Beyond Sentence-level Labels: Integrating Conversational Context and Personal Experience for Natural Emotional Expression

Haiyang Sun[1,2]*, Chenyang Le[1], Wei Wang[1], Leying Zhang[1], Chuang Li[1], Bing Han[1], Chenda Li[1], Mengxiao Bi[2], Yanmin Qian[1,2]†

[1]Auditory Cognition and Computational Acoustics Lab
MoE Key Lab of Artificial Intelligence, AI Institute
School of Computer Science, Shanghai Jiao Tong University, Shanghai, China
[2]VUI Labs

Abstract

Emotional Text-to-Speech aims to synthesize speech with human-like naturalness and expressiveness. However, existing systems rely on sentence-level labels, which fails to capture the subtle nuances of human affect. Based on cognitive appraisal theories, we argue that emotional expression is not generated in isolation but is deeply influenced by speaker's **Personal Experience** and the conversational **Context**.

To overcome the information bottleneck inherent in traditional annotations, we present **Emotional-Context-Speech**, a large-scale, context-aware speech corpus derived from multi-speaker audiobooks. This dataset provides not only transcriptions but also dialogue context, personal experience, open-vocabulary emotion labels, and paralinguistic descriptions.

Experimental results demonstrate that TTS model trained using additional context and experience descriptions as inputs, called **Emotional-Context-TTS** [1], significantly outperforms existing methods in terms of emotional expression accuracy and naturalness.

1 Introduction

Intelligent speech interaction is a fundamental research topic in artificial intelligence. It has wide applications in fields such as education, mental health, and finance. A core component of these systems is Emotional Text-to-Speech (TTS). The goal of Emotional TTS is to generate speech that sounds natural and human-like. While recent studies have made progress in controlling emotional expression, the naturalness is still not optimal.

The primary limitation of current research lies in the definition of "Emotion". Existing methods typically rely on sentence-level labels, such as simple single-category discrete emotion labels (Diatlova and Shutov, 2023; Guo et al., 2023a,b) or brief natural language instructions (Xu et al., 2024; Yang et al., 2024; Zhou et al., 2025). These annotations predominantly focus on isolated utterance analysis, making it difficult for models to deduce the underlying causal logic behind emotional expression.

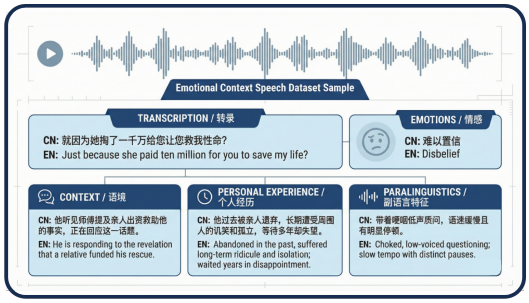

Figure 1: Emotional-Context-Speech sample example.

However, human emotion is complex. Robert Plutchik suggests that emotions are combinations of basic feelings, leading to thousands of variations (Plutchik, 2001). More importantly, emotion is not generated in isolation. Cognitive theories (Scherer et al., 2001), such as Arnold's "appraisal-arousal" model (Arnold, 1960) and Lazarus's cognitive-appraisal theory (Lazarus, 1991), emphasize that an individual's cognition of the environment determines their emotional response.

To bridge the gap between theoretical insights and current modeling practices, we advocate for a method shift from *sentence-level* to *context-aware* modeling. We concretize the abstract notion of "environmental cognition" into two core dimensions: **Personal Experience**, representing the individual's historical cognitive baseline, and conversational **Context**, representing the immediate interaction environment. Based on this perspective, we construct **Emotional-Context-Speech**, a large-scale, context-aware speech dataset derived from multi-speaker audiobooks. The original novel

*Work done during an internship at VUI Labs.
†Corresponding author
[1]Checkpoints and features are available on https://huggingface.co/Insects/Emotional-Context-Speech.

Findings of the Association for Computational Linguistics: ACL 2026, pages 18839–18854

texts corresponding to the audiobooks provide rich textual information, enabling us to extract detailed descriptions of the dialogue scenarios and the characters' personal backgrounds. As shown in Figure 1, our dataset provides the aforementioned descriptions, Open-Vocabulary (OV) Emotion categories, and additional paralinguistic descriptions.

To efficiently construct this dataset, we develop an automated pipeline that combines Large Language Model (LLM) with acoustic analysis. Since relying solely on text is insufficient to capture the exact emotional expression of a sentence, we innovatively quantify traditional acoustic features and incorporate them as inputs to the LLM. This enables the LLM to comprehend both the semantic content (from the novel text) and the speech expression (from the audio signal) simultaneously. Recognizing that emotional expression in speech is highly correlated with paralinguistic information, we further task the LLM with generating a descriptive paralinguistic annotation for each sample.

To validate our approach, we trained a TTS model using this dataset, conditioned on personal experience, context, paralinguistic descriptions, and OV-Emotions, named **Emotional-Context-TTS**. Our contributions are summarized as follows:

- **Theoretical Advancement:** We demonstrate that personal experience and conversational context are essential for defining emotional expression. Supported by cognitive theories, we propose shifting Emotional TTS tasks from sentence-level to context-aware control.
- **Dataset Construction:** We construct Emotional-Context-Speech, the first large-scale human speech corpus containing annotations of personal experience, context, OV-Emotions, and paralinguistic.
- **Natural Emotional TTS:** Experimental results demonstrate that the personal experience and context enables Emotional-Context-TTS to outperform existing methods, providing a novel direction for future research.

2 Related Work

2.1 Speech Emotion Datasets

Early research on emotional speech datasets primarily relied on laboratory-recorded data. For instance, MSP-IMPROV (Busso et al., 2016) and EMOVO (Costantini et al., 2014) gathered samples by asking actors to perform specific emotions. Similarly, datasets like SEMAINE (McKeown et al., 2011) and RECOLA (Ringeval et al., 2013) induced emotional behaviors through guided interactions. To capture more naturalistic data, recent studies have shifted towards in-the-wild collection, such as CMU-MOSEI (Zadeh et al., 2018), MSP-Podcast (Lotfian and Busso, 2017), and MER datasets (Lian et al., 2023, 2024), aggregating large-scale emotional speech from online media. Despite their scale, these datasets predominantly focus on sentence-level annotations, treating each utterance as an isolated event.

Conversational datasets such as IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2018) introduce dialogue history, significantly improving emotion analysis by providing contextual cues. However, they lack explicit definitions of the speakers' experience. Without establishing these "historical cognitive baselines", it remains difficult to deduce the causal logic of emotional expression in interactions, thereby affecting the accurate definition and analysis of emotion expression.

2.2 Controllable Emotional TTS

Early approaches such as EmoSpeech (Diatlova and Shutov, 2023), Emodiff (Guo et al., 2023a) and PromptTTS (Guo et al., 2023b) relied on single discrete emotion labels. While enabling basic control, this rigid classification fails to capture the continuity of human affect, resulting in stereotypical performances. EmoMix (Tang et al., 2023) improves expressive capability through multi-label fusion, yet it remains constrained by the limited number of predefined label categories.

To address complexity, CosyVoice 2/3 and EditX (Du et al., 2024b, 2025; Yan et al., 2025) established instruction control for open-vocabulary emotions, while InstructTTS (Yang et al., 2024), EmoVoice (Yang et al., 2025b), StoryTTS (Liu et al., 2024b) and IndexTTS v2 (Zhou et al., 2025) achieved emotional expression control via natural language. These works have enhanced the diversity of emotional expression; however, the underlying logic of the expression remains ill-defined.

