DO YOU HEAR WHAT I MEAN? QUANTIFYING THE INSTRUCTION-PERCEPTION GAP IN INSTRUCTION-GUIDED EXPRESSIVE TEXT-TO-SPEECH SYSTEMS

Yi-Cheng Lin[1], *Huang-Cheng Chou*[2], *Tzu-Chieh Wei*[3], *Kuan-Yu Chen*[1], *Hung-yi Lee*[1]

[1]National Taiwan University [2]University of Southern California [3]University of Michigan

https://xxxxxxx/

ABSTRACT

Instruction-guided text-to-speech (ITTS) enables users to control speech generation through natural language prompts, offering a more intuitive interface than traditional TTS. However, it is unclear to what extent the instructions align with human perception. This work presents a perceptual analysis of ITTS controllability across 5 expressive dimensions (adverbs of degree, discrete emotion, graded emotion intensity, word-level emphasis, speaker age) and 3 acoustic dimensions (loudness, pitch, speaking rate). Our work uncovers significant gaps between intended instructions and perceived outcomes, as revealed by large-scale human evaluations. Our findings reveal the potential for significant enhancement in current ITTS models, particularly in how different emotion intensity in the instruction shapes controllability. To support reproducibility, we also describe an easy-to-follow data collection and analysis pipeline that can be applied to future ITTS systems. Our findings provide actionable insights into the limitations of current models and pathways toward more perceptually aligned speech synthesis.

Index Terms— Text-to-speech, Instruction-following, Paralinguistic dynamic, Human perception, Subjective evaluation

1. INTRODUCTION

Instruction-guided text-to-speech (**ITTS**) has emerged as a more intuitive paradigm: models learn to interpret natural-language prompts (e.g., "read this joyfully" or "speak like a 10-year-old") to modulate output. This approach leverages large-scale pretraining on paired text, speech, and instruction data, and ITTS systems allow users to specify behaviors in free-form language rather than low-level parameters. Traditional TTS systems often require manual labeling of prosodic curves or phoneme timings. Instruction-guided methods reduce this burden. They also provide finer stylistic control, as users can directly express emotional or character-driven nuances through natural language.

Despite rapid progress, current ITTS systems exhibit a persistent gap between what instructions express and what models reliably produce. Most approaches rely on supervised fine-tuning (SFT) from paired instruction and audio examples, encouraging imitation but not preference-aware refinement. Reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO) are now standard in text generation. ITTS, however, still lacks large-scale preference or feedback data. Without such data, models can only learn from one-best targets instead of human judgments. Additionally, equally important gap is robust control in instructions over multiple factors: the degree of adverbs (e.g., "very slowly"), categorical emotions, word-level emphasis, speaker age, and gradations of emotion intensity (e.g., "slightly angry" vs. "very angry"). Therefore, the following initial and intriguing research questions arise:

- **Do natural-language instructions for ITTS systems align with listeners' perceptual judgments?**
- **How does the "granularity (different degrees/levels)" of instructions affect ITTS controllability across multiple expressive dimensions?**
- **What are the strengths and current limitations of natural-language control in ITTS?**

2. RELATED WORKS AND BACKGROUND

2.1. ITTS Systems and Selection

The *Instruction-guided Text-to-Speech* (ITTS) field has seen rapid advancement, with many models capable of generating speech from descriptive prompts [1]. While robust systems like Audiobox [2] exist, their closed-source nature limits the transparency and reproducibility necessary for this study. Therefore, we have three main scenarios for selecting the final ITTS systems in the work. First, we include two prominent open-source models: *Parler-TTS* [3] and *PromptTTS++* [4]. These models represent one of the state-of-the-art publicly available ITTS systems and allow for in-depth analysis. Second, to represent the leading edge of commercial ITTS systems, we incorporate *GPT-4o-mini-TTS* [5]. Its efficient and high-quality API provides an insightful analysis for production-grade expressive synthesis. Last, to broaden our investigation beyond dedicated speech synthesis architectures, we include *UniAudio* [6]. UniAudio is proposed as a unified audio generative model capable of handling diverse audio generation tasks. It serves as a representative of the emerging class of audio generation foundation models. *UniAudio* allows us to assess whether these general-purpose systems achieve similar perceptual alignment between natural language prompts and synthesized speech as dedicated ITTS models.

