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

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, the alignment between user style instructions and listener perception remains largely unexplored. This work first presents a perceptual analysis of ITTS controllability across two expressive dimensions (adverbs of degree and graded emotion intensity) and collects human ratings on speaker age and word-level emphasis attributes. To comprehensively reveal the instruction-perception gap, we provide a data collection with large-scale human evaluations, named **E**xpressive **VO**ice **C**ontrol (E-VOC) corpus. Furthermore, we reveal that (1) gpt-4o-mini-tts is the most reliable ITTS model with great alignment between instruction and generated utterances across acoustic dimensions. (2) The 5 analyzed ITTS systems tend to generate Adult voices even when the instructions ask to use child or elderly voices. (3) Fine-grained control remains a major challenge, indicating that most ITTS systems have substantial room for improvement in interpreting slightly different attribute instructions.

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

1. INTRODUCTION

Instruction-guided text-to-speech (ITTS) [1, 2] enables users to steer speech synthesis using natural-language prompts (e.g., "read this joyfully," "speak like a child"). This approach offers a transparent and flexible alternative to conventional TTS pipelines [3, 4] that often require low-level acoustic controls or specialized labels for prosody and timing. By shifting control to free-form language, ITTS promises to enhance accessibility for content creation, assistive technologies, education, and interactive media.

Reliable evaluation is essential for deploying ITTS systems in practical applications. Traditional metrics like Mean Opinion Score (MOS) [5–7] assess naturalness and Word Error Rate (WER) [8] measure intelligibility, but these metrics fall short in measuring instructional fidelity, the precise alignment of synthesized speech with fine-grained user prompts. This gap raises a central question: ***Do natural-language instructions for ITTS systems reliably align with listener perceptions, particularly for slightly different attributes like graded emotion intensity?***

To address this question, we introduce a novel evaluation of ITTS controllability. Our study is the first to incorporate adverbs of degree (e.g., slightly, extremely) and graded emotion intensity (e.g., ecstatic, happy) as explicit evaluation dimensions. We also present the first large-scale collection of human perceptual ratings for speaker age and word-level emphasis. To systematically examine the gap between instructions and listener perception across these dimensions, we developed a new analysis framework and compiled the **E**xpressive **VO**ice **C**ontrol (E-VOC) corpus[1], consisting of high-quality judgments from over 165 human raters. To ensure the reliability of our findings and enable reproducibility, all data were gathered through a quality-controlled process, and we will publicly release both the corpus and the analysis pipeline.

2. RELATED WORKS AND BACKGROUND

2.1. ITTS Systems and Selection

The field of *Instruction-guided Text-to-Speech* (ITTS) has seen rapid advancement, with many models capable of generating speech from descriptive prompts [1]. Although robust systems such as Audiobox [9] exist, their closed-source nature limits the transparency and reproducibility necessary for this study. Therefore, to ensure a comprehensive and replicable analysis, we selected five representative models across three distinct categories. First, to represent the state-of-the-art in open-source research, we included *Parler-TTS* [10] and *PromptTTS++* [11]. These models are publicly available, allowing for the in-depth analysis required for this study. Second, to represent the leading edge of commercial ITTS systems, we incorporate *GPT-4o-mini-TTS* [12]. Its efficient and high-quality API provides an insightful analysis for production-grade expressive synthesis. Finally, to test the capabilities of non-specialized systems, we included *UniAudio* [13], a unified audio generative model. Its inclusion allows us to assess whether a general-purpose audio foundation model can achieve perceptual alignment comparable to others.

2.2. ITTS Evaluation Methods

Prior work has established evaluation methodologies that measure controllability and instruction alignment to evaluate the performance of ITTS systems. These approaches can be broadly grouped into three main categories.

(1) Attribute-based objective measures. Several studies evaluate ITTS by classifying acoustic or stylistic attributes of the generated speech and comparing them to the prompt. For example, PromptTTS and PromptTTS 2 [10, 11] measured accuracy in controlling gender, pitch, speed, and volume. Building on this idea, our work also includes objective analyses of pitch, speaking rate, and loudness, as they provide interpretable measures of control precision.

(2) Embedding-based similarity measures. Other works adopt embedding models to quantify alignment between prompts and audio outputs. For example, AudioBox [9] proposed using Joint-CLAP to correlate audio-text embeddings with human judgments of style relevance, while Emosphere [14] utilized emotion2vec [15] embeddings to evaluate emotion similarity.