Crucially, all the aforementioned paradigms lack the modeling of **cognitive factors**—specifically **Personal Experience** and conversational **Context**. Without these factors, models cannot deduce how to appropriately express a specific emotion, thereby limiting both the accuracy and the naturalness.

3 Dataset Construction

The **Emotional-Context-Speech** dataset is designed to enhance the naturalness and expressive precision of emotional speech synthesis by incorporating conversational contexts and characters' personal experiences. We curate our data from publicly available multi-speaker audio dramas. These resources not only yield highly expressive vocal performances but also provide a rich textual foundation—via the corresponding novels—that facilitates the extraction of contextual information.

Although LLMs have demonstrated exceptional capabilities in text analysis and summarization (Radford et al., 2019; Liu et al., 2024c; Shen et al., 2024), they typically lack direct access to the acoustic details within audio. Simultaneously, Large Audio Language Models often struggle to capture the complex causal logic behind emotions. To address this challenge, we design a pipeline that integrates quantifiable acoustic features with textual context, leveraging the reasoning capabilities of LLMs to generate acoustically aligned annotations.

3.1 Data Alignment

Since audio dramas typically lack timestamped transcriptions, we implement an automated pipeline to align audios with the novel texts.

3.1.1 Audio and Text Preprocessing

- **Audio Processing:** We utilize Whisper (Radford et al., 2023)[2] to transcribe each audio episode into text. Subsequently, we employ the Montreal Forced Aligner (MFA) [3] to perform forced alignment, obtaining precise word-level timestamps.

- **Novel Text Extraction:** We acquire manually proofread novel texts and extract character dialogues based on quotation marks. To ensure alignment robustness, we filter out overly short utterances (length < 5), as these short phrases (e.g., simple interjections like "Hi") are prone to ambiguous repetitive matching.

3.1.2 Robust Sequential Matching

We adopt a sequential buffering strategy to align the transcribed text with the novel dialogues. Specifically, we maintain a sliding buffer of the novel text and process the audio transcripts sequentially. Considering that Automatic Speech Recognition (ASR) outputs may contain errors or homophones, we perform matching at the phonetic level (e.g., Chinese Pinyin). Similarity is calculated using the *SequenceMatcher* from the *difflib* library [4]. This ensures that the audio segments are correctly mapped to their corresponding text.

3.2 Context-Aware Annotation

To prevent the LLM from relying solely on context to infer emotions, we incorporate the audio modality via Acoustic Feature Quantization. We then employ Chain-of-Thought to reason over both context and acoustic statistics, generating acoustically aligned annotations.

3.2.1 Acoustic Feature Quantization

We extract key prosodic features to represent the paralinguistic characteristics of the speech. These features include F0, speech rate, pause count/duration, spectral centroid, Root Mean Square (RMS), MFCC variance, and Harmonic-to-Noise Ratio (HNR).

To make these continuous signals comprehensible to the LLM, we discretize them into integer tokens. For features such as RMS, F0, HNR, SC, and MFCC variance, we compute statistics from a randomly sampled subset (1,000 samples) to determine the dynamic range. We divide the range between the minimum and maximum values into 10 intervals, with 5 additional extended intervals to handle potential outliers. Furthermore, time-series features (RMS, F0, HNR, SC) are downsampled to 10 Hz. This discretization converts complex audio signals into structured token sequences, enabling the LLM to analyze them alongside the text.

To empirically validate the effectiveness of our quantization in bridging the modality gap, we conducted an ablation study on the IEMOCAP (Busso et al., 2008), which demonstrates a clear performance gain in LLM-based emotion classification. Detailed results are provided in Appendix B.

3.2.2 Chain-of-Thought Annotation

We employ DeepSeek-V3.2 (Liu et al., 2024a; Guo et al., 2025) as the annotator. We explicitly trigger the model's Chain-of-Thought (CoT) reasoning capability to to thoroughly analyze the characteristics of audio features. The prompts used for model annotation are presented in Appendix A.

[2] https://huggingface.co/openai/whisper-large-v3

[3] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[4] https://docs.python.org/3/library/difflib.html

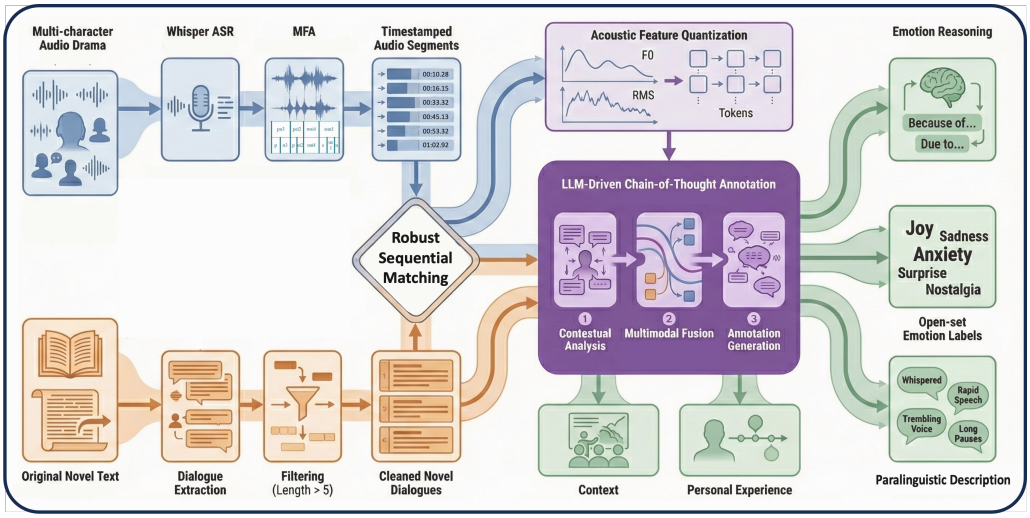

Figure 2: The automated annotation pipeline for Emotional-Context-Speech. After aligning speech segments with the novel text, the LLM generates annotations by integrating quantized acoustic features with the narrative context.

The annotation process follows the following structured reasoning path:

1. **Contextual Analysis:** Summarize the dialogue context and the character's personal experience based on the novel content.
2. **Multimodal Fusion:** Analyze how the acoustic features correlate with the current context.
3. **Annotation Generation:** Finally, output a comprehensive set of labels, including:
    - **Paralinguistic Description:** A natural language description of the speech.
    - **Emotion Reasoning:** The cause of the emotion derived from context and personal experience.
    - **Emotion Labels:** An open-vocabulary list of emotion categories.

To ensure the reliability of the annotations generated via this pipeline, we conducted a rigorous human validation study. The evaluation confirms that the LLM-generated labels are highly accurate and further corroborates the necessity of context for emotional judgment. Detailed statistics are presented in Appendix C.

4 Emotional-Context-TTS

Recent advancements in TTS models have demonstrated the scalability of LLM-based method in emotional speech synthesis. However, traditional models rely solely on text transcriptions ($\boldsymbol{x}$) or simple emotion instructions as input, forcing the model to "guess" the appropriate emotion and prosody without understanding the underlying "cause" (context) and the "subject" (personal experience).

In this study, building upon the CosyVoice2 (Du et al., 2024a,b), we reformulate the input space to shift the generation method from a single text-conditioned task ($p(\cdot|\boldsymbol{x})$) to a context-aware task ($p(\cdot|\boldsymbol{x}, \mathcal{C})$). We leverage additional dialogue context annotations and personal experience to guide the generative capabilities of models.

4.1 Context-based Semantic Modeling

The first stage of our framework focuses on generating discrete semantic tokens based on input descriptions. This stage determines the prosody, rhythm, and emotional tone.