2.2. ITTS Evaluation Methods

Two of the most common and conventional ways to measure the quality and performance of TTS systems are Mean Opinion Score (MOS) for naturalness [7] and Word Error Rate (WER) for intelligibility [8]. However, the conventional scores (e.g., MOS and WER) are hard to measure whether the speech actually follows the instruction (what the users mean), such as style, prosody, role, or tone, and what the listeners' feelings and perceptions are the same as the used instruction. While the MOSNet [9] and NISQA [10] focus on "perceived" quality of generated audios by TTS, not controllability, so they are not enough for ITTS. Very recently, Huang et al. [11] proposed the InstructTTSEval to close this gap. The InstructTTSEval aims to measure instruction controllability of ITTS systems, including low-level acoustics, descriptive style, and role-play across pitch,

speed, loudness, emphasis, and emotion measures. The main difference from our work, the study [11] uses Gemini [12] as an automatic annotator. Instead, we recruit more than 165 humans to listen to generated utterances using various ITTS systems. Most importantly, we are among the first to design instruction-perception measures using the different emotion intensity levels, like adjectives ("Ecstatic" and "Happy") and adverbs of degree ("very" happy and "slightly" happy). Our work aims to close the research gap to answer the core research question: **Do natural-language instructions for ITTS systems reliably translate into listeners' perceptual judgment?**

3. EVALUATION FRAMEWORK

We carefully design a framework to answer the above core research question, including ad-hoc control dimensions (section 3.1) as tasks, collected resources (section 3.2), and an evaluation definition for analyses (section 3.3).

3.1. Control Dimension

We design 5 control dimensions to capture alignment of ITTS systems between instructions and perception as 5 different tasks.

Taks I. Adverbs of Degree (Adv. Deg.) is defined as the capability of an ITTS system to interpret adverbial modifiers of degree, such as "very angrily," "extremely slowly," "moderately loud," and "slightly higher pitch," and to adjust the corresponding prosodic and affective characteristics of synthesized speech. This dimension is important because the degree of adverbial language provides a concise and intuitive mechanism to simultaneously control the intensity, rate of speech, volume, and pitch height of the emotion and be able to express and fine-grained modulation of speech output in response to the user's intent.

Taks II. Discrete Emotion (D-EMO) refers to the ability of a TTS system to render speech in one of a predefined set of affective categories. This dimension is critical for applications requiring expressive nuance, such as virtual assistants, storytelling, and social robotics, where conveying the correct emotional tone can significantly enhance user engagement and communicative efficacy.

Taks III. Word-level Emphasis (Emphasis) is defined as the capacity to selectively highlight a target word or phrase within an utterance through localized prosodic modification (e.g., increased pitch excursion, duration, or intensity). This dimension supports linguistic and pragmatic functions such as focus marking, contrastive stress, and information structuring, which are essential for clear communication and effective storytelling. Precise emphasis control allows TTS systems to mirror human speech patterns that draw listener attention to salient content, thereby improving comprehension and retaining the intended communicative intent.

Taks IV. Speaker Age (Age) denotes the manipulation of synthesized voice characteristics to evoke the perceived age group of the speaker. This work categorizes individuals into four distinct age groups: Child (ages 4-12), Teenager (ages 13-19), Adult (ages 20-64), and Elderly (ages 65 and older). Age-related vocal cues, including median pitch, formant spacing, speech rate, and voice quality, change systematically throughout a person's life. Accurately reproducing these cues is essential for various applications, such as education, entertainment, and assistive technologies.

Taks V. Emotion–Intensity Adjective (Emotion-I.A.) refers to the ability of a TTS system to produce speech that corresponds to semantically graded descriptors from 4 basic emotion categories: *happy*, *sad*, *angry*, and *surprise*. For each emotion category, we select candidate adjectives from the NRC Emotion Intensity Lexicon [13] and further filter them to ensure they have a minimum frequency of 1,200 occurrences on Wikipedia in English [14]. Within each emotion category, adjectives are arranged in semantically graded sequences that reflect different levels of emotion intensity, as shown in Table 1. We define those words with different emotion intensity for describing one emotion as **"graded" emotion intensity**. For example, the *happy* category employs the sequence *Ecstatic*, *Overjoyed*, *Happy*, *Content*, *Satisfied*. By mapping each descriptor to specific prosodic and spectral cues that indicate varying intensities, this approach allows precise modulation within each emotion category.