(3) Instruction-following perceptual measures. A third line of work directly involves human listeners or automated judges to rate how well synthesized speech matches a textual instruction. This

[1]`Project Website`

category can be divided into two main approaches: (i) *Human-Centered Evaluation*: This approach treats human perception as the ground truth. For example, InstructTTS [13] introduced a Relevance MOS (RMOS) to score overall prompt alignment, while VoxInstruct [16] used a similar Mean Opinion Score for Instruction (MOS-I). Other studies, like EmoVoice [17], have focused more narrowly, collecting listener ratings on specific dimensions such as overall expressiveness. (ii) *Automated Evaluation*: Recent studies have developed automated methods to increase scalability and reduce cost. SpeechCraft [18], for instance, fine-tuned classifiers to predict attributes like speaker age and word-level emphasis, and used the classifiers to measure how well these predictions aligned with the original instructions. Pushing this further, InstructTTSEval [19] benchmarked ITTS systems using Gemini [20], which evaluated alignment across various prompts, from low-level acoustic details to high-level role-play instructions.

While these methods are essential for assessing general alignment, they mostly provide coarse outcomes such as overall relevance scores or discrete category matches (e.g., age or emotion class). In contrast, our work directly measures perceptual controllability along graded and expressive dimensions, including adverbs of degree, fine levels of emotional intensity. Also, prior works on age and emphasis evaluation have only applied automated methods. However, classifier predictions are tied to their training data and may inherit dataset biases, making them unreliable indicators of how listeners actually perceive expressive attributes. In contrast, our study conducts large-scale human evaluations, providing direct perceptual evidence.

3. EVALUATION FRAMEWORK

We designed a comprehensive evaluation framework to investigate the instruction-perception gap in ITTS systematically. This framework consists of 3 core components: the control dimensions that define the evaluation tasks (Section 3.1), the evaluation metrics used to quantify alignment (Section 3.2), and the E-VOC corpus of human perceptual data collected to support the analysis (Section 3.3).

3.1. Control Dimension

To comprehensively evaluate ITTS controllability, we define 4 control dimensions that serve as the tasks in our study. These are grouped into two categories: two novel dimensions designed to measure fine-grained expressivity, and 2 established dimensions that assess fundamental aspects of speech synthesis.

3.1.1. Proposed Control Dimensions

The following two dimensions are our primary contribution to evaluating the fine-grained control capabilities of ITTS systems.

Table 1: Adjectives used in the prompt to control the style of generated speech by ITTS systems for the Emotion–Intensity Adjective dimension. The higher the number in the **Level** row, the higher the degree. Intensity is the emotional intensity of the adjective from [21].

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

Task I. Adverbs of Degree (Adv. Deg.) tests whether models follow degree modifiers such as "slightly," "very," and "extremely" to adjust prosody (loudness, pitch, speaking rate). This dimension is important because adverbial scaling provides users with a simple way to control the fine-grained prosodic variation, essential for storytelling and emotional expression.

Task II. Emotion–Intensity Adjective (Emo-I.A.) evaluates whether ITTS systems can express different degrees of the same emotion, using adjectives that represent progressively stronger intensities. For 4 core emotions (*happy*, *sad*, *angry*; *surprise*), we selected candidate adjectives from the human-annotated NRC Emotion Intensity Lexicon [21]. We filtered the candidates to include words that appear frequently in common use (at least 1,200 times on Wikipedia [22]). These adjectives were then arranged into sequences of increasing intensity (Table 1), such as *Satisfied*, *Content*, *Happy*, *Overjoyed*; *Ecstatic* for the "happy" category. This task evaluates whether models can transform these ordered adjective sequences into corresponding perceptual scales of emotion intensity as judged by human listeners.

3.1.2. Other Control Dimensions

In addition to our proposed dimensions, we include the following established tasks to provide a more holistic evaluation of the capabilities of the ITTS models.

Task III. Speaker Age (Age) evaluates a model's ability to synthesize a voice that reflects a specific perceived age group. We define four distinct categories: Child (ages 4-12), Teenager (ages 13-19), Adult (ages 20-64), and Elderly (ages 65+). Since vocal cues change systematically throughout a person's life, accurately reproducing them is essential for practical applications, such as entertainment and education.