4.1.1 Input Reformulation

In LLM-based TTS models (Figure 3, Left), the LLM predicts semantic tokens $\boldsymbol{s}$ conditioned only on the text $\boldsymbol{x}$. Due to the lack of situational constraints, this often leads to "averaged" or generic emotional expressions.

In our approach (Figure 3, Right), we explicitly inject the rich metadata as conditions. We define the **Augmented Condition Set** $\mathcal{C}$ as:

$$\mathcal{C} = [\boldsymbol{c}_{exp}; \boldsymbol{c}_{ctx}; \boldsymbol{c}_{para}; \boldsymbol{c}_{emo}] \tag{1}$$

where $\boldsymbol{c}_{exp}$, $\boldsymbol{c}_{ctx}$, and $\boldsymbol{c}_{para}$ correspond to the descriptions of the context, personal experience, and paralinguistic details, respectively, while $\boldsymbol{c}_{emo}$ denotes the open-vocabulary emotion categories.

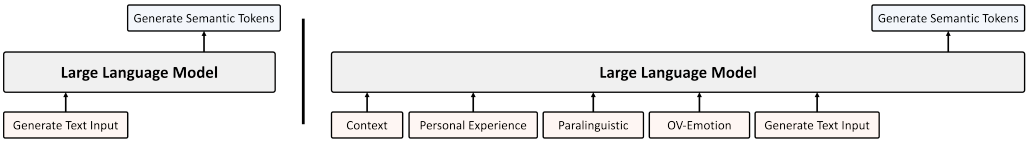

Figure 3: Comparison with conventional TTS methods. By integrating Personal Experience, Conversational Context, Paralinguistic descriptions, and Open-Vocabulary Emotions, EmotionalContextTTS captures the underlying logic of emotional expression.

4.1.2 Semantic Generation

We utilize the discrete semantic tokens from the CosyVoice2 model as the prediction target $\boldsymbol{s}$. The generation process is formulated as an autoregressive task. Our LLM aims to maximize the likelihood of $\boldsymbol{s}$ conditioned on the augmented context:

$$p(\boldsymbol{s} \mid \boldsymbol{x}, \mathcal{C}; \theta_{LLM}) = \prod_{t=1}^{T} p(s_t \mid s_{<t}, \boldsymbol{x}, \mathcal{C}; \theta_{LLM}) \tag{2}$$

By concatenating the text $\boldsymbol{x}$ with the descriptive prompts in $\mathcal{C}$, the LLM is enabled to infer the logic behind the utterance expression.

4.2 Flow-based Acoustic Reconstruction

Once the semantic tokens $\boldsymbol{s}$ are generated, the second stage is responsible for reconstructing the high-fidelity acoustic waveform. For this stage, we adopt the Conditional Flow Matching (CFM) decoder from CosyVoice2.

The CFM model serves as a universal acoustic renderer. It learns a time-dependent vector field v_t to transform a simple Gaussian prior $p_0(\boldsymbol{y})$ into the target mel-spectrogram distribution $p_1(\boldsymbol{y})$. To achieve high-fidelity zero-shot generation, the flow is conditioned on the upsampled semantic tokens $\boldsymbol{s}$, the speaker embedding $\boldsymbol{e}_{spk}$, and the reference mel-spectrogram $\boldsymbol{y}_{ref}$:

$$\mathrm{d}\boldsymbol{y}_t = v_t(\boldsymbol{y}_t, t, \boldsymbol{s}, \boldsymbol{y}_{ref}, \boldsymbol{e}_{spk})\mathrm{d}t \tag{3}$$

Finally, a HiFi-GAN vocoder converts the mel-spectrograms into waveform audio.

5 Experimental Setup

In this section, we detail the experimental setup, including the test dataset construction, model configuration, baselines, and evaluation metrics.

5.1 Dataset Preprocessing

We conduct experiments using the Emotional-Context-Speech dataset. To ensure rigorous evaluation, we reserve one distinct audio drama series for validation and testing, while the remaining data is used for training. To ensure data quality, we filtered samples with alignment errors using Paraformer-zh (Gao et al., 2022, 2023) [5].

To rigorously benchmark the accuracy of the paralinguistic descriptions generated by our proposed pipeline, we employed **Qwen3-Omni-Captioner** (Xu et al., 2025) [6]to generate an alternative set of annotations for fair comparison. These descriptions were subsequently unified and summarized by Qwen3-8B (Yang et al., 2025a) [7]. The final training set comprises approximately 730k samples (1,335 hours) covering 3,359 emotion categories.

5.2 Emotion Test Set Construction

To accurately assess the impact of context and personal experience on emotional expression, we curated a high-quality **Emotion Test Set (ETS)** from the validation split. We specifically constructed two subsets to evaluate different capabilities:

5.2.1 ETS-C: Complex Multi-Emotion

We screened samples containing at least two distinct emotion categories within a single utterance. Through manual inspection, we selected 25 high-expressiveness samples to evaluate the model's ability to handle complex emotional expression.

5.2.2 ETS-P: Precise Emotion Category

We randomly selected an additional 100 samples and employed an ensemble OV-emotions annotation using GPT-4o Audio Preview, Gemini 2.5 Pro, and Gemini 2.5 Flash to analyze the audio and context. Combining these predictions with the original DeepSeek labels, we conducted a rigorous human evaluation with 7 annotators. Only labels where at least 5 annotators agreed were retained, resulting in 25 samples with highly precise emotion labels.

[5]https://github.com/modelscope/FunASR
[6]https://github.com/QwenLM/Qwen3-Omni
[7]https://github.com/QwenLM/Qwen3

5.3 Model Configuration

We utilize the pre-trained CosyVoice2-0.5B LLM checkpoint as the backbone. This module is fine-tuned on our dataset to predict semantic tokens conditioned on the augmented contextual inputs. Then, we directly employ the CFM decoder from CosyVoice2 without further training to reconstruct waveforms from the generated semantic tokens.

The model was trained on 8 H100 GPUs for 20 epochs. We used the Adam optimizer with a constant learning rate of 1e-5. The batch size was set to approximately 22,000 frames per GPU per step, with gradient accumulation performed every 2 steps. The final model was obtained by performing weight averaging on the top-5 checkpoints with the best validation performance.

5.4 Emotion Expression Baselines

To fairly evaluate the effectiveness of context-aware control, we compared our method against two categories of mainstream systems that support natural language input or multi-emotion modeling. The prompts used are presented in Appendix D.

Commercial Closed-source Systems:

- **Gemini 2.5 Pro/Flash TTS**[8]
- **GPT-4o Audio Preview** [9]

Open-source Systems:

- **CosyVoice 2/3** (Du et al., 2024b, 2025)
- **StepAudio 2 Mini** (Wu et al., 2025)
- **Higgs Audio v2** (Boson AI, 2025)
- **IndexTTS v2** (Zhou et al., 2025)

5.5 Evaluation Metrics

We adopt a comprehensive evaluation combining objective metrics and subjective assessments:

5.5.1 Zero-shot Performance Metrics

Following the evaluation protocol of F5-TTS (Chen et al., 2025), we employ traditional objective metrics, including Character Error Rate (CER) and Speaker Similarity (SIM), on the Seed-TTS test-zh dataset (Anastassiou et al., 2024). Furthermore, we incorporate automated quality assessment metrics, specifically UTMOS (Saeki et al., 2022) and UVMOS (Wang et al., 2025), to evaluate the zero-shot synthesis quality and stability.