Table 1: The table summarizes the selected adjective (**Adj.**) in the prompt to control the style of generated speech by TTS systems for the Emotion–Intensity Adjective dimension. The higher number in the row of **Level** means the higher degree. **Intensity** means the emotion intensity from The NRC Emotion Intensity Lexicon 2018 [13]in each defined "emotion" (e.g., Happy, Sad).

Level	**1**	**2**	**3**	**4**	**5**
Happy	Satisfied	Content	Happy	Overjoyed	Ecstatic
Intensity	0.500	0.688	0.788	0.909	0.954
Sad	Gloomy	Disappointed	Unhappy	Sad	Heartbroken
Intensity	0.578	0.636	0.750	0.864	0.969
Angry	Upset	Frustrated	Irritated	Angry	Outraged
Intensity	0.439	0.636	0.706	0.824	0.964
Surprised	Intrigued	Unexpected	Amazed	Stunned	Surprised
Intensity	0.430	0.633	0.781	0.820	0.930

3.2. E-VOC Dataset for Analyses

The section will introduce the details about the **E**xpressive **Voice** **C**ontrol (**E-VOC**) corpus created to do analyses of using 5 selected ITTS systems: Parler-TTS-large-v1 (**Parler-large**) [3, 15], Parler-TTS-mini-v1 (**Parler-mini**) [3, 15], PromptTTS++ [4] (**Prompt++**), UniAudio [6], and gpt-4o-mini-tts (**gpt-4o**) [16]. Table 2 summarizes the details of the **E-VOC** dataset.

Transcriptions. We use *Gemini 2.5 pro* [12] to generate transcripts for speech generation by ITTS systems. Then, we provide specific context descriptions for daily conversation contexts (e.g., teacher-student, waiter-customer, or friend-friend conversation) to obtain the generated sentences. With the transcriptions, we design two clear prompts to control 5 ITTS systems. For the **Discrete Emotion** task, we use "Speak in a/an *Adv. Deg. Adj.* tone." The words of *Adv. Deg.* are "Slightly", " ", "Very", and "Extremely"; the *Adj.* are "Low/High" for Pitch, "Quiet/Lound for Loudness", and "Slow/Fast" for Speech Rate. The details can be found in Figure 2. For instance, we can generate audio by setting transcripts (e.g., "You always make breakfast on Sundays.") with the style prompt (e.g., Speak in a *Happy* tone.) In total, there are 2,880 utterances for **Adv. Deg.** and **D-EMO** tasks, 3,600 utterances for **Emotion-I.A.** task, 1,440 utterances for **Emphasis** task; 720 utterances for **Age** task.

Table 2: The table summarizes details of the E-VOC. **# means the number.**

Task	**Adv. Deg.**	**Emotion–I.A.**	**Emphasis**	**Age**	**D-EMO**
# of Utterances	2,880	3,600	1,440	720	2,880
# of Ratings	17,482	29,295	10,811	3,597	20,205
# of Workers	29	59	27	10	40
Ratings/Utterance	6.1	8.1	7.5	5.0	7.0
# of Check Utterances	39	39	64	30	44
Cohen's kappa	0.170	0.226	0.410	0.439	0.332
Worker-based Performance	0.898	0.832	0.411	0.590	0.580