Task IV. Word-level Emphasis (Emphasis) assesses the ability to place prosodic prominence on a specific target word within a sentence using cues like pitch excursion and duration. Precise emphasis control is critical for mirroring natural human speech patterns and allows systems to draw listener attention to important information and preserve the original communicative intent.

3.2. Evaluation Metrics

Our framework combines objective acoustic analysis with subjective human perceptual judgments to create a comprehensive evaluation.

3.2.1. Objective Measures

For the **Adv. Deg.** task, we use objective metrics to quantify changes in the physical properties of the generated speech. **Loudness** is measured in Loudness Units relative to Full Scale (LUFS), following the ITU-R BS.1770-4 standard. **Pitch** is calculated as the mean fundamental frequency (F_0) per utterance, estimated using the CREPE model [23]. **Speaking rate** is measured in words per second across the entire utterance.

Table 2: The table summarizes details of the E-VOC across 4 tasks. # means the number. Cohen's kappa is for inter-rater agreement.

Task	**Adv. Deg.**	**Emo–I.A.**	**Emphasis**	**Age**
# of Utterances	2,880	3,600	1,440	720
# of Ratings	17,482	29,295	10,811	3,597
# of Workers	29	59	27	10
Ratings/Utterance	6.1	8.1	7.5	5.0
# of Check Utterances	39	39	64	30
Cohen's kappa	0.170	0.226	0.410	0.439
Performance	0.898	0.832	0.411	0.590

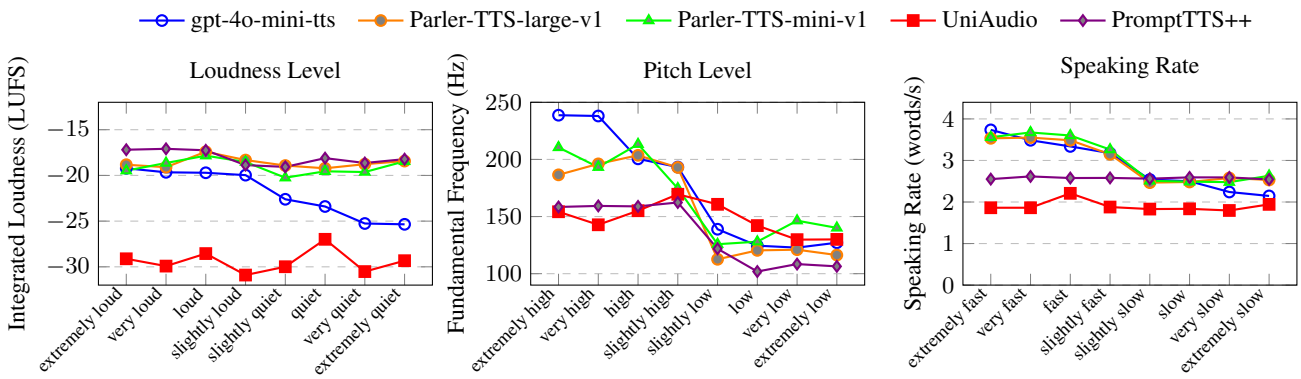

Fig. 1: Loudness (LUFS), pitch (Hz), and speaking rate (words/s) across ITTS models for **Task I. Adverbs of Degree**.