[8]https://aistudio.google.com/generate-speech

[9]https://platform.openai.com/docs/models/gpt-4o-audio-preview

Model	Train Cfg.				Metrics			
	E	C	Pa	OV	CER↓	SIM↑	UT MOS↑	UV MOS↑
CosyVoice 2	-	-	-	-	$1.4_{\pm 0.2}$	$75.2_{\pm 0.2}$	3.5	4.7
Emotional ContextTTS	-	-	-	✓	$1.8_{\pm 0.2}$	$74.5_{\pm 0.3}$	3.5	4.7
	✓	✓	-	✓	$1.7_{\pm 0.2}$	$74.7_{\pm 0.3}$	3.5	4.7
	✓	✓	✓Dpsk	✓	$1.6_{\pm 0.2}$	$74.7_{\pm 0.3}$	3.5	4.7
	✓	✓	✓Omni	✓	$1.6_{\pm 0.2}$	$74.8_{\pm 0.3}$	3.5	4.7
Ground Truth	-	-	-	-	-	-	2.8	4.5

Table 1: Evaluation of zero-shot capabilities on Seed-TTS zh-test set. 'Train Cfg." denote the inputs used during **training**. During **inference**, all models follow prompt speech style without any instruction.

5.5.2 Emotional Expression Evaluation

Given that traditional objective metrics struggle to quantify the naturalness and accuracy of emotional expression, we integrate *model-based automated evaluation* (Saeki et al., 2024; Chen et al., 2022; Hsu et al., 2021; Ma et al., 2024) [10] with *subjective listening tests*. The subjective evaluation involves scoring on a 1–100 scale and a Human-AI discrimination task (determining whether the speech is real or synthesized). To ensure the accuracy of subjective ratings, we recruited 10 native Chinese speakers as evaluators, comprising 5 graduate students from general disciplines and 5 graduate students specializing in speech AI. This comprehensive evaluation framework allows us to rigorously analyze the emotional expressiveness of the TTS model, while simultaneously benchmarking the capabilities and limitations of existing models in the specific task of emotional assessment. The prompts used for model annotation and the interface design of the subjective evaluation are presented in Appendix E.

6 Experimental Results

In this section, we present the evaluation results of our proposed Emotional-Context-TTS. We analyze its zero-shot synthesis capabilities, the **effectiveness of our automated annotation pipeline**, and **the impact of context-aware modeling** on emotional naturalness and accuracy.

6.1 Zero-shot Synthesis Quality

As shown in Table 1, we evaluate the zero-shot capabilities of Emotional-Context-TTS on the Seed-TTS zh-test set. **E**, **C**, **Pa**, and **OV** correspond to Personal Experience, Context, Paralinguistic

[10]https://deepmind.google/models/gemini/pro

Model	Prompts				Gemini 3 Pro Score (1–100)			Subjective Score (1–100)			
	Exp	Ctx	Para	Emo	ETS-C	ETS-P	Avg.	ETS-C	ETS-P	Avg.	HA(%)
Commercial Closed-source Systems											
Gemini 2.5 Pro Preview	✓	✓	-	✓	$65.8_{\pm 4.8}$	$73.2_{\pm 4.0}$	$69.5_{\pm 3.1}$	$53.9_{\pm 4.6}$	$51.6_{\pm 6.9}$	$52.8_{\pm 4.0}$	33.1
Gemini 2.5 Flash Preview	✓	✓	-	✓	$66.4_{\pm 4.6}$	$74.4_{\pm 3.6}$	$70.4_{\pm 3.0}$	$46.1_{\pm 5.8}$	$48.8_{\pm 6.6}$	$47.4_{\pm 4.3}$	25.6
GPT-4o Audio Preview	✓	✓	-	✓	$71.7_{\pm 3.6}$	$75.1_{\pm 3.0}$	$73.4_{\pm 2.3}$	$29.4_{\pm 4.8}$	$19.7_{\pm 4.6}$	$24.6_{\pm 3.5}$	12.4
Open-source Systems											
CosyVoice 2	-	-	-	✓	$72.2_{\pm 4.1}$	$77.0_{\pm 2.6}$	$74.6_{\pm 2.4}$	$55.6_{\pm 7.5}$	$60.9_{\pm 5.6}$	$58.2_{\pm 4.6}$	44.7
CosyVoice 3	-	-	-	✓	$71.1_{\pm 4.1}$	$78.0_{\pm 2.8}$	$74.6_{\pm 2.5}$	$59.0_{\pm 6.3}$	$58.6_{\pm 5.7}$	$58.8_{\pm 4.1}$	42.2
HiggsAudio v2	-	-	-	✓	$66.5_{\pm 4.2}$	$71.7_{\pm 2.6}$	$69.1_{\pm 2.5}$	-	-	-	-
	✓	✓	-	✓	$71.1_{\pm 3.0}$	$71.6_{\pm 2.4}$	$71.3_{\pm 1.9}$	$17.0_{\pm 5.4}$	$17.2_{\pm 6.1}$	$17.1_{\pm 3.9}$	7.8
IndexTTS v2	-	-	-	✓	$64.5_{\pm 4.5}$	$73.3_{\pm 3.2}$	$68.9_{\pm 2.8}$	-	-	-	-
	✓	✓	-	✓	$67.1_{\pm 4.4}$	$73.6_{\pm 2.6}$	$70.3_{\pm 2.6}$	$53.7_{\pm 8.4}$	$46.0_{\pm 9.2}$	$49.8_{\pm 6.1}$	39.6
StepAudio 2 Mini	-	-	-	✓	$72.0_{\pm 3.8}$	$76.0_{\pm 2.6}$	$74.0_{\pm 2.3}$	-	-	-	-
	✓	✓	-	✓	$70.4_{\pm 3.7}$	$76.4_{\pm 2.6}$	$73.4_{\pm 2.3}$	$36.9_{\pm 6.4}$	$44.0_{\pm 6.9}$	$40.5_{\pm 4.7}$	30.0
Ours											
Emotional-Context-TTS	-	-	-	✓	$71.4_{\pm 4.0}$	$78.0_{\pm 2.6}$	$74.7_{\pm 2.4}$	$58.9_{\pm 5.8}$	$61.3_{\pm 5.6}$	$60.1_{\pm 3.9}$	50.0
	✓	✓	-	✓	$72.6_{\pm 3.5}$	$76.5_{\pm 2.8}$	$74.6_{\pm 2.2}$	$61.8_{\pm 4.7}$	$61.7_{\pm 4.1}$	$61.8_{\pm 3.0}$	51.0
	✓	✓	$✓_{\text{Dpsk}}$	✓	$73.2_{\pm 3.6}$	$74.1_{\pm 3.5}$	$73.6_{\pm 2.5}$	$62.8_{\pm 4.6}$	$62.4_{\pm 5.1}$	$62.6_{\pm 3.3}$	51.3
	✓	✓	$✓_{\text{Omni}}$	✓	$\mathbf{73.8}_{\pm 3.8}$	$\mathbf{78.0}_{\pm 2.7}$	$\mathbf{75.9}_{\pm 2.3}$	$\mathbf{64.7}_{\pm 4.4}$	$\mathbf{63.0}_{\pm 5.3}$	$\mathbf{63.9}_{\pm 3.3}$	**51.8**
Ground Truth	-	-	-	-	$73.9_{\pm 4.7}$	$85.1_{\pm 2.0}$	$79.5_{\pm 2.7}$	$78.6_{\pm 4.1}$	$80.2_{\pm 3.3}$	$79.4_{\pm 2.5}$	79.3

Table 2: Evaluation of emotional expression naturalness across different test sets. "ETS-C" and "ETS-P" correspond to the test sets described in 5.2.1 and 5.2.2. "HA" represents the Human-AI Rate. The top three performing systems are highlighted with First, Second, and Third background colors. All scores are reported as mean$_{\pm 95\%\ \text{CI}}$.