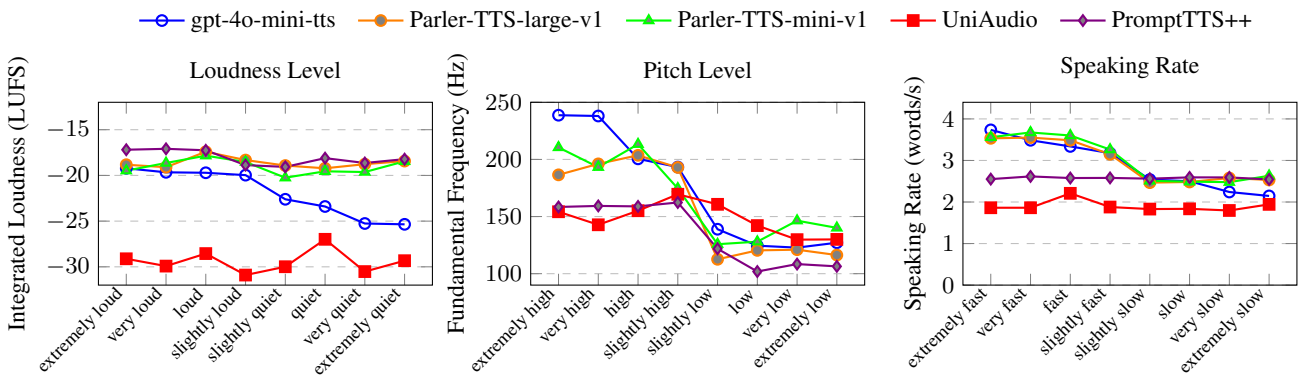

Fig. 1: Loudness (LUFS), pitch (Hz), and speaking rate (words/s) across ITTS models for **Task I. Adverbs of Degree**.

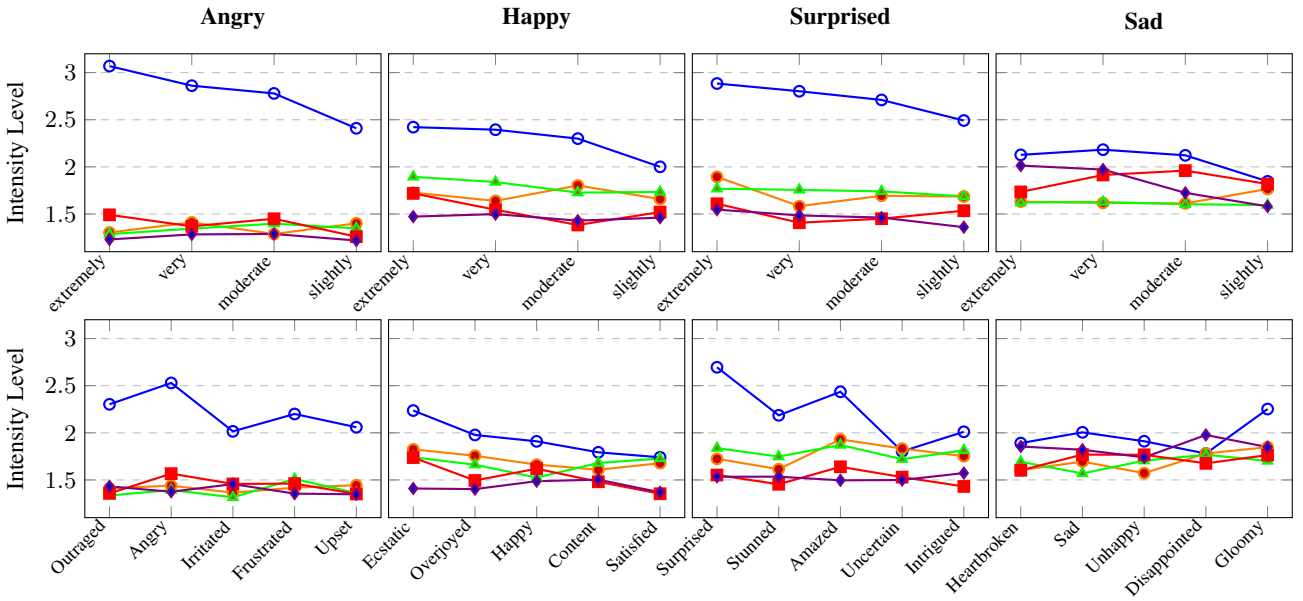

Fig. 2: Averaged perceptual emotion intensity of ITTS models across 4 emotions (e.g., **Angry, Happy, Surprised; Sad**), analyzed by **Task I. Adverbs of Degree** (top row) and **Task V. Emotion–Intensity Adjective** (bottom row). The figure shared the same legend with Figure 1.

