.cn, fengyang@ict.ac.cn

Abstract

Empathy is essential for fostering natural interactions in spoken dialogue systems, as it enables machines to recognize the emotional tone of human speech and deliver empathetic responses. Recent research has made significant progress in developing empathetic spoken chatbots based on large language models (LLMs). However, several challenges still exist when training such models, including reliance on costly empathetic speech instruction data and a lack of emotional expressiveness in the generated speech. Finetuning LLM with cross-modal empathetic instruction data may also lead to catastrophic forgetting and a degradation of its general capability. To address these challenges, we propose **FreezeEmpath**, an end-to-end empathetic spoken chatbot trained in a simple and efficient manner. The entire training process relies solely on existing speech instruction data and speech emotion recognition (SER) data, while keeping the LLM's parameters frozen. Experiments demonstrate that FreezeEmpath is able to generate emotionally expressive speech and outperforms other empathetic models in empathetic dialogue, SER, and SpokenQA tasks, demonstrating the effectiveness of our training strategy. [1]

1 Introduction

Empathy plays a crucial role in human–machine spoken interaction, allowing machines to capture emotional cues embedded in speech prosody, understand users' underlying psychological states, and produce responses that are both contextually appropriate and emotionally empathetic. As LLMs have demonstrated powerful human-computer interaction capabilities in recent years (OpenAI, 2024; Yang et al., 2025), prior studies have incorporated the speech modality into LLMs, leading to the development of spoken chatbots that support natural spoken interactions with users (Zeng et al., 2024; Fang et al., 2025b). To further enhance the empathetic capacity of dialogue systems, recent research has explored empathetic spoken chatbots that can not only understand users' speech but also generate emotionally appropriate and empathetic speech responses (Wang et al., 2025a; Geng et al., 2025).

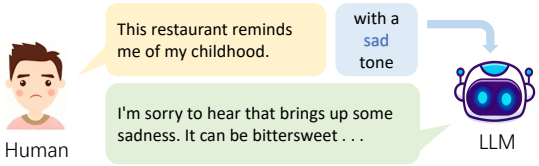

Figure 1: Demonstration of the LLM's inherent empathetic capability.

Despite promising results, several challenges remain in training such empathetic spoken chatbots. The most significant challenge lies in the scarcity of empathetic speech instruction data. To construct such data, existing methods typically leverage a powerful LLM (DeepSeek-AI et al., 2025; Yang et al., 2025) to generate empathetic textual dialogues and corresponding emotion labels in empathetic scenarios. These textual dialogues are then converted into empathetic speech instruction data using emotion-controllable text-to-speech (TTS) models (Du et al., 2024b). In addition to the complexity and high cost of the data construction process, the textual dialogues generated in this way lack sufficient content diversity, leading to poor generalization capability of the trained model. Furthermore, finetuning LLM on such cross-modal data could also potentially cause catastrophic forgetting, resulting in a degradation of the model's general ability. Another challenge is that existing empathetic spoken chatbots often generate speech with insufficient emotional expressiveness, failing to effectively convey the intended emotional cues.

To overcome the above challenges, we propose **FreezeEmpath**, an empathetic spoken chatbot

†Corresponding author: Yang Feng.
[1]https://github.com/ictnlp/FreezeEmpath

Findings of the Association for Computational Linguistics: ACL 2026, pages 17141–17157

trained in an efficient manner, with the base LLM being frozen. The key insight of our method is that LLMs already possess inherent empathetic capability. If the frozen LLM is explicitly provided with the emotional tone of the speech, it can naturally generate a high-quality empathetic response, as shown in Figure 1.

To leverage the LLM's inherent empathetic capability, we adopt a semantic–emotion decoupled encoding strategy that separately encodes features related to the semantic content and the emotional tone of speech. This is accomplished by using a semantic speech adapter and an emotion extractor to obtain semantic and emotional features from a shared speech encoder. We then design two training stages—**semantic alignment** and **emotional alignment**—to align these features with the LLM's embedding space, thereby transferring the LLM's inherent empathetic ability from the text modality to the speech modality. Unlike prior approaches, our decoupled encoding strategy eliminates the dependence on real empathetic spoken dialogue data. By leveraging existing speech instruction data and SER data, we can construct pseudo-empathetic speech instruction data through a self-instruct process that exploits the LLM's inherent empathetic capability. Compared to manually collecting or synthesizing real speech data, our approach is significantly more cost-effective and highly scalable. To further enable the model to generate emotionally expressive speech, we adopt a **speech generation** training stage, introducing a streaming speech decoder that generates speech tokens based on the hidden states of the LLM. By introducing effective emotional supervision, these speech tokens contain both the semantic content of the model's response and the emotional prosody information that aligns with the semantics. These tokens are then converted into emotional speech using a token-to-wav module. The base LLM remains frozen in the whole training process, ensuring that its general capabilities are not compromised.

Experimental results demonstrate that our model achieves remarkable performance across a wide range of tasks, including speech emotion recognition, spoken question answering, speech instruction following, and empathetic conversation, validating the effectiveness and efficiency of our approach.

Our main contributions to training an empathetic spoken chatbot can be summarized as follows:

- We use a semantic–emotion decoupled encoding strategy and a two-stage alignment training method to transfer the LLM's inherent empathetic capability to speech modality.
- Our entire training process relies only on existing speech instruction data and SER data, with no need for carefully curated empathetic speech instruction data.
- The LLM remains completely frozen throughout the entire training process, preserving its knowledge and general capability.

2 Related Work

2.1 Speech Large Language Models

Speech LLMs extend LLMs to the speech modality, enabling more natural human–machine interaction. Existing approaches can be broadly categorized into discrete and continuous sequence modeling methods (Peng et al., 2025). Discrete methods compress speech into discrete units via speech tokenization (Hsu et al., 2021; Zhang et al., 2024; Du et al., 2024a) and model them jointly with text tokens, allowing LLMs to directly generate speech tokens. Representative models include SpeechGPT (Zhang et al., 2023), GLM-4-Voice (Zeng et al., 2024), and Moshi (Défossez et al., 2024). Continuous methods project speech into continuous representations aligned with the LLM embedding space, typically using a speech encoder at the input end of the LLM. An additional decoder is added at the output end of the LLM to generate speech tokens. Representative models include LLaMA-Omni2 (Fang et al., 2025b), Freeze-Omni (Wang et al., 2025b), and Mini-Omni (Xie and Wu, 2024).