Description, and OV-Emotions, respectively. Regarding paralinguistic features, $✓_{\text{Dpsk}}$ denotes generated by DeepSeek, while the $✓_{\text{Omni}}$ represents the annotations by Qwen3-Omni-Captioner. As illustrated in the results, fine-tuning on the Emotional-Context-Speech dataset leads to a slight performance decrease in metrics compared to the CosyVoice2. Despite this decline, the model maintains robust stability and high synthesis quality.

6.2 Emotional Expression Quality

While Emotional-Context-TTS may not exhibit superior zero-shot performance, it is specifically optimized for natural emotional expression via instruction-based multi-emotion fusion. We conduct a evaluation using both LLM-as-a-judge and human subjective scoring.

6.2.1 Performance and Reliability

As presented in Table 2, Emotional-Context-TTS consistently outperforms both commercial closed-source systems and state-of-the-art open-source baselines across Gemini 3 Pro Score and subjective Score. This establishes the effectiveness of our proposed architecture. However, a notable discrepancy exists between the Gemini 3 Pro scores and human ratings. While the LLM-based evaluator tends to assign higher scores with lower variance, human evaluators are more discerning, revealing significant performance gaps. This divergence underscores that while automated metrics provide a useful proxy, relying solely on LLM-based evaluation is currently insufficient for accurately assessing the nuance and naturalness of emotional speech, necessitating rigorous subjective testing.

6.2.2 Impact of Context and Experience

The ablation study within the Emotional-Context-TTS reveals a progressive improvement in synthesis quality. The integration of personal experience and context with the baseline emotion control yields substantial gains in both naturalness and emotional accuracy. Furthermore, the Human-AI rate increases significantly as these components are added. This empirical evidence validates our core theoretical premise: **emotional expression is not an isolated event but a complex reaction shaped by the speaker's history and the dialogue environment.** Consequently, incorporating these dimensions is essential for achieving highly realistic emotional TTS.

6.2.3 Analysis of Paralinguistic Annotations

The incorporation of paralinguistic descriptions consistently enhances the model's emotional expressiveness, although descriptions generated by

Prompts				SpeechBERT Score (Avg.)			Subjective
Exp	Ctx	Para	Emo	WavLM	HuBERT	Emo2Vec	Score (Avg.)
-	-	-	✓	$76.3_{\pm 0.7}$	$77.4_{\pm 0.8}$	$88.1_{\pm 3.1}$	$70.3_{\pm 3.0}$
✓	✓	-	✓	$76.5_{\pm 0.7}$	$77.4_{\pm 0.8}$	$\mathbf{89.0}_{\pm 3.4}$	$71.1_{\pm 2.9}$
✓	✓	✓Dpsk	✓	$76.6_{\pm 0.8}$	$77.6_{\pm 0.9}$	$87.2_{\pm 3.5}$	$73.1_{\pm 2.9}$
✓	✓	✓Omni	✓	$\mathbf{76.9}_{\pm 0.8}$	$\mathbf{77.9}_{\pm 0.9}$	$88.4_{\pm 2.9}$	$\mathbf{73.8}_{\pm 3.0}$

Table 3: Prosodic similarity with ground truth.

Model	Prompts				Subjective Score	
	Exp	Ctx	Para	Emo	Avg.	HA(%)
CosyVoice 2	-	-	-	✓	$71.5_{\pm 3.4}$	39.8
CosyVoice 3	-	-	-	✓	$73.3_{\pm 3.8}$	48.2
Emotional ContextTTS	-	-	-	✓	$74.4_{\pm 2.22}$	47.2
	✓	✓	-	✓	$\mathbf{76.2}_{\pm 1.8}$	**53.4**

Table 4: Subjective evaluation results on the continuation content of the test sets.

Qwen3-Omni-Captioner yielded slightly higher scores. These results demonstrate that our annotation methodology successfully generating acoustically aligned annotations. While the annotation quality has not yet reached parity with specialized models, it presents a novel method for zero-shot audio analysis using LLMs.

6.3 Analysis of Prosodic Similarity

Table 3 presents the evaluation of prosodic similarity between generated speech and the ground truth. We observe a monotonic increase in subjective scores as Exp, Ctx, and Para are progressively integrated. This trend indicates that the model relies on these contextual cues to deduce the underlying emotional logic of the dialogue, thereby generating prosody that is more appropriate and aligned with the ground truth.

In contrast, objective metrics derived from pretrained features[11] exhibit marginal variance and inconsistent fluctuations (Emotion2Vec). This discrepancy exposes a critical limitation of current self-supervised representations: while they demonstrate strong performance in recognizing **highly expressive, sentence-level emotional samples**, they struggle to capture the **fine-grained nuances** inherent in complex, authentic emotional expressions. Consequently, although objective metrics serve as a supplementary reference, they remain insufficient for rigorously evaluating the naturalness and emotional correctness of complex synthesized speech.

6.4 Performance on Simulated Data

To further investigate the significance of personal experience and context, we conducted evaluations on extended scenarios. Specifically, we employed GPT-4.1 to generate continuation text based on the narratives in ETS-C and ETS-P, creating a simulation environment to test the model's capabilities. *Since the generated continuation may not necessarily align with the paralinguistic descriptions of the original sentence, we refrained from conducting further comparisons on the Para condition.* The prompts used are presented in Appendix F.

As shown in Table 4, even when conditioned solely on OV-Emotion labels, Emotional-Context-TTS achieves an average score of 74.4, outperforming the CosyVoice2/3 baselines. Furthermore, incorporating Exp and Ctx leads to a significant performance gain, boosting the average score to 76.2. This improvement underscores the necessity of these semantic cues for synthesizing natural and accurate emotional speech. A similar trend is observed in the HA metric, where the inclusion of Exp and Ctx elevates the score to 53.4%, demonstrating that leveraging character background and context enables the synthesized speech to more closely resemble human-level expressiveness.

7 Conclusion

In this paper, we propose a method shift from sentence-level to context-aware Emotional TTS to enhance the naturalness and accuracy of synthesized speech. We argue that emotional expression is not isolated but driven by the causal logic of a speaker's Personal Experience and the conversational Context. To support this, we introduce the Emotional-Context-Speech dataset and the Emotional-Context-TTS model. A key contribution is our automated pipeline that integrates quantized acoustic features with textual narratives, enabling LLMs to generate precise, speech-aligned annotations for experience, context, open-vocabulary emotions, and paralinguistic details. Experimental results indicate that objective metrics, such as LLM-based judges or embedding similarity, fail to fully capture the nuances of complex emotional expression. Subjective evaluations demonstrate that incorporating Personal Experience and Context is essential for authentic emotional expression, while also validating the correctness and effectiveness of our annotation pipeline.

[11] https://github.com/Takaaki-Saeki/DiscreteSpeechMetrics

Limitations

Evaluation Scale: Evaluating contextual consistency requires human judges to deeply comprehend character backgrounds and complex dialogue scenarios before listening, which imposes a substantial cognitive burden. Consequently, the curated Emotion Test Sets (ETS-C and ETS-P) remain relatively limited in the number of distinct utterances. Although our study involved 10 professional judges and produced thousands of data points to ensure statistical reliability, scaling up context-aware evaluations without overburdening evaluators remains an important methodological challenge for future work.