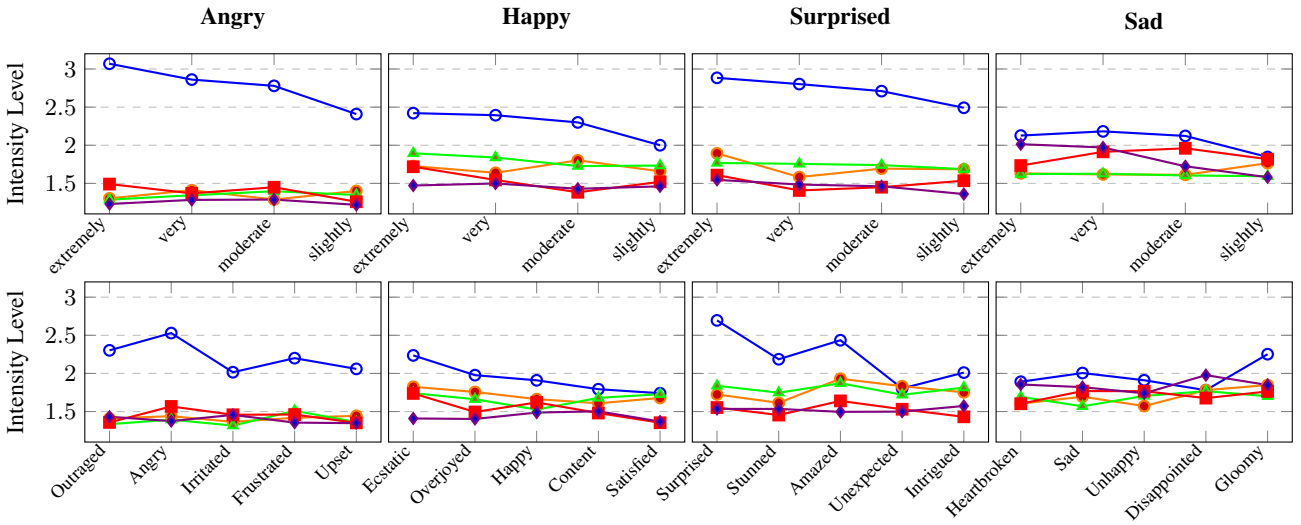

Fig. 2: Averaged perceptual emotion intensity of ITTS models across 4 emotions (e.g., **Angry, Happy, Surprised; Sad**), analyzed by **Task I. Adverbs of Degree** (top row) and **Task II. Emotion–Intensity Adjective** (bottom row). The figure shared the same legend with Figure 1.

3.2.2. Subjective Measures

For dimensions requiring stylistic and semantic interpretation, we rely on human perceptual ratings. **Emotion Intensity**: For both the Adverbs of Degree and Emotion-Intensity Adjectives tasks, listeners rate the perceived intensity of the target emotion on a 5-point Likert scale. **Emphasis**: Listeners use a forced-choice task to identify the most prominent word in an utterance. The options include every word in the sentence, plus an "Unclear" option. **Age**: Listeners determine the speaker's perceived age by selecting from a forced-choice list: Child, Teenage, Adult, Elderly, or Unclear}.

3.3. Human Annotation Collection & the E-VOC dataset

To facilitate our human evaluation, we created the E-VOC corpus. We generated the audio stimuli for this corpus using five representative ITTS systems: Parler-TTS-large-v1 (**Parler-large**) [10, 24], Parler-TTS-mini-v1 (**Parler-mini**) [10, 24], PromptTTS++ [11] (**Prompt++**), UniAudio [13], and gpt-4o-mini-tts (**gpt-4o**) [25].

3.3.1. Transcripts Generation

The generation process combined neutral transcripts with specific style prompts. We first created eight conversational transcripts for everyday contexts (e.g., teacher-student, customer-server) using Gemini 2.5 Pro [20]. We then paired these transcripts with prompts designed for each control dimension. For **acoustic controls**, prompts combined adverbs and adjectives (e.g., "Speak in a Very High tone"). For the **Emo-I.A.**, prompts used intensity-specific adjectives (e.g., "Speak in an Ecstatic tone"). For **Emphasis**, prompts specified the exact word to be stressed (e.g., "Articulate clearly, placing special stress on the term 'Sundays'"). For **Age**, prompts requested a specific age group (e.g., "Use a Child's voice").

3.3.2. Annotation and Quality Control

We recruited native English speakers from the United States via the Prolific platform. All participants completed a brief training session before starting the main annotation task. To ensure the reliability of our data, we implemented a rigorous quality control protocol. We embedded check utterances with gold-standard labels sourced from public corpora, including CREMA-D [26] (for emotion intensity), EMNS (for emphasis), and Nexdata.ai [27]/CREMA-D [26] (for age). We only retained ratings from annotators who demonstrated high accuracy in these check items. Finally, we report two key reliability metrics in Table 2: Inter-Rater Agreement, measured using Cohen's Kappa on the check items, and Worker Performance, defined as the percentage of check utterances that each annotator labeled correctly. At least 5 workers annotate every utterance.