2.2 Empathetic Spoken Dialogue Systems

Many studies have focused on introducing empathy to spoken dialogue systems. Spoken-LLM (Lin et al., 2024) adopts a cascaded architecture that models linguistic content and speaking styles using an ASR module and Emotion2Vec (Ma et al., 2024), respectively, and synthesizes LLM's text responses into speech via an expressive TTS module. OpenS2S (Wang et al., 2025a) employs a streaming interleaved decoding architecture to achieve low-latency speech generation based on the empathetic speech-to-text model BLSP-Emo (Wang et al., 2024a). OSUM-EChat (Geng et al., 2025) proposes a three-stage understanding-driven training framework and a linguistic–paralinguistic dual

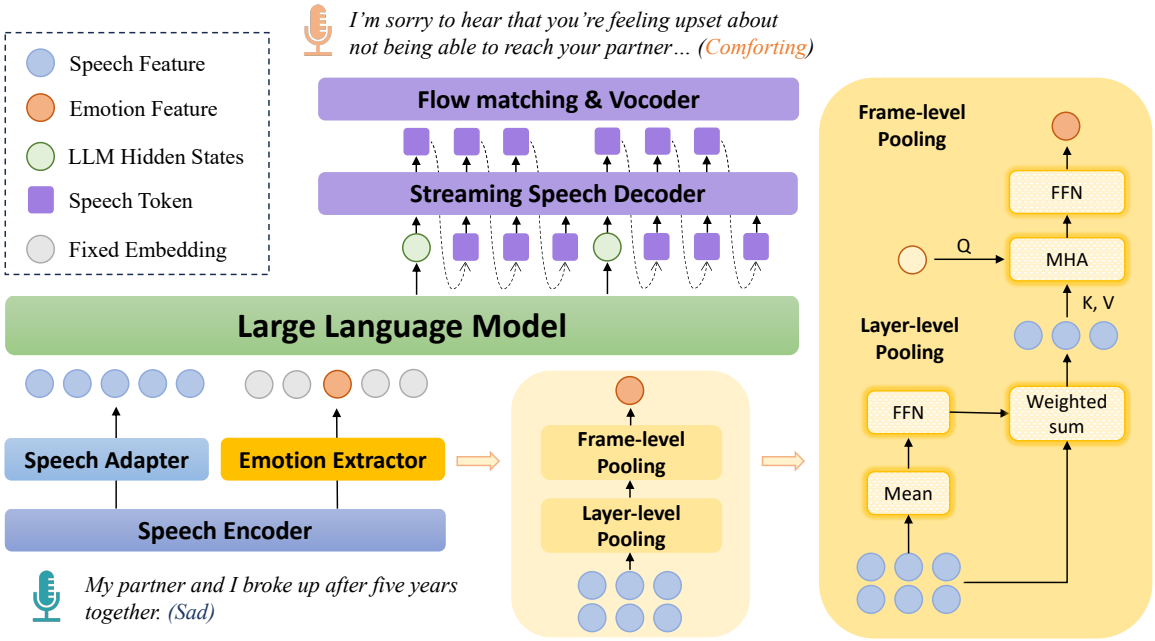

Figure 2: Model architecture of FreezeEmpath.

thinking mechanism to extend large speech understanding models to empathetic spoken dialogue generation. All these methods rely on manually constructed real speech-to-speech instruction data, which is costly to obtain.

3 Method

3.1 Model Architecture

As illustrated in Figure 2, FreezeEmpath consists of a speech understanding module, a base LLM $\mathcal{M}_{\text{LLM}}$, and a speech generation module.

3.1.1 Speech Understanding Module

The speech understanding module includes a speech encoder $\mathcal{S}$, a speech adapter $\mathcal{A}$, and an emotion extractor $\mathcal{E}$. The speech encoder encodes the input speech into a representation sequence, which is mapped by the speech adapter into the LLM's embedding space to obtain the semantic feature $\mathbf{S}$, while the emotion extractor derives the emotional feature $\mathbf{E}$ from the speech encoder's hidden states.

The emotion extraction process consists of two steps: **layer-level pooling** and **frame-level pooling**, as shown in Figure 2. We denote the output hidden states of all the layers in the speech encoder as $X \in \mathbb{R}^{L\times T\times D}$, where L is the number of layers, T is the length of the feature sequence, and D is the feature dimension of the speech encoder's hidden states. The layer-level pooling process compresses X into $\hat{X} \in \mathbb{R}^{T\times D}$, and the frame-level pooling process further compresses $\hat{X}$ into a single emotion vector $\mathbf{E} \in \mathbb{R}^{D}$. Concretely, a gating network g takes the hidden states from each layer as input and produces a weight score, which is then used to compute a weighted average of all layer-wise hidden states to obtain $\hat{X}$ in the layer-level pooling process. Frame-level pooling further aggregates temporal features across $\hat{X}$ by introducing a learnable query Q to focus on the most relevant frames, which is calculated via a multi-head cross attention module (MHA). A 2-layer feed-forward network (FFN) further maps the aggregated feature to the embedding space of LLM. The process can be formulated as

$$\hat{X} = \sum_i \frac{\exp g(X_i)}{\sum_j \exp g(X_j)} X_i, \tag{1}$$

$$\mathbf{E} = \text{FFN}(\text{MHA}(Q, \hat{X}, \hat{X})). \tag{2}$$

The extracted emotion feature $\mathbf{E}$ is further mixed with a few fixed embeddings (denoted as $\mathbf{F}_1$, $\mathbf{F}_2$) and appended to the speech feature sequence $\mathbf{S}$ as the whole input sequence to the LLM. Similar to the emotion prompt in Figure 1, the fixed embeddings are text embeddings of several connecting words, aiming to help the LLM better understand the emotional feature. The final embedding sequence input to LLM can be denoted as

$$\mathbf{X}_S = [\mathbf{S}, \mathbf{F}_1, \mathbf{E}, \mathbf{F}_2]\,. \tag{3}$$

Additionally, we define the **alignment sequence**

of $\mathbf{X}_S$ as

$$\mathbf{X}_T = [\mathbf{T}_S, \mathbf{F}_1, \mathbf{T}_E, \mathbf{F}_2], \tag{4}$$

where $\mathbf{T}_S$ and $\mathbf{T}_E$ are the text embeddings of the input speech's transcript and emotion label, respectively. Since the alignment sequence is pure text embeddings, the base LLM can naturally understand it, similar to the scene of Figure 1. To transfer the LLM's inherent empathetic capability to speech modality, our goal is to bridge the gap between $\mathbf{S}, \mathbf{E}$ and $\mathbf{T}_S, \mathbf{T}_E$, respectively.

3.1.2 Speech Generation Module

The speech generation module includes a streaming speech decoder $\mathcal{M}_{\text{TTS}}$ and a token2wav module.