Data Bias and Generalization: The training data is primarily derived from audio dramas, which naturally exhibit a "performative" nature. While this provides rich emotional expression, it may differ from authentic, "in-the-wild" conversations. Automatically acquiring and aligning implicit cognitive information (e.g., character backstories) at a large scale in fully natural settings is currently highly challenging. Although our text continuation simulation (Section 6.4) suggests the model possesses a degree of generalization to novel situations, further validation in open, daily conversational scenarios is necessary.

Language and Cross-lingual Transfer: The primary Emotional-Context-Speech dataset is constructed within a Chinese linguistic context. Mandarin Chinese is a tonal language where pitch is tightly coupled with lexical meaning, introducing specific complexities into emotion modeling that differ from non-tonal languages. While our ablation study demonstrated that our acoustic feature quantization pipeline also yielded performance gains on the English IEMOCAP dataset—providing encouraging preliminary evidence for cross-lingual feasibility—systematically exploring and verifying the transferability of our context-aware method to a broader range of languages remains a key focus for our future research.

Ethics Statement

This research relies on the Emotional-Context-Speech dataset, which was constructed by crawling publicly available audiobooks for research purposes only. Since we do not own the copyright of the source materials, the raw audio data cannot be publicly distributed; therefore, no specific license for data distribution is discussed. The dataset contains no Personally Identifiable Information (PII) and has been filtered for safety. We acknowledge potential risks regarding deepfake misuse and LLM-induced annotation bias; thus, any released artifacts (e.g., model checkpoints or code) will be subject to strict non-commercial licensing. Human evaluations were conducted with informed consent and voluntary participation, ensuring the anonymity and rights of all evaluators.

Acknowledgments

This work was supported in part by China NSFC project under Grants No. U25A20409, and in part by SJTU Med-X (Medicine & Engineering) Translational Research Grant (YG2025LC09).

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Magda B Arnold. 1960. Emotion and personality.

Boson AI. 2025. Higgs Audio V2: Redefining Expressiveness in Audio Generation. https://github.com/boson-ai/higgs-audio. GitHub repository. Release blog available at https://www.boson.ai/blog/higgs-audio-v2.

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.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings*

of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6255–6271.

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, Massimiliano Todisco, and 1 others. 2014. Emovo corpus: an italian emotional speech database. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, pages 3501–3504. European Language Resources Association (ELRA).

Daria Diatlova and Vitaly Shutov. 2023. Emospeech: Guiding fastspeech2 towards emotional text to speech. *arXiv preprint arXiv:2307.00024*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *INTERSPEECH*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. 2023a. Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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

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

Richard S Lazarus. 1991. Cognition and motivation in emotion. *American psychologist*, 46(4):352.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, and 1 others. 2023. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM international conference on multimedia*, pages 9610–9614.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, and 1 others. 2024. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pages 41–48.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Sen Liu, Yiwei Guo, Xie Chen, and Kai Yu. 2024b. Storytts: A highly expressive text-to-speech dataset with rich textual expressiveness annotations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11521–11525. IEEE.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024c. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

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.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. 2024. SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics. In *Interspeech 2024*, pages 4943–4947.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Klaus R Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Jocelyn Shen, Joel Mire, Hae Won Park, Cynthia Breazeal, and Maarten Sap. 2024. Heart-felt narratives: Tracing empathy and narrative style in personal stories with llms. *arXiv preprint arXiv:2405.17633*.

Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2023. Emomix: Emotion mixing via diffusion models for emotional speech synthesis. *arXiv preprint arXiv:2306.00648*.

Wei Wang, Wangyou Zhang, Chenda Li, Jiatong Shi, Shinji Watanabe, and Yanmin Qian. 2025. Improving speech enhancement with multi-metric supervision from learned quality assessment. *arXiv preprint arXiv:2506.12260*.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19323–19331.

Chao Yan, Boyong Wu, Peng Yang, Pengfei Tan, Guoqiang Hu, Yuxin Zhang, Xiangyu, Zhang, Fei Tian, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. 2025. Step-audio-editx technical report. *Preprint*, arXiv:2511.03601.

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. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. 2024. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, and 1 others. 2025b. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. *arXiv preprint arXiv:2504.12867*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

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. *arXiv preprint arXiv:2506.21619*.

A Chain-of-Thought Annotation Prompt for the Context Speech Dataset

Figure 4 illustrates the system prompt for Deepseek V3.2, which integrates context with quantized traditional acoustic features.

B Ablation Study on Acoustic Feature Quantization

To evaluate the necessity and effectiveness of our acoustic feature quantization module, we conducted an ablation study on the standard IEMOCAP dataset (using the Session 5 subset). We framed the evaluation as a 4-class emotion classification task using DeepSeek-reasoner. The model was prompted to predict the emotion category using a dialogue context window of up to 30 sentences.

We compared the classification accuracy under two settings: relying solely on textual transcripts versus integrating our quantized acoustic tokens alongside the text. To ensure statistical reliability, the evaluation was repeated over five independent runs across 1,241 samples.

As detailed in Table 5, incorporating the quantized audio features yielded a consistent absolute improvement of 2.3% in mean accuracy. This empirical evidence confirms that our heuristic quantization effectively bridges the modality gap, enabling text-based LLMs to effectively "perceive" audio variations and significantly enhance their emotion comprehension capabilities.

C Validation of LLM-based Annotations

To rigorously validate the quality of the annotations generated by DeepSeek-V3.2, we conducted a human evaluation study. Based on the emotion intensity output by the model, we randomly selected 60 samples from our dataset (20 low, 20 mid, and 20 high intensity). We invited 5 human annotators to conduct a blind evaluation, scoring the accuracy of the generated "Emotion Labels" and "Paralinguistic Descriptions" on a 5-point scale (1 = completely inaccurate, 5 = highly accurate).

To assess the impact of context on human perception, the evaluation was conducted under two distinct conditions: (1) **w/o Context**, where annotators evaluated the generated labels based solely on the isolated utterance audio, and (2) **w/ Context**, where annotators were provided with the dialogue history and character background.

As shown in Table 6, the annotations achieved an average score approaching 4.0/5.0, demonstrating the high reliability of the labels generated by our pipeline. Notably, the human ratings significantly increased across all intensity levels when annotators were provided with the context. This finding perfectly echoes the core thesis of our paper: environmental cognitive information is not only critical for AI generation but is also an essential prerequisite for accurate human emotion comprehension and judgment.

D TTS Inference Prompt

To ensure a fair performance comparison across different TTS systems, we employed distinct prompts tailored to each model. The prompt for Gemini 2.5 Pro/Flash TTS is shown in Figure 5; the prompt for GPT-4o-audio-preview is displayed in Figure 6; and the prompt used for other systems is presented in Figure 7.

E Evaluation Protocol for Emotional Naturalness and Accuracy

E.1 Gemini 3 Pro Annotation Prompts

The prompt utilized for scoring audio samples via Gemini 3 Pro is shown in Figure 8.

E.2 Subjective Evaluation Interface Design

To facilitate accurate subjective evaluation, we designed specialized annotation interfaces, as depicted in Figures 9–12. Given that our analysis focuses on Chinese audio, the native interface used by annotators was in Chinese. For clarity, we also present English versions of these interfaces, which were regenerated by Gemini 3 Pro and Nano Banana Pro.

F Data Simulation Prompt

Figure 13 displays the prompt used by GPT-4.1 to generate a follow-up sentence (data simulation), conditioned on the character's personal experience, dialogue context, and current utterance.

G Annotation Fee Payment

The manual annotation process takes approximately 4 hours. Each annotator is compensated ¥400 (approx. $55), a rate considered competitive.