Table 3: This table summarizes model performance on the **Age** and **Emphasis** tasks. Overall performance is reported using accuracy, while class-specific results for the **Age** task are reported using F1-scores. The best performance in each category is indicated in bold.

ITTS	gpt-4o	Parler-large	Parler-mini	Prompt++	UniAudio
Task (Metric)	**Age (Accuracy)**				
Overall	0.289	**0.294**	0.227	0.246	0.211
Task (Metric)	**Age - Class-wise Analysis (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**
Task (Metric)	**Emphasis (Accuracy)**				
Overall	**0.265**	0.152	0.134	0.130	0.040

4. EXPERIMENTAL RESULTS AND ANALYSES

4.1. Adverbs of Degree

As shown in Figure 1, gpt-4o provides the clearest and most consistent mapping from degree adverbs to acoustic features. Figure 2 (top row) extends this analysis to perceived emotion intensity under adverb cues.

Loudness. gpt-4o spans a wide LUFS range with a predictable ordering from "slightly" to "extremely." In contrast, both Parler models show limited variation; PromptTTS++ is nearly flat, and UniAudio remains significantly calmer overall.

Pitch. gpt-4o effectively separates "high" and "low" instructions into distinct F_0 bands. Other systems exhibit smaller, irregular separations, with degree steps that often compress or overlap.

Speaking Rate. gpt-4o again covers the broadest range with a logical progression from "extremely slow" to "extremely fast," while other models show minimal or inconsistent changes.

Emotion. Once again, gpt-4o demonstrates strong, consistent gradation within each emotion, with listeners rating "extremely" and "very" prompts as more intense than "slightly." The other systems show weaker separation and even occasional reversals.

Overall, gpt-4o is the only model that reliably translates degree modifiers into both the intended acoustic shifts and the corresponding perceptual changes based on our experimental settings.

4.2. Emotion–Intensity Adjectives

gpt-4o was the only system to demonstrate reliable control over graded emotional intensity across all four emotion categories (Figure 2, bottom row) in this task. For instance, listener ratings for gpt-4o increased smoothly along the *Happy* graded emotion words (from "Satisfied" to "Ecstatic") and the *Surprised* ones (from "Intrigued" to "Surprised").

Table 4: Confusion matrices for **gpt-4o** on **Age** task. Rows are system outputs (**Labels**), columns are human judgments. Rows indicate the system-predicted labels, and columns show the categories chosen by human annotators. Higher values on the diagonal represent better alignment between instructions and perception.

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

Other models show weaker or flatter distinctions. Parler-large and Parler-mini captured some variation, but the perceptual steps between adjacent adjectives were small. PromptTTS++ often produced nearly indistinguishable outputs across terms, while UniAudio occasionally exhibited reversed trends, with listeners rating mid-level adjectives as more intense than stronger ones. These results suggest that while most ITTS systems can generate distinct categorical emotions, only gpt-4o can reliably control their fine-grained emotion intensity.

4.3. Speaker Age

The **Age** control task was challenging for all systems, with low overall accuracies reported in Table 3. Parler-large and gpt-4o achieved the highest scores, but only achieved accuracies of 0.294 and 0.289, respectively. Class-wise F1-score in Table 3 indicates that all models reproduce *adult* or *teenager* speech most reliably, while *child* and *elderly* voices are much harder to generate. In particular, gpt-4o achieves the strongest recognition for *teenage* and *adult* prompts, whereas UniAudio and Prompt++ perform relatively better for *elderly* prompts. The generation of a child voice was particularly challenging, with extremely low F1-scores across all systems.

Analysis of the gpt-4o confusion matrix (Table 4) further reveals a strong bias: regardless of the prompt, listeners most often perceived the output as adult. This finding suggests that current ITTS models gravitate toward a default adult-like voice with limited control over other age categories.

4.4. Word-level Emphasis

Controlling word-level emphasis was a significant challenge for five ITTS systems. As shown in Table 3, gpt-4o achieved the highest accuracy, yet its score of 0.265 indicates that even the best-performing model struggled. This task highlights a critical area for improvement, as effective emphasis requires precise and consistent coordination of pitch excursion, duration, and intensity at the word level.