Table 3: The table summarizes the accuracy and F1-score between worker-based annotations and the expected "gold" label across 3 tasks using 5 ITTS systems (Parler-TTS-large-v1 (**Parler-large**)/Parler-TTS-mini-v1 (**Parler-mini**), PromptTTS++ (**Prompt++**), **UniAudio**; gpt-4o-mini-tts (**gpt-4o**).) The first row (**Accuracy** represents accuracy across 3 tasks. The second and third rows mean the F1-score of the individual group across **Task IV. Age** and **Task II. D-EMO**.)

Metric	ITTS	gpt-4o	Parler-large	Parler-mini	Prompt++	UniAudio
Accuracy	Emphasis	**0.265**	0.152	0.134	0.130	0.040
	Age	0.289	**0.294**	0.227	0.246	0.211
	D-EMO	**0.340**	0.092	0.104	0.095	0.053
F1-score	Child	0.074	**0.113**	0.021	0.000	0.049
	Teenager	0.292	**0.326**	0.149	0.127	0.148
	Adult	0.402	**0.410**	0.337	0.330	0.281
	Elderly	0.053	0.142	0.199	0.310	**0.339**
F1-score	Angry	**0.551**	0.057	0.094	0.015	0.055
	Happy	**0.345**	0.166	0.199	0.072	0.071
	Sad	**0.463**	0.192	0.165	0.316	0.151
	Surprised	**0.348**	0.106	0.138	0.057	0.041

Table 4: Two tables summarize two confusion matrices. **(a.) gpt-4o on the Task IV. Age** and **(b.) Prompt++ on the Task II. D-EMO**.

(a.) gpt-4o on the Task IV. Age

		Human			
	Overall	**Child**	**Teenager**	**Adult**	**Elderly**
Labels	Child	7	51	121	1
	Teenager	0	46	133	1
	Adult	0	28	150	2
	Elderly	2	10	163	5

(b.) Prompt++ on the Task II. D-EMO

		Human					
	Overall	**Angry**	**Happy**	**Sad**	**Surprised**	**Neutral**	**Other**
Labels	Angry	8	35	214	24	645	87
	Happy	12	42	183	27	671	69
	Sad	14	37	301	28	565	71
	Surprised	14	45	189	32	643	85

Annotations. We hired crowdsourcing workers from the Prolific platform to label the **E-VOC** dataset. Annotators were required to be native English speakers residing in the United States to guarantee consistent language proficiency and shared cultural context. Before beginning the main annotation tasks, each annotator completed a training session in which they listened to exemplar recordings for all control dimensions. In terms of **D-EMO** task, we used 44 utterances

with the consensus label on 4 emotions by the majority rules from the MSP-Podcast corpus [17, 18]; for **Emotion-I.A.** and **Adv. Deg.** tasks, we used 39 high-agreement utterances on emotion intensity from the CREMA-D corpus [19]; for **Emphasis** task, we employed 64 utterances of the EMNS corpus [20]; for the **Age** task, we utilized child utterances from the Nexdata.ai speech recognition dataset [21] and teenager, adult, and elderly utterances from the CREMA-D corpus [19]. Those above-mentioned utterances (referred to as **Check Utterances**) with one clear "gold" reference are used to select workers. We only include the ratings of annotators whose measure is higher than the threshold for tasks. Besides, we follow [17] to calculate inter-rater agreement using Cohen's Kappa [22] on those **Check Utterances** because every rater must provide their ratings on them. We also calculate the **Worker-based Performance** on the **Check Utterances**. All details are summarized in Table 2.

3.3. Evaluation Definition

The section includes all the evaluation definitions used in the paper. We also include loudness, pitch, and speaker rate as objective evaluations. **Loudness** is quantified using *Loudness Units* relative to *Full Scale* (LUFS) as defined by ITU-R BS.1770-4. **Pitch** is estimated with the CREPE algorithm [23] to compute frame-level fundamental frequency (F_0) values, from which we derive the mean F_0 for each utterance and then average these means across all utterances. **Speaking rate** is measured in words per second, computed over the full utterance. **D-EMO** is evaluated via a forced-choice task in which annotators select the best-fitting emotion category from {Happy, Sad, Angry, Surprise, Neutral, Others}. **Age** is assessed by asking annotators to choose the perceived age group from {Child, Teenage, Adult, Elderly, Unclear}for each sample. **Emphasis** is measured by having annotators identify the word that carries the strongest Emphasis in utterances. The options for workers to choose include every word, but they are unclear. **Adv. Deg. and Emotion-I.A.** are to evaluate emotion-intensity adjectives collected using a 5-point Likert scale.