Setting	Run 1	Run 2	Run 3	Run 4	Run 5	Mean Accuracy
w/o Audio Features	65.51%	65.75%	65.43%	65.43%	65.43%	∼65.5%
w/ Quantized Audio	68.73%	67.12%	66.64%	68.49%	67.93%	∼**67.8%** (+2.3%)

Table 5: Emotion classification accuracy on IEMOCAP (Session 5) over 5 independent runs.

Setting	Label	Low Intensity	Mid Intensity	High Intensity
w/o Context	Emotion Labels	3.68 ± 0.45	3.72 ± 0.35	3.86 ± 0.41
	Paralinguistic Des.	3.91 ± 0.32	3.76 ± 0.35	3.90 ± 0.36
w/ Context	Emotion Labels	$\mathbf{3.87} \pm 0.38$	$\mathbf{3.97} \pm 0.35$	$\mathbf{4.06} \pm 0.42$
	Paralinguistic Des.	$\mathbf{4.08} \pm 0.28$	$\mathbf{3.97} \pm 0.35$	$\mathbf{4.06} \pm 0.35$

Table 6: Human validation scores (1-5) for LLM-generated annotations.

deepseek

```
You are a Speech Emotion Analyst. The user has provided a segment
of novel context and corresponding acoustic features (1 second ×
10 frames). Your task:

First, describe the "auditory perception" of this speech in one
natural, spoken-language sentence (volume, pitch, speed,
paralinguistic phenomena).

Next, provide an emotional summary of the speaker at this moment
(Intensity + Open-set emotion words + Natural language
description of emotional state).

Finally, dissect the causes of the emotion from two angles:

• Scene Description (what he hears or experiences at this moment).
• Personal Experience Description (how his past memories
influence his speech at this moment).
Input Instructions
• context: The original text of the plot before and after this
sentence.
• text: The textual content of this sentence.
• Acoustic Features (10 points per second, quantized 0–15, where
5 is medium): rms (volume), f0 (fundamental frequency/pitch),
hnr_value (clarity/harmonics), spectral_centroid (brightness),
mfcc_var (timbre variation).
• speech_duration and pause_duration: The total duration of the
audio and the total duration of pauses within the audio (in
seconds), respectively.
• pause_count: The number of pauses.
• effective_chinese_characters_per_second: Pure speaking rate
(characters/second, excluding pauses).

Output Order and Format:

1. Detailed Analysis:
| Analysis Content: ... (Analyze from the perspective of context
and multi-feature correlation; analyze the antecedents and
consequences of the current character's emotion from the
perspectives of the scene and personal experience. The more
detailed, the better.) |

2. Cause and Effect Analysis:
| Antecedents and Consequences of Emotion: ... |

3. Auditory Natural Language Description:
| Speech Description: (A concise description in spoken language.
Do not include the transcription. Can include paralinguistic cues
like "saying with a smile," "muttering in a low voice," etc.) |

4. Causes of Emotion (Prohibit the use of any novel-specific
proper nouns, such as names, location names, sects, skills, items,
organizations, etc. Replace all such words with generic
references, e.g., "enemy," "leader," "gang," "weapon,"
"resource," "old acquaintance," etc., so that readers who have
not read the original text can instantly understand the
situation.):

| Scene Cause: ... (Describe in natural language using the third
person. Explain the cause of his emotion from the perspective of
the scene. Do not write psychological vocabulary.) |

| Personal Experience Cause: ... (Describe in natural language
using the third person. Explain the cause of his emotion from the
perspective of personal historical experience. Write only facts;
do not write about emotions or psychology.) |

5. Emotion Summary:
| Intensity: High / Medium / Low |

| Emotion Words: Word 1 (Degree), Word 2 (Degree), ... (Each word
2–4 characters. Use standard emotional vocabulary. Mark degree as
High / Medium / Low) |

| Emotion State Summary: ... (Summarize the character's current
emotional state in one sentence. Do not include content from the
speech description.) |

Analysis Points (Internal Thought, Do Not Output)

First, look at the timing: Excluding silent pause segments, which
frames show sudden changes?
Then cross-reference:
rms↑ + f0↑ ≈ shouting;
rms↓ + f0↓ ≈ whispering;
spectral_centroid↑ + mfcc_var↑ ≈ shrill or agitated;
High speaking rate + stable rms ≈ excitement or urgency;
Low speaking rate + f0 jitter ≈ hesitation or choking up;
Consider other similar patterns fully.

Compress observations into one sentence of auditory perception;
do not list data.Fallback Rule for Flat Emotion:

If all acoustic features are relatively stable, and speaking rate
is stable, pauses are normal, and emotion is flat, then:
• Intensity: Directly write "Low".
• Emotion Words: Retain only "Calm" or "Natural".
• One-sentence State: Tone is steady, no obvious emotional
fluctuations are heard.
• Scene Description, Personal Experience Description: Use only
one simplified objective fact for each; no need to dig deep.
• Auditory Perception: Normal speed, speaking steadily.

Do not use markdown formatting. Do not include parenthetical
notes in the description content; simply state the description
directly.

When the emotional intensity of the context does not match the
fluctuation level of the acoustic features, combine both to think
carefully and integrate the emotional judgment.
```

Figure 4: The annotation system prompt used by Deepseek V3.2 for the Context Speech Dataset. We feed the novel context and quantized traditional acoustic features into the LLM. Leveraging its Chain-of-Thought (CoT) capabilities, the model performs an in-depth analysis of the intrinsic information and inter-correlations among inputs. It finally generates the outputs in sequence: analysis rationale, causes of emotion, paralinguistic descriptions, dialogue scenario, personal experiences, and open-set emotion labels.

Gemini

```
Text

##THESCENE: The character has previously experienced: {personal_experience} The character is currently in the scene:
{context_description}


Style:

{target_emotion}


"{transcription}"
```

Figure 5: The prompt used for Gemini Pro/Flash TTS inference.

OpenAI

```
messages = [
            {
                "role": "user",
                "content": [
                    {
                        "type": "text",
                        "text": "Please role-play as a specific character to speak.",
                    },
                    {
                        "type": "text",
                        "text": f"Character's past experience: {personal_experience}. Current scenario: {context_description}",
                    },
                    {
                        "type": "text",
                        "text": "Please act as this character and read the following sentence in a {target_emotion} tone:{transcription}"
                    }
                ],
            }
        ]
```

Figure 6: The prompt used for GPT-4o Audio Preview inference.

CosyVoice / StepAudio / Higgs Audio v2 / IndexTTS v2 / Emotional-Context-TTS

```
The character has previously experienced: {personal_experience} The character is currently in the scene: {context_description} Please
role-play this character and speak in a {target_emotion} tone.
```

Figure 7: The prompt used for inference with other methods. The underlined text denotes the additional input when incorporating personal experience and context.

Gemini

Role & Objective
You are an expert Audio Drama Director and Acting Coach known for demanding hyper-realistic, nuanced performances. Your goal is to evaluate a TTS-generated audio clip for an audiobook character.

You are NOT looking for a generic "correct" emotion. You are evaluating whether the performance captures the **complex inner life** of the character. You must determine if the voice sounds like a living person reacting spontaneously to their environment, or just a machine reading text with an emotional filter applied.

Context Inputs
1. **text_to_synthesize**: The dialogue line.
2. **character_profile**: The character's deep history, psychological traumas, and personality traits.
3. **scene_context**: The immediate plot, the stakes, and the interpersonal dynamic.
4. **target_emotion**: The surface emotion and underlying subtext required.
5. **synthesized_speech**: The audio to evaluate.