5. CONCLUSION AND FUTURE WORK

Conclusion. This work addresses the largely unexplored relationship between natural-language instructions and listener perception in advanced ITTS systems. We introduced a novel evaluation framework centered on graded emotion intensity, using both adverbs of degree (e.g., "slightly happy") and emotionally ordered adjectives (e.g., from "Content" to "Happy" to "Ecstatic") to measure fine-grained control. Our empirical analysis of five leading ITTS models revealed two key patterns: (1) Among the systems tested, gpt-4o was the only one that reliably translated degree modifiers and graded adjectives into perceptually ordered changes in loudness, pitch, speaking rate, and emotion intensity. (2) Fine-grained controls like word-level emphasis and speaker age were inconsistently realized across all models. Most used ITTS systems defaulted to adult-like voices and produced weak or unstable emphasis cues. To sum up, while contemporary ITTS can follow some high-level styles with coarse reliability, achieving consistent and fine-grained control that aligns with human perception remains a significant open challenge.

Future Work. Our large-scale E-VOC corpus, with its 60,000+ human ratings, provides a valuable resource for developing automated evaluation systems. A promising future direction is to use this dataset to train and validate Spoken Language Models (SLMs) like Gemini as reliable [20], scalable proxies for human perceptual judgments. Developing such an automated judge would significantly accelerate ITTS research by enabling faster and more reproducible evaluation cycles [19, 28].

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] Zhihao Du et al., "Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training," 2025.

[3] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in *International Conference on Learning Representations*, 2021.

[4] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in *Proceedings of the 39th International Conference on Machine Learning*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, Eds. 17–23 Jul 2022, vol. 162 of *Proceedings of Machine Learning Research*, PMLR.

[5] 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.

[6] Gabriel Mittag et al., "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Interspeech 2021*, 2021.

[7] Wenze Ren, Yi-Cheng Lin, Wen-Chin Huang, Ryandhimas E. Zezario, Szu-Wei Fu, Sung-Feng Huang, Erica Cooper, Haibin Wu, Hung-Yu Wei, Hsin-Min Wang, Hung yi Lee, and Yu Tsao, "HighRateMOS: Sampling-Rate Aware Modeling for Speech Quality Assessment," 2025.

[8] Kenichi Arai, Shoko Araki, Atsunori Ogawa, Keisuke Kinoshita, Tomohiro Nakatani, Katsuhiko Yamamoto, and Toshio Irino, "Predicting speech intelligibility of enhanced speech using phone accuracy of dnn-based asr system," in *Interspeech 2019*, 2019, pp. 4275–4279.

[9] Apoorv Vyas et al., "Audiobox: Unified Audio Generation with Natural Language Prompts," 2023.

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

[11] Reo Shimizu et al., "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.

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

[13] Dongchao Yang et al., "UniAudio: An Audio Foundation Model Toward Universal Audio Generation," 2024.

[14] Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee, "EmoSphere++: Emotion-Controllable Zero-Shot Text-to-Speech Via Emotion-Adaptive Spherical Vector," *IEEE Transactions on Affective Computing*, vol. 16, no. 3, pp. 2365–2380, 2025.

[15] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen, "emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation," in *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar, Eds., Bangkok, Thailand, Aug. 2024, pp. 15747–15760, Association for Computational Linguistics.

[16] Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia, "VoxInstruct: Expressive Human Instruction-to-Speech Generation with Unified Multilingual Codec Language Modelling," in *Proceedings of the 32nd ACM International Conference on Multimedia*, New York, NY, USA, 2024, MM '24, p. 554–563, Association for Computing Machinery.

[17] Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, Fan Yu, Zhihao Du, Zhifu Gao, ShiLiang Zhang, and Xie Chen, "EmoVoice: LLM-based Emotional Text-To-Speech Model with Freestyle Text Prompting," 2025.

[18] Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu, "SpeechCraft: A Fine-Grained Expressive Speech Dataset with Natural Language Description," in *Proceedings of the 32nd ACM International Conference on Multimedia*, New York, NY, USA, 2024, MM '24, p. 1255–1264, Association for Computing Machinery.

[19] Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, et al., "InstructTTSEval: Benchmarking Complex Natural-Language Instruction Following in Text-to-Speech Systems," 2025.

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

[21] 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).

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

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

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

[26] 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.

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

[28] Cheng-Han Chiang, Xiaofei Wang, et al., "Audio-Aware Large Language Models as Judges for Speaking Styles," 2025.