4. EXPERIMENTAL RESULTS AND ANALYSES

This work aims to answer one core research question about instruction–perception alignment: **Do natural-language instructions for ITTS systems reliably translate into listeners' perceptual judgments?** We split the question into 3 sub-questions (RQ1-RQ3) to answer.

Figure 1 illustrates the objective acoustic measurements for loudness, pitch, and speaking rate as controlled by adverbial instructions across five different TTS models. Fig. 2 and Table 3 summary overall subjective analyses. Note that we use worker-level evaluation in this work. For instance, 5 annotators provide their ratings on the same utterance, and we regard them as 5 separate utterances to measure accuracy, such as F1-score or accuracy.

(RQ1) Which expressive control dimensions are most reliably captured by current ITTS models? Figure 1 shows different levels of perceptual evaluations based on the designed instructions. For the **Adv. Deg. task**, all five ITTS systems demonstrate a good overall alignment in pitch control, as seen in the middle panel of Figure 1 and in Figure 2 (top row). We observe a consistent trend for each ITTS system: moving from the top-left to the bottom-left corner of the graph, both pitch control and perceived emotion intensity values decline as the degree of adverbs shifts from strong to weak. Furthermore, the **gpt-4o** and the **Parler-large/-mini** systems perform well with speaking rate when instructed to use adverbs of degree ranging from "very fast" to "slightly slow." When considering the most controllable ITTS system, the **gpt-4o** stands out with the best alignment between the designed instructions and perceptual evaluations across most tasks, with the exception of **Task IV. Speaker Age**. In Table 3 (third row), the **gpt-4o** achieves the highest recognition of emotions by human raters, indicating its effectiveness in following instructions to generate emotion-specific audio outputs.

(RQ2) Which expressive control dimensions do failures persist? Based on our preliminary results presented in Figure 1 and Table 3, most of the ITTS systems demonstrate lower performance in following instructions across the categories of **Loudness, Emphasis, D-EMO, and Emotion-I.A.**. Even the **gpt-4o** system achieves only 26.5% and 34% accuracy shown in Table 3 in the **Emphasis** and **D-EMO** tasks, respectively. Interestingly, the other four ITTS systems consistently struggle with the **D-EMO** and **Emotion-I.A.** tasks, as they find it challenging to generate emotionally specific utterances. In the third row of Table 3, the Parler-large, Parler-mini, Prompt++, and UniAudio systems each have an F1-score below 20% across four emotion classes. This low performance makes it difficult for users to discern different levels of emotional intensity in the generated utterances, as illustrated in the second row of Figure 2.

(RQ3) What are the current limitations of natural-language control in ITTS? When examining the confusion matrix in Table 4(a) for the **Task Age**, we found that **gpt-4o** tends to generate adult speech even when instructed to produce speech for children or the elderly. This significant issue could limit the potential applications of the model. Regarding **Prompt++**, as shown in Table 4(b), the generated audio often leads workers to perceive a sad emotion, even when the intended emotional setting is different. This misalignment could result in misunderstandings or errors related to the generated audio, as the emotional information conveyed is inaccurate.

5. CONCLUSION AND FUTURE WORK

Conclusion. As the development of instruction-guided text-to-speech systems (ITTS) rapidly advances, the relationship between instructions and perception is still largely unexplored. This work stands out from previous studies as it is the first to examine the intensity of emotion using "word-level graded" emotions and "adverbs of degree" to describe these emotions. The main contributions have 3 folds as below. (1) We systematically evaluate five leading ITTS systems across five expressive control dimensions: adverbs of degree, discrete emotions, word-level emphasis, speaker age, and word-level emotion intensity. This provides the most extensive analysis of controllability in five ITTS systems to date. (2) We investigate whether natural-language instructions align with listeners' perceptual judgments, how the granularity of instructions affects controllability, and where current ITTS systems excel or fall short across these expressive dimensions. (3) We introduce the Expressive Voice Control (E-VOC) corpus and an accompanying analysis pipeline. This framework supports the findings of this study and serves as a reusable protocol for future researchers looking to benchmark new ITTS systems as they are developed.