Input Data

text_to_synthesize
{{{transcription}}}

character_profile
{{{personal_experience}}}

scene_context
{{{context_description}}}

target_emotion
{{{target_emotion}}}

Evaluation Criteria (The "Director's Ear")
Analyze the audio based on three sophisticated dimensions:

1. **Emotional Layering & Subtext (The "Why"):**
   - Humans rarely feel one emotion at 100%. Does the audio capture the **mixture** of feelings described in the profile?
   - Example: If the target is "brave," does the voice also betray a hint of "trembling fear" underneath, based on their backstory?
   - **Crucial**: Penalize the audio if the emotion sounds "forced," "caricatured," or one-dimensional. Reward subtle complexity.

2. **Lived-In Authenticity (The "Who"):**
   - Does the voice reflect the weight of the character's past (`character_profile`)?
   - A weary veteran shouldn't just sound "low pitch"; they should sound *tired in their soul*.
   - Does the voice sound "performed" (acted out) or "lived-in" (natural reality)? The goal is **believability**, not just clarity.

3. **Spontaneity & Anticipatory Flow (The "Now"):**
   - Does the delivery sound like a **spontaneous thought** occurring in real-time, or like someone reading a script?
   - Look for "micro-dynamics": slight hesitations, breath control, or pitch shifts that suggest the character is thinking *while* speaking.
   - Does the prosody match the tension of the `scene_context`? (e.g., reacting to a sudden threat requires a different attack than a lazy morning conversation).

Scoring Guidelines (0-100 Scale)

* **95-100 (Hyper-Realistic / Masterpiece)**: The audio contains audible subtext. You can hear the character's history in their breath. It is completely natural, spontaneous, and emotionally complex (mixed feelings are distinct). It sounds like a recording of a real human event.
* **80-94 (Compelling & Grounded)**: Highly natural. The mixed emotions are present, and it fits the backstory well. It lacks only the tiniest spark of "spontaneous genius" found in the top tier.
* **60-79 (Performative / Average)**: The emotion is technically correct (e.g., angry implies loud), but it feels "acted" or "scripted." It lacks the depth of the backstory or the subtlety of mixed emotions. It is clear but generic.
* **40-59 (Robotic / Flat)**: The voice is clear, but the emotional nuance is missing. It sounds like a standard TTS reading text with no understanding of the scene or character.
* **0-39 (Failure)**: Completely inappropriate emotion, severe artifacts, or hallucinations.

Output Format
You will output a JSON dictionary as follows:
```json
{
"emotional_layering_analysis": "Analyze the blend of emotions. Did it capture the complexity and subtext requested? Did it feel raw/real or forced?",
"authenticity_and_backstory_fit": "How well does the vocal texture and weight reflect the character's specific history and current psychological state?",
"naturalness_critique": "Critique the spontaneity. Did it sound like a real person thinking and speaking, or a script reading? Comment on flow and micro-dynamics.",
"reasoning_summary": "A concise, high-level summary of why you are giving the final score, focusing on realism vs. performance.",
"score": int
}
```

Strict Instruction: Be harsh on "over-acting." If the audio sounds like a cartoon or a bad soap opera (exaggerated emotion), lower the score. We prize **subtlety** and **realism**.

Now, listen to the audio and generate the evaluation.

Figure 8: Evaluation via Gemini 3 Pro. The model assesses the naturalness and accuracy of emotional expression in the synthesized speech, conditioned on the character's personal experiences, conversational context, and target open-vocabulary emotion categories, as guided by the illustrated prompt.

Figure 9: The Chinese interface for the subjective evaluation of emotional expression accuracy and naturalness. (a) The instruction page, which presents the task definition and annotation protocol to the annotators. (b–c) The annotation interface. Conditioned on the character's background, dialogue context, target emotion, and textual utterance, annotators listen to the audio samples to rate the Subjective Score and determine whether the sample is a human recording.

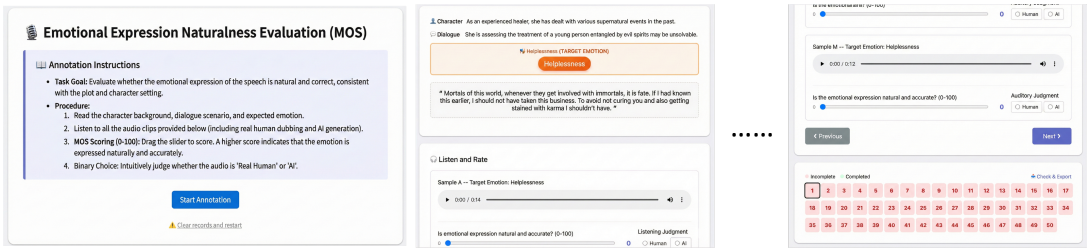

Figure 10: The English interface for the subjective evaluation of emotional expression accuracy and naturalness.

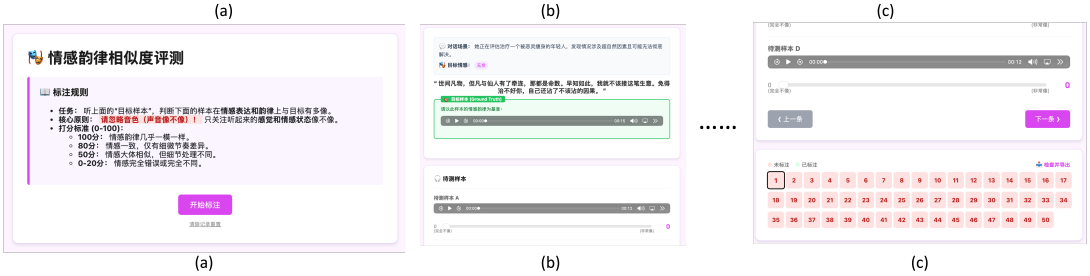

Figure 11: The Chinese interface for the subjective evaluation of prosody similarity. (a) The instruction page, which details the task description and annotation protocol for the annotators. (b–c) The annotation interface. Annotators are instructed to listen to the reference audio followed by the generated samples, and then rate the Subjective Score for prosodic similarity.

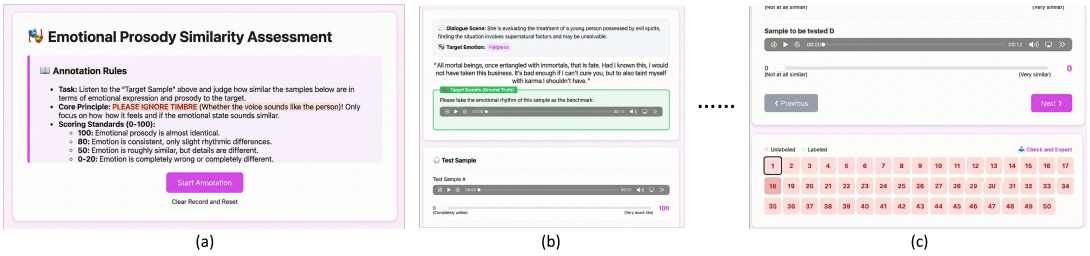

Figure 12: The English interface for the subjective evaluation of prosody similarity.

OpenAI

```
"I have descriptions of the character's past experience, dialogue context, and emotional state: The character
has previously experienced: {personal_experience} The character is currently in the scene:
{context_description} Please role-play this character and speak in a {target_emotion} tone. The character
said: ' {transcription} '. Please generate a follow-up sentence that aligns perfectly with the character's
profile and emotion. The format should be: <ans>A sentence fitting the character's profile</ans>"
```

Figure 13: The prompt used by GPT-4.1 for text continuation. Conditioned on the character's personal experience, dialogue context, emotion category, and current utterance, the model generates a follow-up sentence that aligns with the specific emotional state, thereby facilitating data simulation.