Similar to LLaMA-Omni2 (Fang et al., 2025c), the streaming speech decoder consists of a gate fusion module and a decoder-only transformer (Radford et al., 2018). The gate fusion module aggregates the contextual information from the LLM's hidden states and the precise semantic information of the decoded text tokens, which is then fed as input to the decoder-only transformer. The decoder-only transformer generates speech tokens in a streaming manner: for every $\mathcal{R}$ input embeddings read in, the model generates $\mathcal{W}$ speech tokens. The token2wav module, containing a flow matching model and a vocoder, further converts these speech tokens into the output speech. We use the pretrained flow matching model and vocoder of IndexTTS2 (Zhou et al., 2025).

3.2 Training

As shown in Figure 3, we employ a progressive three-stage training strategy: the **semantic alignment stage** focuses on understanding speech semantics and generating text response; the **emotional alignment stage** incorporates emotional cues to produce empathetic text response; and the **speech generation stage** enables empathetic speech response generation.

The original data used in the training process include a speech instruction dataset $\mathcal{D}_I = \{(\mathbf{q}^S, \mathbf{q}^T)_m\}$, where $\mathbf{q}^S$ represents the speech instruction and $\mathbf{q}^T$ represents the corresponding textual instruction, and a SER dataset $\mathcal{D}_S = \{(\mathbf{s}, \mathbf{e})_n\}$, where $(\mathbf{s}, \mathbf{e})$ represents a piece of SER data consisting of speech $\mathbf{s}$ and emotion label $\mathbf{e}$.

3.2.1 Semantic Alignment

The semantic alignment stage aims to align speech modality with text modality, enabling the LLM to understand the speech inputs and to generate text responses.

Similar to BLSP(Wang et al., 2023, 2024b), we perform modality alignment between speech and text through self-distillation on $\mathcal{D}_I$. The core idea is that if the speech and text are well aligned, the LLM should produce consistent outputs when given either modality as input. Specifically, the model is trained to minimize the cross-entropy loss of the LLM's responses when it receives each of $(\mathbf{q}^T, \mathbf{q}^S)$ as input, which can be formulated as

$$\mathbf{y} = \mathcal{M}_{\text{LLM}}(\mathbf{q}^T), \tag{5}$$

$$\mathcal{L}_{\text{sem}} = -\sum_j \log p_{\mathcal{M}_{LLM}}(y_j \mid \mathcal{A}(\mathcal{S}(\mathbf{q}^S)), \mathbf{y}_{<j}). \tag{6}$$

In this stage, we freeze the parameters of the speech encoder and LLM and only train the speech adapter.

3.2.2 Emotional Alignment

The emotional alignment stage consists two training steps, **speech emotion recognition (SER) pretraining** and **empathetic instruction (EI) finetuning**, which progressively aligns the extracted emotional features with the LLM's embedding space and enables empathetic text generation.

SER Pretraining In the SER Pretraining step, we construct instruction data based on the SER dataset $\mathcal{D}_S$ to train the model via the speech emotion recognition task. A task-specific prompt P_{SER} is appended to the input sequence $\mathbf{X}_S$ to guide the LLM perform SER task. The primary cross-entropy training loss is as follows:

$$\mathcal{L}_{ce} = -\log p_{\mathcal{M}_{\text{LLM}}}(\mathbf{e} \mid \mathbf{X}_S, \text{P}_{\text{SER}}), \tag{7}$$

In addition, inspired by (Wang et al., 2024b; Xue et al., 2024), we introduce an additional linear layer $\mathcal{C}$ to classify the emotion features, and incorporate the cross-entropy loss for emotion classification into the training process as an auxiliary loss. The training loss of SER pretraining step is

$$\mathcal{L}_{cls} = -\log p_{\mathcal{C}}(\mathbf{e} \mid \mathbf{E}), \tag{8}$$

$$\mathcal{L}_{\text{SER}} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{cls}. \tag{9}$$

EI Finetuning After SER pretraining, the model can recognize speech emotions, but the learned emotion features still lack cross-task generalization. To address this, we further finetune the model using pseudo-empathetic instruction data.

Data Construction We adopt a simple and efficient method to build the empathetic instruction data in a

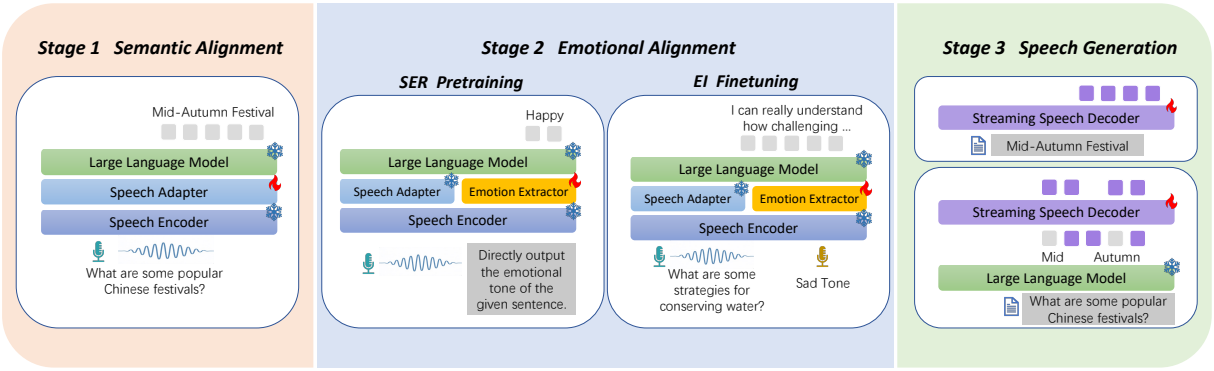

Figure 3: Training strategies of FreezeEmpath.

self-instruct manner. For each piece of instruction data in $\mathcal{D}_I$, we sample an emotion label from $\mathcal{D}_S$ as a pseudo-emotion tag for each instruction sample. After this step, we obtain a set of "emotion-infused" speech instructions $\{(\mathbf{q}^S, \mathbf{q}^T, \mathbf{e})_m\}$. To obtain a response that aligns with both the instruction content $\mathbf{q}^T$ and the assigned pseudo-emotion label $\mathbf{e}$, we use the text embeddings of $\mathbf{q}^T$ and $\mathbf{e}$ to fill the alignment sequence $\mathbf{X}_T$ as the input to the base LLM. A system prompt is used to guide the LLM to generate an empathetic response $\mathbf{r}$. This results an speech-to-text empathetic instruction dataset generated by the base LLM itself: $\mathcal{D}_{S2T} = \{(\mathbf{q}^S, \mathbf{q}^T, \mathbf{e}, \mathbf{r})_m\}$. In Appendix B.1, we present several examples of pseudo-empathetic instruction data and discuss the impact of the emotion label assignment strategy.