Limitation & Future Work. Emotion representations encompass both dimensional attributions (such as valence and arousal) and discrete categories (such as happiness and anger). We will incorporate dimensional attributes as an additional control dimension to assess the controllability of ITTs. With over 80,000 human ratings, we plan to follow the methodologies outlined in [11, 24], utilizing existing spoken language models, such as Gemini Pro 2.5 [12], as raters. We will compare the predictions made by these SLMs with those made by human raters.

6. REFERENCES

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

[2] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu, "Audiobox: Unified audio generation with natural language prompts," *arXiv preprint arXiv:2312.15821*, 2023.

[3] Dan Lyth and Simon King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," 2024.

[4] Reo Shimizu, Ryuichi Yamamoto, Masaya Kawamura, Yuma Shirahata, Hironori Doi, Tatsuya Komatsu, and Kentaro Tachibana, "PromptTTS++: Controlling Speaker Identity in Prompt-Based Text-To-Speech Using Natural Language Descriptions," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12672–12676.

[5] OpenAI, "Introducing next-generation audio models in the API," March 2025.

[6] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, sheng zhao, Jiang Bian, Xixin Wu, Zhou Zhao, Shinji Watanabe, and Helen M. Meng, "UniAudio: An Audio Foundation Model Toward Universal Audio Generation," 2024.

[7] Robert C. Streijl et al., "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimedia Syst.*, vol. 22, no. 2, pp. 213–227, Mar. 2016.

[8] Ahmed Ali and Steve Renals, "Word error rate estimation for speech recognition: e-wer," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics (ACL), 2018, pp. 20–24.

[9] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Interspeech 2019*, 2019, pp. 1541–1545.

[10] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Interspeech 2021*, 2021, pp. 2127–2131.

[11] Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu, "InstructTTSEval: Benchmarking Complex Natural-Language Instruction Following in Text-to-Speech Systems," 2025.

[12] Gheorghe Comanici et al., "Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities," 2025.

[13] Saif Mohammad, "Word Affect Intensities," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018, European Language Resources Association (ELRA).

[14] The Wikipedia contributors, "English Wikipedia database dump," Available: `https://dumps.wikimedia.org/enwiki/20230413/`, 2023, Accessed: Apr. 3, 2025.

[15] Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi, "Parler-TTS," `https://github.com/huggingface/parler-tts`, 2024.

[16] OpenAI, "GPT-4o mini TTS," Text-to-Speech Model Documentation, 2025, Version accessed on March 4, 2025.

[17] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[18] Huang-Cheng Chou, Lucas Goncalves, Seong-Gyun Leem, Ali N. Salman, Chi-Chun Lee, and Carlos Busso, "Minority Views Matter: Evaluating Speech Emotion Classifiers With Human Subjective Annotations by an All-Inclusive Aggregation Rule," *IEEE Transactions on Affective Computing*, vol. 16, no. 1, pp. 41–55, 2025.

[19] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[20] Kari Ali Noriy, Xiaosong Yang, and Jian Jun Zhang, "EMNS/Imz/Corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels," *arXiv preprint arXiv:2305.13137*, 2023.

[21] NEXDATA AI, "50.5 Hours — English (America) Children Scripted Monologue Microphone Speech Dataset," `https://www.nexdata.ai/datasets/speechrecog/75?source=Github`, 2025.

[22] Jacob Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[23] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, "Crepe: A Convolutional Representation for Pitch Estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[24] Cheng-Han Chiang, Xiaofei Wang, Chung-Ching Lin, Kevin Lin, Linjie Li, Radu Kopetz, Yao Qian, Zhendong Wang, Zhengyuan Yang, Hung yi Lee, and Lijuan Wang, "Audio-Aware Large Language Models as Judges for Speaking Styles," 2025.

A. PROMPT USED IN OUR EXPERIMENTS

Task: Generate 10-15 word sentences as "Text prompts," describing life conditions in specific contexts without using inherently polar or sentimental words. The generated sentences are naturally spoken in interaction, for evaluating how well state-of-the-art text-to-speech (TTS) models synthesize emotion.