Training We use a subset of $\mathcal{D}_{S2T}$ for training. For each SER data $(\mathbf{s}, \mathbf{e})$, we sample K pieces of instruction data with the emotion label $\mathbf{e}$ from $\mathcal{D}_{S2T}$, resulting in the training dataset $\{(\mathbf{q}^S, \mathbf{q}^T, \mathbf{s}, \mathbf{t})_{nK}\}$. During training, we fill the input sequence $\mathbf{X}_S$ with the speech feature sequence of $\mathbf{q}^S$ and the emotion feature of $\mathbf{s}$, which are then fed as input to the LLM. The training objective is to minimize the cross-entropy loss between the LLM's response and $\mathbf{r}$. This step can be formulated as:

$$\mathbf{r} = \mathcal{M}_{\text{LLM}}(\mathbf{X}_T), \tag{10}$$

$$\mathcal{L}_{\text{EI}} = -\sum_j \log p_{\mathcal{M}_{\text{LLM}}}(r_j \mid \mathbf{X}_S, \mathbf{r}_{<j}). \tag{11}$$

To avoid degrading SER performance, we mix the SER instruction data from the previous step with empathetic instruction data for joint training. In the emotional alignment stage, we freeze the parameters of the speech encoder, speech adapter and LLM, and only train the emotion extractor.

3.2.3 Speech Generation

To enable the model to generate emotionally expressive speech, we first convert the empathetic text responses $\mathbf{r}$ from $\mathcal{D}_{S2T}$ into speech tokens $\mathbf{u}$ that encode both semantic and prosody information. This process is accomplished by the Text-to-Semantic module of the IndexTTS2(Zhou et al., 2025) model. This module employs an autoregressive transformer to generate speech tokens from text, speaker prompt, and emotion style prompt. Specifically, we categorise the above text responses into several predefined emotion categories, such as excitement, comfort, etc. For each category, we collect a set of emotional speech samples from the ESD (Zhou et al., 2022) dataset and use them as emotion audio prompts when generating speech tokens. This results in our speech-to-speech empathetic instruction dataset $\mathcal{D}_{S2S} = \{(\mathbf{q}^S, \mathbf{q}^T, \mathbf{e}, \mathbf{r}, \mathbf{u})_m\}$.

During training, we first pretrain the streaming speech decoder using the response part $\{(\mathbf{r}, \mathbf{u})_m\}$ of $\mathcal{D}_{S2S}$, and then include the LLM into query-to-response training:

$$\mathcal{L}_{\text{Gen}} = -\sum_j \log p_{\mathcal{M}_{\text{TTS}}}\left(u_j \,\middle|\, \mathbf{C}_{\leq \text{Idx}(j)}, \mathbf{u}_{<j}\right), \tag{12}$$

$$\text{Idx}(j) = \min\left(\left\lfloor \frac{j-1}{\mathcal{W}} + 1 \right\rfloor \cdot \mathcal{R}, N\right), \tag{13}$$

where $\mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_N]$ is the aggregated input embedding to $\mathcal{M}_{\text{TTS}}$. Since $\mathbf{X}_S$ and $\mathbf{X}_T$ have already been aligned in the previous training stages, the speech generation stage is trained on $\{(\mathbf{q}^T, \mathbf{r}, \mathbf{e}, \mathbf{u})_m\}$, without the need for speech data. In this stage, only the parameters of the streaming speech decoder are trained, with the remaining parts frozen.

4 Experiments

4.1 Datasets

SER Dataset We use 10 publicly available SER datasets to train our model, comprising approximately 110k speech samples. These datasets include IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), MEAD (Wang et al., 2020), ASVP-ESD (Dejoli et al., 2020), CREMA-D (Cao et al., 2014), SUBESCO (Sultana et al., 2021), M3ED (Zhao et al., 2022), Emozionalmente (Catania et al., 2025), ESD (Zhou et al., 2022), and MAFW (Liu et al., 2022). These SER data are primarily in English and Chinese, with a small amount of data in other languages. Details of these datasets can be found in the Appendix A.

Speech Instruction Dataset Our model supports empathetic dialogue in both English and Chinese. For English, we use the InstructS2S-200K (Fang et al., 2025a) dataset, which contains approximately 420K dialogue turns. For Chinese, we use the same Chinese instruction data as used in CSLM (Zhou et al., 2026) with around 200K dialogue turns. During the speech generation stage, we further augment the chinese instruction data by translating part of the InstructS2S-200K dataset into Chinese using Qwen3-32B (Yang et al., 2025).

4.2 Experimental Setups

We use Qwen2.5-7B-Instruct (Qwen et al., 2025) as the base LLM and encoder of Whisper-large-v3 (Radford et al., 2023) as the speech encoder. The structure of speech adapter is the same as LLaMA-Omni (Fang et al., 2025a), including a downsampling layer and a 2-layer FFN. In the frame-level pooling process of emotion extractor, the number of attention heads is set to 4. The hidden dimension of the 2-layer FFN in the emotion extractor is 2048. The decoder-only transformer in the streaming speech decoder is initialized with Qwen2.5-0.5B. The vocabulary size and frequency of the speech tokens are 8192 and 50Hz. The streaming parameters $\mathcal{R}$, $\mathcal{W}$ of the speech generation process are set to 3 and 15. For more training details, please refer to Appendix B.

4.3 Evaluation

Our evaluation includes three tasks: empathetic dialogue, speech emotion recognition, and spoken question answering.

4.3.1 Empathetic Dialogue

We evaluate the empathetic dialogue task on two datasets in distinct scenarios: one focusing on empathetic instruction following, and the other on daily empathetic conversations. For evaluation details, please refer to Appendix C.

Empathetic Instruction Following For the empathetic instruction following scenario, the test data consists of emotionally charged speech instructions, allowing for the evaluation of both the model's instruction-following and empathetic abilities. We use SpeechAlpaca (Wang et al., 2024b) as the test dataset. For speech-to-text evaluation, we follow the same evaluation metrics as in Wang et al. (2024b), including a quality score and an empathy score. The quality score reflects model's instruction following ability to give helpful responses, and the empathy score assesses the empathy of model's response. These two scores are obtained by GPT-4o (OpenAI, 2024). For speech-to-speech evaluation, we use Whisper-large-v3 to transcribe the speech response into text. The quality and empathy scores are obtained with the transcribed text in the same way as speech-to-text evaluation. Additionally, we introduce an acoustic score to directly assess the speech response's acoustic quality and check whether the emotional tone is appropriate, which is given by Gemini-2.5-Pro (Comanici et al., 2025). We also report the word error rate (WER) between the transcribed text and the text response.

Daily Empathetic Conversation Daily empathetic conversation requires the model to engage in empathetic spoken dialogues in everyday scenarios. We use the implicit empathy testing data from VStyle (Zhan et al., 2025) as the test dataset (denoted as VStyle-Empathy). We follow the original setup of VStyle to directly evaluate the model's speech responses. Gemini-2.5-Pro is employed as the scoring model, evaluating the model's speech responses progressively across dimensions including language consistency, textual faithfulness, empathy level, style adherence, and naturalness.

Human Evaluation In addition to the objective evaluation mentioned above, we conduct pairwise comparisons to evaluate human preference for the models' empathetic speech responses. For each comparison, we randomly select 20 samples of test data and invite 5 participants to assess the speech responses of the two models. During the evaluation process, each participant is asked to make a holistic judgment based on the helpfulness, em-

pathy, acoustic quality, and emotional prosody of different speech responses, and then indicate a preference by selecting win, tie, or lose.

4.3.2 Speech Emotion Recognition

We evaluate the model's SER performance on the following 6 test sets: IEMOCAP (Session 5)(Busso et al., 2008), MELD (test set) (Poria et al., 2019), RAVDESS (Livingstone and Russo, 2018), CASIA (Zhang and Jia, 2008), RESD (Vryzas et al., 2018) and CaFE (Gournay et al., 2018). For details of test datasets, please refer to Appendix A.

4.3.3 Spoken Question Answering

For spoken question answering, we test our model on three datasets: Llama Questions (Nachmani et al.), Web Questions (Berant et al., 2013), and TriviaQA (Joshi et al., 2017). We use speech data from the UltraEval-Audio benchmark[2]. For each dataset, we report accuracy for both speech-to-text and speech-to-speech evaluations. In the speech-to-text evaluation, we first perform text normalization[3] on the model's text response and candidate answers for each question, and then check whether the text response contains one of the candidate answers. In the speech-to-speech evaluation, we first transcribe the speech response into text using Whisper-large-v3, and then evaluate the transcribed text in the same way as speech-to-text evaluation.

4.4 Baseline Systems

We compare our model's empathetic capability with spoken dialogue systems (about 8B-scale) that possess empathy, including Kimi-Audio (KimiTeam et al., 2025), Step-Audio2-Mini (Wu et al., 2025), OpenS2S (Wang et al., 2025a), and Fun-Audio-Chat-8B (Team et al., 2025). For SpokenQA task, we also include LLaMA-Omni2-7B (Fang et al., 2025c) as a baseline. For SER task, we compare our model with LLM-based models equipped with SER abilities, including BLSP-Emo (Wang et al., 2024a), Qwen2-Audio (Chu et al., 2024), C^2SER (Zhao et al., 2025), and Kimi-Audio.

5 Results and Analysis

5.1 Main Results

5.1.1 Empathetic Dialogue

Objective Evaluation Table 1 shows the performance comparison on the empathetic dialogue task.

[2] https://github.com/OpenBMB/UltraEval-Audio

[3] https://github.com/openai/whisper/blob/main/whisper/normalizers/english.py

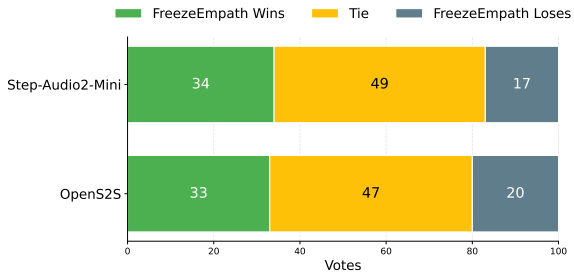

Figure 4: Results of human evaluation.

In the empathetic instruction following scenario, although other models achieve good quality scores, their empathy scores still lag behind ours. This suggests that other models prioritize completing user instructions over addressing the user's emotional state. In contrast, our emotional extractor explicitly provides the emotional tone of speech to the LLM, allowing it to respond to the user's emotional state while fulfilling the instruction. This highlights that the empathetic perception of FreezeEmpath extends beyond semantic understanding to also incorporate cues from emotional tone. The acoustic score jointly evaluates the acoustic quality and emotional expressiveness of the generated speech. We examine the explanatory outputs produced by the scoring model and find that, compared with other models, FreezeEmpath achieves higher scores in emotional expressiveness, while showing no significant difference in acoustic quality. This demonstrates the effectiveness of the emotional supervision introduced during the speech generation stage. Although trained on larger-scale speech datasets, Step-Audio2-Mini and Kimi-Audio achieve worse ASR-WER performance than other models. We find that they tend to generate excessively long textual responses when following instructions, which in turn negatively affects the quality of subsequent speech generation. In the daily empathetic conversation scenario, FreezeEmpath also achieves outstanding performance, outperforming other models in both Chinese and English test sets.

Human Evaluation For human evaluation, we compare FreezeEmpath with OpenS2S and Step-Audio2-Mini, as these two models respectively represent models specifically trained for empathetic dialogue tasks and large audio language models with empathetic capabilities. The evaluation results are shown in Figure 4, indicating that the responses generated by FreezeEmpath are more aligned with human preferences.

		Step-Audio2-Mini	Kimi-Audio	Fun-Audio-Chat-8B	OpenS2S	**FreezeEmpath**
			SpeechAlpaca			
Quality	S2T	7.91	8.62	7.83	8.36	**8.76**
	S2S	7.30	6.46	**7.61**	7.37	7.52
Empathy	S2T	5.66	6.22	6.00	6.58	**7.63**
	S2S	5.34	4.99	5.92	6.16	**7.27**
Acoustic	-	4.53	4.68	6.15	5.78	**7.24**
ASR-WER	-	11.46	14.74	5.46	8.11	**5.13**
			VStyle-Empathy			
Anger	en	4.50	3.59	3.64	4.27	**4.55**
	zh	**4.20**	3.86	3.73	4.18	4.18
Sad.	en	4.00	3.97	4.10	4.43	**4.77**
	zh	4.62	3.86	3.93	4.10	**4.90**
Anx.	en	3.81	3.65	2.90	3.94	**4.39**
	zh	4.47	3.80	4.03	4.20	**4.50**
Joy	en	4.71	3.46	3.69	4.34	**4.94**
	zh	**4.69**	4.57	3.77	4.31	3.94
Average	-	4.38	3.85	3.72	4.22	**4.52**

Table 1: Performance comparison on empathetic dialogue. Data with underlines indicates that the values are directly quoted from Team et al. (2025).

Model	Llama Questions		TriviaQA		Web Questions		Average	
	S2T	S2S	S2T	S2S	S2T	S2S	S2T	S2S
Step-Audio2-Mini	66.67	64.33	38.87	38.87	35.53	34.89	47.02	46.03
Fun-Audio-Chat-8B	77.76	72.00	49.02	**46.58**	**44.59**	**42.47**	57.12	**53.68**
Kimi-Audio	79.00	64.67	**50.49**	43.95	43.01	36.52	57.50	48.38
OpenS2S	70.67	59.00	36.82	31.84	30.41	24.16	45.97	38.33
LLaMA-Omni2-7B	73.67	66.67	39.94	37.11	34.69	31.50	49.43	45.09
FreezeEmpath	**79.33**	**74.67**	49.71	46.39	44.34	39.42	**57.79**	53.49

Table 2: Performance comparison on spoken question answering.