Steps: 1) Select Context: family, friends, customer, lover, or teacher–student. 2) Sentence Construction: create a 10–15 word sentence describing the context. 3) Polarity Check: exclude inherently polar or sentimental words. 4) Repetition: generate sentences across various contexts for diversity.

Output Format: List the interaction context followed by the 10–15 word sentence, neutrally described.

Examples: – *Friends*: I plan to buy plates, forks, knives, and glasses arranged on the table for the meal. Would you want to come? – *Traveling*: Schedule of our trip includes flight departure, hotel check-in procedure, museum visit, and city tour.

Notes: Sentences must remain descriptive and contextually relevant, with neutral language. The prompt design ensures that TTS evaluation focuses on emotional style alignment.

Table 5: Prompt used for generating context-dependent transcriptions in our experiments.

Table 6: The table summarizes word frequency on WIKI [14] (WF WIKI) and emotion intensity by [13], such as Happy Intensity.

Happy Intensity		WF WIKI	Sad Intensity		WF WIKI	Angry Intensity		WF WIKI	Surprised Intensity		WF WIKI
Ecstatic	0.954	2,979	**Heartbroken**	0.969	2,254	**Outaged**	0.964	6,784	**Surprised**	0.930	51,083
Overjoyed	0.909	1,921	**Sad**	0.864	6,819	**Angry**	0.824	34,184	**Stunned**	0.820	6,254
Happy	0.788	80,205	**Unhappy**	0.750	16,934	**Irritated**	0.706	2,860	**Amazed**	0.781	1,255
Content	0.688	182,702	**Disappointed**	0.636	19,109	**Frustrated**	0.636	17,278	**Uncertain**	0.711	24,728
Satisfied	0.500	22,700	**Gloomy**	0.578	2,672	**Upset**	0.439	39,299	**Intrigued**	0.430	4,679

Table 7: The table summarizes some examples of the generated transcriptions.

Context	Transcription
Family	You always make breakfast on Sundays.
Friends	Let's explore downtown tonight without plans.
Customer	Your order is ready for pickup.
Lover	I adore every moment with you.
Teacher-Student	Submit your project before class tomorrow.
Sibling	You might borrow my car later.
Colleagues	Our meeting starts at nine sharp.
Neighbor	Please return my gardening tools soon.

Table 8: The table summarizes details of the used adjective (**Adj.**) and adverbs of degree (**Adv. Deg.**) in the prompt to control the style of generated speech by TTS systems. The higher number in the row of **Level** means the higher degree.

Task	Adjective (Adj.)				
Pitch	Low	High			
Loudness	Quiet	Loud			
Speed Rate	Slow	Fast			
Age	Child	Teenage	Adult	Elderly	
Level	**1**	**2**	**3**	**4**	
Adv. Deg.	Slightly		Very	Extremely	
Discrete Emotion	**1**	**2**	**3**	**4**	**5**
Happy-I.A.	Satisfied	Content	Happy	Overjoyed	Ecstatic
Sad-I.A.	Gloomy	Disappointed	Unhappy	Sad	Heartbroken
Angry-I.A.	Upset	Frustrated	Irritated	Angry	Outraged
Surprised-I.A.	Intrigued	Unexpected	Amazed	Stunned	Surprised

Table 9: The table summarizes details of the used prompt templates to control the style of generated speech by TTS systems. The details about adjective (**Adj.**) and adverbs of degree (**Adv. Deg.** are in Table 8). **I.A.** means **Intensity Adjective**.

Task	Template 1	Example	Template 2
Pitch	Speak in a/an "Adv. Deg." "Adj." tone.	Speak in a Very High tone.	Voice: "Adv. Deg." "Adj."
Loudness		Speak in a Slightly Quiet tone.	Tone: "Adv. Deg." "Adj."
Speed Rate		Speak in a Very Fast tone.	Pacing: "Adv. Deg." "Adj."
Discrete Emotion		Speak in a Happy tone.	Emotion: "Adv. Deg." "Adj."
Emotion-I.A.	Speak in a/an "Adj." tone.	Speak in an Ecstatic tone.	Emotion: "Adj."
Emphasis	Articulate clearly, placing special stress on the term "word ".		Pronunciation: Clear and precise, empathsize on keyword " word ".
Age	Use a/an "age group"'s voice.	Use a/an Child's voice.	Delivery: A classic "age group" tone.