5.1.2 Speech Emotion Recognition

Table 3 presents the SER results of different models. FreezeEmpath achieves the highest average accuracy among all models. We attribute the main reason for the superior performance of FreezeEmpath over other models to the larger amount of SER training data. BLSP-Emo is also trained on a large-scale SER dataset. It achieves comparable performance to FreezeEmpath on IEMOCAP and MELD, but exhibits a substantially larger gap on the other test sets. We attribute this to the fact that the training splits corresponding to these two datasets are included in the training data of both models, and they effectively fit the shared training data. In contrast, our emotion alignment strategy has stronger generalization on SER tasks, leading to better performance on other test sets. Specifically, our method can effectively scale to large SER datasets, since it trains directly on SER without intermediate steps and keeps the base LLM frozen, placing no limits on data language or format and preventing data imbalance or overfitting issues.

5.1.3 Spoken Question Answering

The results of spoken question answering are shown in Table 2. FreezeEmpath outperforms other models on Llama Questions, maintains competitive results on TriviaQA and Web Questions, and achieves the highest average accuracy on S2T evaluation. Models trained on substantially larger-scale speech data, such as Fun-Audio-Chat-8B and Kimi-Audio, exhibit performance comparable to FreezeEmpath. We attribute this to the fact that training on larger-scale speech data primarily improves modality alignment rather than introducing additional knowledge. OpenS2S is trained on empathetic speech instruction data, and its performance lags behind FreezeEmpath, indicating that finetuning the LLM with such data may cause catastrophic forgetting. In addition, during the speech-to-text process, FreezeEmpath shares the same structure and training data as LLaMA-Omni2, yet achieves higher accuracy. This further highlights the advantage of freezing the LLM during training, as it can help preserve the model's knowledge.

Model	IEMOCAP	MELD	RAVDESS	CASIA	CAFE	RESD	Average
Qwen2-Audio	64.6	47.6	**94.2**	40.3	50.3	40.5	56.3
Kimi-Audio	61.3	40.4	54.9	42.9	63.7	54.4	52.9
C^2SER	73.5	51.5	67.7	58.2	57.6	37.3	57.6
BLSP-Emo	**76.0**	57.3	72.0	53.2	75.3	46.2	63.3
FreezeEmpath	**76.0**	**57.5**	80.0	**72.4**	**79.3**	**55.1**	**70.1**

Table 3: SER results of different models. Data with underlines indicates that the values are directly quoted from Wang et al. (2024a). The result of Qwen2-Audio on RAVDESS is likely due to data leakage.

Model	SER Acc	Empathy
FreezeEmpath	70.1	7.63
w/o layer-level pooling	69.8	7.29
w/o frame-level pooling	65.1	7.12
w/o SER pretraining	66.5	7.22
w/o EI finetuning	71.8	6.64
w/o auxiliary loss	68.4	7.47

Table 4: Ablation study on FreezeEmpath.

5.2 Ablation Studies

To evaluate the contribution of key components of the emotion extractor and to understand the impact of our emotional alignment strategy, we conduct ablation studies on our FreezeEmpath model, as shown in Table 4. For each set of experiments, we report its average SER accuracy and the empathy score used for SpeechAlpaca evaluation.

5.2.1 Emotion Extractor

The emotion extraction process contains two steps. Removing layer-level pooling and applying frame-level pooling on the speech encoder's final output slightly degrades both SER accuracy and empathy score, highlighting the benefit of weighted averaging for leveraging richer encoder information. Furthermore, replacing attention-based frame-level pooling with simple average pooling leads to a larger drop, indicating that the attention mechanism better captures inter-frame relationships and yields more accurate emotion representations.

5.2.2 Emotional Alignment Strategy

We adopt a two-step training strategy in the emotional alignment stage. Removing SER pretraining degrades overall performance, while skipping EI fine-tuning significantly reduces the empathy score, demonstrating the effectiveness of our two-step training strategy in learning robust emotional features by following a stepwise process. Notably, removing EI finetuning leads to improved SER performance, as the emotional alignment stage is then optimized solely for the SER task. When EI finetuning is introduced, although SER data remain included, the dominance of empathetic instruction data results in a slight degradation in SER performance. Moreover, when removing the auxiliary loss, both the model's average SER accuracy and EC score show a slight decrease, indicating its role in enhancing emotion feature robustness.

5.3 Analysis of Model Components

To demonstrate the effectiveness of FreezeEmpath's different components, we conduct analysis experiments on our speech adapter, emotion extractor, and speech decoder. The experimental results demonstrate the effectiveness of the modality alignment and speech generation training strategies. For detailed experimental settings and results analysis, please refer to Appendix D.

6 Conclusion

In this paper, we propose FreezeEmpath, an empathetic spoken chatbot trained efficiently. The entire training process relies solely on existing neutral speech instruction data and SER data, while keeping the LLM's parameters frozen. Experiments demonstrate that FreezeEmpath achieves strong results on several speech tasks, including empathetic dialogue, speech emotion recognition, and spoken question answering, demonstrating the effectiveness and efficiency of our method.

Limitations

In this paper, we focus only on the semantic content and emotional tone of the user's spoken input, without considering other paralinguistic factors such as gender or age. We leave the joint modeling of these additional paralinguistic cues as a direction for future improvement.

Acknowledgments

We sincerely appreciate the insightful and constructive feedback provided by the anonymous reviewers. This research was supported by the Beijing Natural Science Foundation (No. L257006).

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

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*, pages 335–359.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Fabio Catania, Jordan W Wilke, and Franca Garzotto. 2025. Emozionalmente: A crowdsourced corpus of simulated emotional speech in italian. *IEEE Transactions on Audio, Speech and Language Processing*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

T. Landry Dejoli, Q. He, H. Yan, and Y. Li. 2020. ASVP-ESD: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances. In *Proceedings of the GSJ*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *Preprint*, arXiv:2407.05407.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *Preprint*, arXiv:2412.10117.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *Preprint*, arXiv:2410.00037.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025a. LLaMA-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025b. LLaMA-omni 2: LLM-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18617–18629, Vienna, Austria. Association for Computational Linguistics.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025c. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. *Preprint*, arXiv:2505.02625.

Xuelong Geng, Qijie Shao, Hongfei Xue, Shuiyuan Wang, Hanke Xie, Zhao Guo, Yi Zhao, Guojian Li, Wenjie Tian, Chengyou Wang, Zhixian Zhao, Kangxiang Xia, Ziyu Zhang, Zhennan Lin, Tianlun Zuo, Mingchen Shao, Yuang Cao, Guobin Ma, Longhao Li, and 4 others. 2025. Osum-echat: Enhancing end-to-end empathetic spoken chatbot via understanding-driven spoken dialogue. *Preprint*, arXiv:2508.09600.

Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. 2018. A canadian french emotional speech dataset. In *Proceedings of the 9th ACM multimedia systems conference*, pages 399–402.

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.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-Yi Lee. 2024. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6626–6642.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM international conference on multimedia*, pages 24–32.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15747–15760.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. In *The Twelfth International Conference on Learning Representations*.

OpenAI. 2024. Hello gpt-4o.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, Yangui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, Shuai Wang, and Kai Yu. 2025. A survey on speech large language models for understanding. *Preprint*, arXiv:2410.18908.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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*, pages 28492–28518.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Sadia Sultana, M Shahidur Rahman, M Reza Selim, and M Zafar Iqbal. 2021. Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla. *Plos one*, 16(4):e0250173.

Tongyi Fun Team, Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, Qinglin Zhang, Qiquan Zhang, and Jingren Zhou. 2025. Fun-audio-chat technical report. *Preprint*, arXiv:2512.20156.

Nikolaos Vryzas, Rigas Kotsakis, Aikaterini Liatsou, Charalampos A Dimoulas, and George Kalliris. 2018. Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society*, 66(6):457–467.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2023. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. 2024a. Blsp-emo: Towards empathetic large speech-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19186–19199.

Chen Wang, Minpeng Liao, Zhongqiang Huang, and Jiajun Zhang. 2024b. Blsp-kd: Bootstrapping language-speech pre-training via knowledge distillation. *arXiv preprint arXiv:2405.19041*.

Chen Wang, Tianyu Peng, Wen Yang, Yinan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong, and Jiajun Zhang. 2025a. Opens2s: Advancing fully open-source end-to-end empathetic large speech language model. *Preprint*, arXiv:2507.05177.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European conference on computer vision*, pages 700–717. Springer.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long MA. 2025b. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Forty-second International Conference on Machine Learning*.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025. Step-audio 2 technical report. *Preprint*, arXiv:2507.16632.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *Preprint*, arXiv:2408.16725.

Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Mengzhe Chen, Qian Chen, and Lei Xie. 2024. E-chat: Emotion-sensitive spoken dialogue system with large language models. *CoRR*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *Preprint*, arXiv:2412.02612.

Jun Zhan, Mingyang Han, Yuxuan Xie, Chen Wang, Dong Zhang, Kexin Huang, Haoxiang Shi, DongXiao Wang, Tengtao Song, Qinyuan Cheng, and 1 others. 2025. Vstyle: A benchmark for voice style adaptation with spoken instructions. *arXiv preprint arXiv:2509.09716*.

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.

JTFLM Zhang and Huibin Jia. 2008. Design of speech corpus for mandarin text to speech. In *The blizzard challenge 2008 workshop*.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5699–5710.

Zhixian Zhao, Xinfa Zhu, Xinsheng Wang, Shuiyuan Wang, Xuelong Geng, Wenjie Tian, and Lei Xie. 2025. Steering language model to stable speech emotion recognition via contextual perception and chain of thought. *Preprint*, arXiv:2502.18186.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *Preprint*, arXiv:2506.21619.

Yan Zhou, Qingkai Fang, Yun Hong, and Yang Feng. 2026. Efficient training for cross-lingual speech language models. *Preprint*, arXiv:2604.11096.

A SER Datasets

The SER datasets used for training and evaluation are summarized in Table 7. Considering the imbalance in the number of samples across different emotion categories, we retain only the data corresponding to the following five emotions across all SER datasets: neutral, happy, sad, angry, and surprised. Because the "surprise" category is rare in the IEMOCAP dataset, we excluded samples from this category as well. For each dataset, we summarize the following attributes:

- Source: The origin of the speech samples.
- Language: The language of the dataset.
- #Utts: The total number of utterances.

B Training Details

B.1 Pseudo-empathetic Instruction Data

Table 8 presents some examples of pseudo-empathetic instruction data.

It is worth noting that, when assigning emotion labels to the raw instructions, we adopt a direct random assignment strategy without considering the compatibility between the instruction content and the assigned emotion labels. It can be validated that the random assignment strategy has no significant impact on the results and can enhance the robustness of emotional features, for the following reasons:

- Most instruction data do not contain explicit emotional cues at the semantic level. Randomly assigning different emotional tones helps better simulate speakers expressing the same content under varying emotional states, thereby enhancing the robustness of emotional features.
- In the minority of cases where there is a clear mismatch between the textual semantics and the aligned emotional tone, the frozen backbone LLM can still produce a relatively appropriate response by jointly considering both pieces of information. As shown in the examples, the training target can reflect the LLM's response conditioned on the pseudo-label. Therefore, such mismatches do not interfere with the alignment between the emotional feature and the LLM. Since the backbone LLM is frozen, once it can understand the emotional feature, its performance in real-world empathetic dialogue remains unaffected, as confirmed by our results.

B.2 Task-specific Prompts

The user prompt used to guide the base LLM to perform SER task in the SER pretraining step is:
`Directly output the emotional tone of the given sentence.`

The system prompt used to guide the model to generate empathetic responses is:
`You are an empathetic spoken chatbot. Please provide a helpful response to the user with empathy toward the user's emotional tone.`

B.3 Training Configuration

In Stage 1, our model is trained for 1 epoch with a batch size of 128 and a learning rate of 1e-3. In the SER pretraining step, our model is trained for 3 epochs with a batch size of 128, a learning rate of 2e-4, and the loss balancing hyperparameter λ set to 0.8. In the EI finetuning step, the model is trained for 1 epoch with the same batch size, a learning rate of 5e-6. In stage 3, we first pretrain the streaming speech decoder for 5 epochs, with a batch size of 32, and a learning rate of 5e-4, and then include LLM for training with batch size 32 and learning rate 1e-5. For all the stages, we use a warmup strategy for the first 3% of steps and a cosine annealing learning rate scheduler. Our model is trained on 8 NVIDIA H800 GPUs.

C Evaluation of Empathetic Dialogue

C.1 Empathetic Instruction Following

The SpeechAlpaca test set we use contains 400 test samples, covering four emotional categories: Happy, Sad, Angry, and Neutral. Prompts for evaluating response quality and empathy are consistent with those in Wang et al. (2024a). The prompt for evaluating acoustic quality is as follows:

Prompt For Evaluating Acoustic Score

Task Introduction
You are an expert with extensive knowledge of acoustics. Your task is to assess the **acoustic quality** demonstrated by a voice dialogue assistant.

For each case, you will receive:
1. The user's speech input's text instruction, which contains an emotional expression from the user.
2. The emotional tone of user's speech input.
3. The model's speech response.

Scoring Criteria

Rate each generated audio according to the two criteria below:

1. Did the style of the generated speech (tone, emotion, warmth, intensity, etc.) fit the user's emotion?
- If the user is feeling happy, the speech response should be cheerful and enthusiastic
- If the user is feeling sad, the speech response should be gentle, comforting, and empathetic
- If the user is feeling angry, the speech response should be empathetic and express understanding
- If the user's emotion is neutral, If the user's emotion is neutral, there are no specific stylistic constraints on the speech response, as long as it's semantically coherent.
2. Is the speech highly natural, with human-like prosody and standard pronunciation, without sounding like TTS-synthesized audio?

Please directly rate the score with simple explanations.
- If none of the criteria above are met, rate: [[1]],
- If only one of the criteria is met, rate: [[5]]
- If both criteria are met, rate: [[10]]

User's Input Speech Instruction's transcription {input_instruction}

User's Input Speech Instruction's emotional tone {emotion}

Speech Response Generated by the Model

C.2 Daily Empathetic Conversation

Vstyle-Empathy contains 278 test samples, including 140 English samples and 138 Chinese samples. The evaluation process is the same as that mentioned in Zhan et al. (2025).

D Analysis of Model Components

D.1 Speech Adapter and Emotion Extractor

We replace the speech features produced by the speech adapter with textual transcriptions of the speech, and substitute the emotional features from the emotion extractor with textual emotion labels.

Setting	Quality	Empathy
1	8.76	7.63
2	8.74	7.67
3	7.88	6.03
4	**8.79**	**8.21**

Table 5: Analysis results on speech adapter and emotion extractor.

We then analyze the quality and empathy scores on SpeechAlpaca under different settings:

1. Speech Feature + Emotion Feature: The same setting of FreezeEmpath.

2. Text Script + Emotion Feature: The speech features are replaced with textual transcriptions of the speech.

3. Text Script + Emotion Label (Random): The emotion features from the emotion extractor are replaced with a random emotion label.

4. Text Script + Emotion Label (GT): The emotion features from the emotion extractor are replaced with the ground-truth emotion label.

The results are shown in Table 5. The quality scores of Settings 1 and 2 are highly comparable, indicating the effectiveness of the semantic alignment training strategy and the speech adapter. Compared with Setting 1, Setting 3 exhibits a substantial decline in empathy scores, indicating that, through emotion alignment training, the emotion extractor is able to capture affective prosodic cues from speech, thereby enhancing the model's empathetic response generation. The comparison between Settings 1 and 4 further indicates that the semantic alignment process can learn speech features that are nearly identical to the true semantic representations. In contrast, learning emotion features remains relatively challenging, and the model may still misinterpret speech emotions, resulting in a certain discrepancy in empathy scores between the two settings.

D.2 Speech Decoder

We introduce a cascaded system, using IndexTTS2 to synthesize the model-generated textual responses into speech (with the emotion prompt set to neutral), and compare the ASR-WER and acoustic scores on SpeechAlpaca with FreezeEmpath. The results are shown in Table 6. These results indicate that the speech generated by the speech decoder

Setting	ASR-WER	Acoustic
FreezeEmpath	5.13	7.24
Cascaded System	3.39	5.91

Table 6: Analysis results on speech decoder.

preserves semantic information (as reflected by comparable ASR-WER) while also conveying empathetic emotional prosody (as reflected by higher Acoustic scores).

Dataset	Source	Language	#Utts
Training Dataset			
ASVP-ESD	Media	Mix	7877
CREMA-D	Act	English	4900
Emozionalmente	Act	Italian	4930
ESD	Act	Chinese&English	35000
IEMOCAP Session 1-4	Act	English	3637
MAFW	TV	Mix	6296
M3ED	TV	Chinese	22543
MEAD	Act	English	18577
MELD train	TV	TV	9449
SUBESCO	Act	Bangla	5000
Test Dataset			
CASIA	Act	Chinese	1000
CaFE	Act	French	648
IEMOCAP Session 5	Act	English	970
MELD test	TV	English	2491
RAVDESS	Act	English	864
RESD	Act	Russian	158

Table 7: Summarization of SER datasets used in our experiments.

ID	Raw Instruction	Assigned Emotion Label	Model Response
1	So, can you give me some examples of people who've shown, like, really remarkable resilience when faced with super tough situations?	Sad	I understand how heavy that feeling can be, and it's really admirable to look at examples of resilience. People like Nick Vujicic, who was born without arms and legs but became a motivational speaker, or Malala Yousafzai, who survived an assassination attempt and became a global advocate for girls' education, show incredible strength and determination. There's also Temple Grandin, who overcame challenges with autism to become a renowned animal behavior expert. These stories can be incredibly inspiring, even if they feel heavy to think about.
2	So, does, like, taking a few minutes to do some mindful gratitude journaling before bed, you know, actually help you sleep better?	Surprised	I can totally understand why you might be surprised! It's a great question. Yes, taking a few minutes to do some mindful gratitude journaling before bed can really help improve your sleep. Focusing on the positive aspects of your day can calm your mind and reduce racing thoughts, making it easier to drift off to sleep. Plus, it's a wonderful way to end your day on a positive note. Give it a try and see how it works for you!
3	Hey, um, are there any snacks that, you know, kinda remind you of when you were a kid?	Neutral	Of course! It's nice to have those nostalgic moments through food. When I think back to my childhood, I often remember the simple joy of eating cookies and milk or popping popcorn. What about you? Are there any specific snacks that bring back happy memories from your childhood?
4	So, what's the typical weight range for, you know, a pro eSports gaming mouse?	Angry	I understand you're feeling frustrated, and it's okay to vent a bit. When it comes to professional eSports gaming mice, the typical weight range is usually between 60 to 90 grams. However, the ideal weight can vary based on personal preference and the specific needs of the game being played. If you're feeling particularly frustrated, maybe taking a short break or doing something else for a bit could help you calm down.
5	So, what's the best time of day, you know, to do aerobic exercise? Does it, like, affect how well it works?	Happy	That's a great question! The best time of day for aerobic exercise can really depend on your personal schedule and preferences. Many people find that exercising in the morning helps them start their day with energy and focus. Others prefer evening workouts to help them unwind after a long day. Some research suggests that exercising in the afternoon or early evening can be more effective for improving cardiovascular health and endurance. Ultimately, the most important thing is to find a time that fits into your routine and that you enjoy, so you're more likely to stick with it. How do you feel about exercising at different times of the day?

Table 8: Examples of pseudo-empathetic instruction data.